Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2021 Section "Clinical Information Systems"

Fabregat A, Magret M, Ferré JA, Vernet A, Guasch N, Rodríguez A, Gómez J, Bodí M

A Machine Learning decision-making tool for extubation in Intensive Care Unit patients

Invasive Mechanical Ventilation (IMV) is central to treating patients who are unable to maintain adequate pulmonary ventilation and oxygenation to allow patients to recover. Although IMV can be a life-saving procedure, it also bears significant risks such as ventilator-induced lung injuries or infections as well as long-term problems after recovery. One of the critical decisions regarding IMV is weaning. This includes, amongst other steps, the removal of the endotracheal tube. Patients that need to be reintubated bear several risks and problems associated, including increased mortality (25%-50%). The goal of the current work was to create a machine learning (ML) model that can increase the successful extubation rate in adult Intensive Care Unit (ICU) patients. The model is based on data routinely collected from patients' health record data. Patients included were admitted to an ICU in a Spanish hospital between 2015 and 2019, and received at least 12 consecutive hours of IMV. Variables used for prediction were categorized in four types: T1: time series data (averaged over 20 minutes) (e.g., heart rate); T2: derived variables from T1 (e.g., respiratory rate); T3: discrete event information, (e.g., Glasgow coma scale (GCS)); T4: demographics and admission information (e.g., gender). In total, 20 predictors were used. The resulting dataset had a strong imbalance with regard to successful extubations (1,108 versus 100). Therefore, randomly selected data points of the most frequent class (successful extubation) were removed for the training data set

and/or a weight was assigned to data points. Seven-fold cross-validation was determined appropriate. Extubation/Reintubation was basically identified by finding gaps on the IMV monitor signal larger than 48 hours. As several possible errors have an influence on this gap, a comparison with medical records was necessary to correct the numbers (final dataset: 647 successful and 50 failed). Three different ML classifiers were compared: support vector machine (SVM) with radial basis, gradient boosting machine (GBM) with Bernoulli loss, and Linear Discriminant Analysis (LDA). Mean Accuracy and AU-ROC were used to determine performance. The following scores were achieved: SVM 94.6% and 98.3%; GBM 87% and 96%; LDA 72% and 79%. The results suggest that the top five predictors in descending order of importance are time, GCS, body mass index, respiratory rate-oxygenation index, and plateau pressure. On the other hand, the least relevant predictors in descending order of importance are Spanish Society of Intensive, Critical Medicine and Coronary Units classification code for ICU admission reason, gender, total cumulative dose, total given dose, and ventilation mode. The models should not be applied as a general-purpose predictor of success for programmed extubations or as a monitoring alarm system but as a support tool to validate the medical staff's decision. With the predictive accuracy achieved, the rate of failed extubation (currently 9%) could be reduced to a theoretical 1%. The results suggest that ML tools are especially well suited to support the decision-making protocol based on spontaneous breathing trials to decide about extubation.

Kempa-Liehr AW, Lin CYC, Britten R, Armstrong D, Wallace J, Mordaunt D, O'Sullivan M

Healthcare pathway discovery and probabilistic machine learning

The success of electronic health records has also driven several other research areas such as knowledge management in healthcare, which basically involves four steps: (1) data access; (2) knowledge discovery; (3) knowledge translation and interpreta-

tion, as well as (4) knowledge description, integration and sharing. An important role hereby is played by healthcare pathways that incorporate the operational knowledge of a healthcare organization by defining the execution sequence of clinical activities as patients move through a treatment process. In many cases, these pathways result from clinician-led practice rather than explicit design, which leads to several problems (e.g., lack of update). The study aims to combine healthcare pathway discovery with predictive models of individualized recovery times after appendicectomy. Particular emphasis is set on easy to interpret models for clinicians. The predictive model takes the stochastic volatility of pathway performance indicators into account and can replicate the dominant mode as well as the fat tail of the empirical recovery time distribution. To mine the pathways, the ProM software was used. First, healthcare pathway variations were discovered and then reduced (clustering, merging consecutive activities, condense repetitive patterns) to meaningful models. In a second step, conformance of these models with actual patient traces is evaluated, including new findings into the model leads to an iterative approach between pathway discovery and conformance analysis. The third step involves data enrichment, which comprises two stages: healthcare pathway performance evaluation and healthcare pathway performance analysis. The main objectives of evaluating healthcare pathway performance are to understand the strengths and weaknesses of the current pathway design. Analyzing the performance of healthcare pathways with respect to pathway variants and other possible influencing factors like demographics or patient-specific pathway characteristics (e.g., surgery duration) is the final step of the proposed process mining pipeline. For the appendicitis model, 13 pathway variants were discovered, whereas the top four variants accounted for approximately 88% of the patient traces. In a next step, it was analyzed if the variants are relevant features or covariates for explaining the stochastic volatility of postoperative length of stay. To build two probabilistic machine learning models, 415 individual patient traces were used. The two models showed promising results to explain the length of stay. Summarizing, the pro

posed process mining pipeline successfully constructed concise pathway models for the appendicitis case study and, therefore, supported the generation of probabilistic machine learning models.

