doi:10.20944/preprints201911.0073.v1

Beyond Traditional Covariates in Medical Informatics

Uri Kartoun
Center for Computational Health at IBM Research
Cambridge, MA 02142, USA
Corresponding: uri.kartoun@ibm.com

Abstract

Deep behavioral covariates (DBCs) introduced in this perspective form a new class of covariates that have the potential to enhance the performance of predictive models and improve analytics in clinical decision support applications. DBCs can measure how engaged a patient tends to be and how he or she tends to respond to events, and they may be highly predictive of the patient's outcomes for a planned treatment. DBCs may potentially serve as a standard to measure patient engagement and activation and may form highly efficient mechanisms for improving patient outcomes.

Keywords: deep behavioral covariates; clinical informatics; predictive modeling; electronic medical records; machine-learning; data-mining

I. BEYOND MACHINE LEARNING

Machine learning is a group of techniques that allows computers to not only process data faster than humans but also to process it more intelligently. Machine learning allows a computer to observe large collections of data elements and provide accurate predictions on the occurrence of future events [e.g., Obermeyer and Emanuel, 2016; Kartoun et al., 2017(a); Kartoun, 2017(b)]. Existing methodologies to improve accuracy in prediction tasks often focus on which algorithm to apply [e.g., Beam and Kohane, 2018; Xiao et al., 2018]. The literature, however, barely emphasizes the importance of constructing a wider range of more advanced covariates that may capture the patient's condition better. Covariates capable of dynamically capturing patient behavior may enhance the performance of the most sophisticated computational algorithms [e.g., LeCun et al., 2015; Hinton, 2018] as well as of simple brute-force mechanisms [e.g., Kartoun, 2017 (c)].

II. PREDICTIVE MODELING IN HEALTH CARE

Remarkable hardware-based advances in health care include medicinal contact lenses [Ciolino et al., 2016], tooth sensors [Tseng et al., 2018], and ingestible sensors [Kalantar-Zadeh et al., 2018]. Other astonishing advances include pharmaceutical-related mechanisms such as synthesized antibiotics [Parmar et al., 2016] and cancer vaccines [Sagiv-Barfi et al., 2018]. To advance health care further, predictive modeling, a domain that combines computer science and statistics, has received significant increased attention, yielding highly valuable scientific publications and applications. This rapid growth may be explained by factors such as the greater accessibility of medical records enabled by the ever-improving software and hardware security mechanisms, the development of friendlier online collaborative methods, and increased speeds of data storage and transfer.

One of the most intriguing challenges in applying predictive modeling to health care is the ability to identify individuals at high risk for a future undesirable medical outcome [e.g., Cheng et al., 2016; Devinsky et al., 2016; Ng et al., 2016; Sulieman et al, 2016; Gottlieb et al., 2017; Wang et al., 2017]. When a high-risk individual is identified, the clinical teams are better positioned to allocate resources more accurately and can attempt to prevent or delay undesirable outcomes. Another challenge is the ability to accurately identify the most informative set of covariates associated with a patient and thereby the most effective means of predicting the outcome [e.g., Kartoun, 2018(a)]. Accessible covariates in EMRs typically include demographic details, laboratory observations, comorbidities, and features extracted from clinical narrative notes. Typically, hundreds of such covariates are available at the patient level to develop a predictive model.

Many publications describe predictive models for a variety of patient situations, such as readmission, the development of new diseases, monitoring the progression of a patient's current disease, and mortality. Predicting patient outcomes may include forecasting whether a patient's condition will improve or decline and the magnitude of such changes. Predictions are achieved by applying a computational algorithm capable of calculating the probability for each possible outcome given a set of selected covariates. For example, within a defined follow-up time window, the algorithm could estimate the extent to which a patient will recover, partially or fully.

Frequently, patient records include both structured and unstructured data. Structured data are organized in tables and include elements such as date of birth, gender, laboratory measurements, and International Classification of Diseases codes (standard disease classifications defined and published by the World Health Organization). At a higher level of heterogeneity, unstructured data contain elements such as a patient's natural language description of his or her symptoms or a physician's or nurse's clinical narrative about diagnoses, treatment options, and laboratory results. A patient's EMR can be parsed using one or more natural language processing (NLP) techniques to identify and extract a wide variety of covariates that may help predict patient outcomes. Processing notes can identify covariates capable of indicating whether the patient is adherent for diet or medications. Moreover, more complex covariates such as the frequency of appointments with health care professionals (or trends or changes in the frequency) or trends or changes in laboratory measurements can also be extracted and used to better predict patient outcomes. Covariates extracted from patient health care data can serve as useful predictors to trigger decision-making, such as to advise a physician on various treatments. Covariates representing attributes such as smoking status and alcohol use have proven indicative of how likely a treatment is to be successful in mitigating or controlling a medical disorder. Current methods identify a relatively limited and rigid set of covariates that may help predict outcomes.

III. TRADITIONAL COVARIATES USED IN PREDICTIVE MODELING

The variety of covariate types could be categorized into classes, ranked by their level of ease in collection and construction. Covariates collected by looking at the patient, such as gender or race, are the easiest to capture because they require neither any interaction with the patient nor any measurements or calculations. Covariates such as age, marital status, mood status, pain level, alcohol use, and tobacco use require only minimal interaction with the patient. Capturing covariates such as weight, height, and blood pressure can be accomplished within a few minutes because they require the use of easily accessible equipment available in any physician's office, including the use of EMR data management systems that can quickly provide details regarding the patient's current and past comorbidities. At an intermediate level of complexity are the patient's laboratory test results (such as creatinine, hemoglobin, and glucose)—such covariates require analysis of urine or blood samples by an external laboratory with more advanced measurement devices than those found in a standard outpatient setting. Processing measurements stored in the EMR can be challenging given the heterogeneity of measurement units and reference values that are often unique to a certain measurement. Of a higher complexity are covariates representing the genetic profile of the patient, often requiring an external specialized laboratory and associated with high costs to process.

Another set of covariates commonly used to classify and predict patient outcomes include those extracted from clinical narrative notes (such as progress notes, operation notes, and discharge summaries)—such covariates require the use of advanced NLP and machine leaning algorithms and often require time-consuming manual chart reviews, typically performed by physicians and nurses. Additional covariates rely on the use of wearable devices that are often integrated with the patient's EMR [Kartoun et al., 2017(d)]—such covariates require the application of advanced time series classification techniques on the collected data that represent large collections of continuous and often noisy values.

IV. CHARATERISTICS OF DEEP BEHAVIORAL COVARIATES

At a significantly higher complexity than the covariate classes specified above comes a new class of covariates, referred to exclusively in the current manuscript as “deep behavioral covariates” (DBCs). Such covariates are composed of behavior-related data elements captured over several points of time throughout the patient's longitudinal horizon and can assess the patient's behavioral dynamics throughout his or her interaction with one or more care systems. An example for a class of DBCs is the association between stimuli and responses [Kartoun et al., 2018(b)]. Such DBCs may be composed of sub-data elements, such as completion status, associated with traditional data elements such as encounter type, laboratory value, or procedure. DBCs may also be composed of multiple data elements distinguished by type and time, such as an abnormal laboratory value followed by an increased number of office visits. Furthermore, such covariates may be extracted from multiple encounter types, such as online messages, telephone calls, and in-person appointments, as well as from patient portals, such as heart rate data uploaded by a patient to an online portal. These covariates may also depend in part on data captured via wearable and ingestible sensors.