Li Y, Nair P, Lu XH, Wen Z, Wang Y, Dehaghi AAK, Miao Y, Liu W, Ordog T, Biernacka JM, Ryu E, Olson JE, Frye MA, Liu A, Guo L, Marelli A, Ahuja Y, Davila-Velderrain J, Kellis M

Inferring multimodal latent topics from electronic health records

Nat Commun 2020;11(1):2536

Electronic health records (EHRs) are heterogeneous collections of patient health information that would support multiple uses such as risk prediction, clinical recommendations, or individual therapeutic concepts. However, raw data in EHRs is in many cases not directly processable, especially when building formal models. Different challenges such as non-standardized clinical notes, heterogeneous data types, missing standardization, or diagnosis-driven lab tests pose challenges. Appropriate and effective computational methods have the potential to overcome those challenges and provide access to an encyclopedia of diseases, disorders, injuries, and other related health conditions, uncovering a modular phenotypic network. The paper introduces MixEHR to: (1) distill meaningful disease topics from otherwise highly sparse, biased, and heterogeneous EHR data; and (2) provide clinical recommendations by predicting undiagnosed patient phenotypes based on their disease mixture membership. MixEHR builds on collaborative filtering and latent topic modeling and can model various EHR categories with separate discrete distributions. A variational inference algorithm that scales to large-scale EHR data was created. The model was applied to three EHR datasets: (1) Medical Information Mart for Intensive Care (MIMIC)-III (50,000 intensive care unit admissions); (2) Mayo Clinic EHR dataset containing 187 patients, including with 93 bipolar disorders and 94 controls; (3) The Régie de l'assurance maladie du Québec Congenital Heart Disease Dataset (Quebec CHD Database; more than 80,000 patients with congenital heart disease). The authors followed a probabilistic joint matrix factorization approach. The high dimensional and heterogeneous clinical record was projected onto a low dimension probabilistic meta-phenotype signature, reflecting the patient's mixed memberships across diverse latent disease topics. Factorization is carried out at two levels. At the lower level, data-type-specific topic models, learning a set of basis matrices for each data type, were applied. A common loading matrix that connects the multiple data types for each patient was used at the higher level. The approach was used, among others, to define a disease comorbidity network, create patient risk prioritization, EHR code predictions, or mortality predictions from the given datasets. Overall, the MixEHR approach's accuracy scores top compared to other existing approaches. MixEHR can infer expected phenotypes of a patient conditioned only on a subset of clinical variables that are perhaps easier and cheaper to measure. Currently, data are a set of two-dimensional matrices of patients by measurements in the model. To model higher dimensional objects such as patient by lab test by diagnoses, MixEHR could be extended to a probabilistic tensor-decomposition framework.

Weemaes M, Martens S, Cuypers L, van Elslande J, Hoet K, Welkenhuysen J, Goossens R, Wouters S, Houben E, Jeuris E, Jeuris K, Laenen L, Bruyninckx K, Beuselinck K, André E, Depypere M, Desmet S, Lagrou K, Van Ranst M, Verdonck AKLC, Goveia J

Laboratory information system requirements to manage the COVID-19 pandemic: A report from the Belgian national reference testing center

J Am Med Inform Assoc 2020;27(8):1293–9

The paper describes the challenges faced by the Belgian National Reference Center for COVID-19 testing at the University Hospitals Leuven, when demand passed allocated surge capacity during the initial phases of the COVID-19 pandemic. This includes the design, implementation and requirements of laboratory information system (LIS) functionality related to managing increased test demand during the COVID-19 crisis. In particular, all phases in laboratory testing were streamlined: the pre-laboratory phase (test ordering, sample packaging, and shipping); the pre-analytical phase (sample registration, tracking, and test prioritization); and the post-analytical phase (automated reporting and facilitating data-driven policy-making). Apart from COVID-19 testing, the laboratory concerned performs more than 12,000,000 lab tests a year. The LIS is in-house developed and maintained by a dedicated team. The system includes a computerized physician order entry (CPOE) module for in-house test ordering, which is fully integrated into the electronic health record (EHR). All external orders were initially paper-based and required that request forms accompany the sample. In the course of the analysis, 17 major challenges were identified in the different phases of the testing process. Selected solutions included: a COVID-19 specific CPOE module was linked to both the LIS and EHR, allowing to automatically retrieve demographic information, which dramatically improved metadata completeness; a "COVID-19 status" button on the main page of the EHR of each patient was displayed, showing in real-time the results of SARS-CoV-2 laboratory testing; a database with contact details and preferred reporting methods (e.g., fax, email, electronic mailbox system) of every laboratory in Belgium was compiled, to enable automated test reporting (resulted in more than 98% automated reporting). To successfully implement such changes in a short time, several prerequisites apply. The authors, therefore, recommend that crisis management teams not only consist of staff focused on increasing analytical capacity but also information technology-staff and to apply change management frameworks. To summarize, the most effective solutions reported were to streamline sample ordering through a CPOE system and reporting by developing a database with contact details of all laboratories in Belgium. In addition, the implementation of R/Shiny-based statistical tools facilitated epidemiological reporting and enabled explorative data mining.