An event may refer to anything that occurs at one or more points in time in the patient's medical history. Examples of events may include receiving a referral; the completion of a visit or appointment; failing to attend a visit or appointment; canceling or rescheduling an appointment; ordering or requesting health maintenance; completing or failing to complete health maintenance by a set date; ordering, requesting, or scheduling a laboratory measurement or a procedure; and completing, rescheduling, canceling, or missing a scheduled laboratory report or procedure. Similarly, an event may include receiving, recording, or providing a biometric reading or finding such as blood pressure, heart rate, glucose level, or any other physical measurements related to health maintenance, laboratory reports, or procedures, as well as receiving a diagnosis or reviewing a laboratory report or other item in person or through an online portal.

A stimulus-response association can be determined within a single event in the patient data. An appointment, procedure, or a laboratory test may be identified in the EMR, along with a corresponding status label indicating the completion status of each event. A data mining algorithm may identify a stimulus event of scheduling an appointment, procedure, or test and determine the response event based on the status label. Each event may have a label such as "completed," "canceled," "deferred," "not done," "declined," "ordered," "pending," "active," and the like. If the event is a procedure and the status is "completed," the data mining algorithm may identify a response event indicating that the procedure was completed. Similarly, if the status indicates that it was “canceled,” “declined,” “not done,” or similar, the data mining algorithm may create a link indicating that the appointment was not completed. In this way, a variety of DBCs will be formed, such as the percentage of scheduled appointments completed, using a single identified event in the patient data and the corresponding label.

The event type (stimulus, response, or independent) and the links between events are identified based on a set of predefined rules corresponding to DBCs. A health care provider or a professional may define a DBC as the ratio between scheduled appointments and completed appointments. Generally, any relationship between stimulus events and response events can be used to create the DBC. Each such covariate generally relates to a certain behavior or behavioral patterns of the patient and helps determine how the patient tends to respond to a particular stimulus. A stimulus event is an event in the patient’s medical history that may trigger some sort of response by the patient. Receiving a medical diagnosis might be classified as a stimulus event because the patient may want to schedule a follow-up visit, fill a prescription, or change a habit. A stimulus event may be initiated or completed by the patient (e.g., recording high blood pressure at home), or it may be caused or initiated without action by the patient (e.g., receiving a diagnosis from a physician). A response event corresponds to the patient’s response to a stimulus. For example, picking up a prescription from a pharmacy may be classified as a response event in response to the stimulus event of receiving the prescription from a health care provider. The events then are extracted to identify stimulus events and response events, as well as to determine the associations or links between these events.

Response events include only actions taken by the patient, such as scheduling an appointment, and do not include things that happen to the patient, such as receiving a diagnosis. Notably, each response event is associated with at least one stimulus event, but a stimulus event may be associated with any number of response events, including none. Additionally, a single event may be identified as a stimulus event, response event, or independent event depending on various factors, including the context of the event. An event corresponding to scheduling an appointment with a specialist may be a stimulus event (with a corresponding response event upon attending, canceling, or rescheduling the appointment) or a response event (in response to a stimulus event corresponding to receiving a diagnosis or referral).

A single event may be both a stimulus event and a response event. For instance, recording or uploading a blood pressure reading may be a response to instructions from a doctor to periodically record blood pressure, as well as a stimulus to schedule an appointment (e.g., if the reading is abnormal or outside the preferred range). Additional examples of stimulus-response associations may include receiving a physician referral and scheduling or completing the referral visit; having health maintenance ordered by a provider and completing or failing to complete the maintenance; ordering, requesting, or scheduling a laboratory or procedure and completing, rescheduling, or canceling the laboratory or procedure; scheduling an appointment and completing the appointment; receiving a medical reading or a diagnosis and opening or reviewing the results; and recording an abnormal reading and continuing to record or upload results (at the same rate or an increased or decreased rate).

Deep behavioral associations may be identified between events that appear to be clinically unrelated, such as receiving the first diagnosis of a disorder and scheduling or completing an unrelated appointment or procedure. For example, a patient may receive a diabetes diagnosis and subsequently schedule or complete a colonoscopy. While these two events are seemingly unrelated, they may in fact constitute a stimulus-response pair because they demonstrate the patient’s continuing interest in his or her general health, despite (or perhaps because of) the recent adverse diagnosis. Thus, an association between two clinically unrelated events (events that pertain to different, unrelated medical disorders) is identified based on determining that a stimulus event is the patient measuring or receiving some abnormal value or adverse result (e.g., from a procedure, diagnosis, or laboratory test). In such a scenario, one or more corresponding response events may be the scheduling or completion of any health care event, such as an appointment or a procedure. This association is identified because it may indicate that the patient remains engaged and active in maintaining his or her health, even upon receiving bad news. When an adverse event occurs, one or more corresponding responses that involve preemptive or preventative care may be identified, such as scheduling a colonoscopy or checkup, when no apparent symptoms of sickness correspond to the appointment.

V. EXPECTED BENEFITS TO ENHANCE FUTURE RESEARCH AND APPLICATIONS

After the DBCs are computed, they may also be integrated to determine patient outcomes. A data mining process then can analyze longitudinal data to identify points in time when a physician can provide decisions regarding a patient's disorder. For instance, the process can identify all office encounters associated with abnormal blood pressure values for a population of patients suffering from hypertension. The process can then extract a large collection of covariates, including traditional ones as well as DBCs. Further, the model may identify an outcome for each patient in the population. A binary outcome is one in which a measurement such as blood pressure is either under control or abnormal within a given follow-up window. A complex outcome could be a set of new symptoms. The process then uses a feature selection algorithm to identify the most informative covariates capable of predicting the outcome [e.g., Kartoun et al., 2018(a)]. When a patient is seen by a physician, that patient's covariates (including his or her DBCs) are used to query a predictive model to help the physician choose the potentially most efficient treatment. Such models are data structures that combine associations between past patients, covariates, and outcomes. A patient's covariate representing appointment completion rate may indicate how likely he or she is to keep up with treatments and appointments in the future, which may influence which treatment plan the physician follows. Thus, a treatment plan requiring frequent check-ins with a physician may be more suitable for a patient with a history of attending appointments as planned, which will lead to a higher probability that the disorder will be managed if this treatment plan is selected.

DBCs form a new class of covariates that have the potential to enhance the performance of predictive models and improve analytics in clinical decision support applications. DBCs can measure how engaged a patient tends to be and how he or she tends to respond to events, and they may be highly predictive of the patient's outcomes for a planned treatment. DBCs may potentially serve as a standard to measure patient engagement and activation and may form highly efficient mechanisms for improving patient outcomes.

Keywords: Deep behavioral covariates, clinical informatics, predictive modeling, electronic medical records, machine-learning, data-mining.

Conflict of interest statement

The author has declared that no competing interests exist.

Funding statement

The author received honoraria and travel funding from The American Association for the Study of Liver Diseases (October 2017).

References

Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317–8.

Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: a deep learning approach. Proceedings of the 2016 SIAM International Conference on Data Mining 2016:432–40.

Ciolino JB, Ross AE, Tulsan R, Watts AC, Wang RF, Zurakowski D, Serle JB, Kohane DS. Latanoprost-eluting contact lenses in glaucomatous monkeys. Ophthalmology 2016;123(10):2085–92.

Devinsky O, Dilley C, Ozery-Flato M, Aharonov R, Goldschmidt Y, Rosen-Zvi M, Clark C, Fritz P. Changing the approach to treatment choice in epilepsy using big data. Epilepsy Behav 2016;56:32–7.

Gottlieb A, Yanover C, Cahan A, Goldschmidt Y. Estimating the effects of second-line therapy for type 2 diabetes mellitus: retrospective cohort study. BMJ Open Diabetes Res Care 2017;5(1):e000435.

Hinton G. Deep learning-a technology with the potential to transform health care. JAMA 2018;18;320(11):1101–2.

Kalantar-Zadeh K, Berean KJ, Ha N, Chrimes AF, Xu K, Grando D, Ou JZ, Pillai N, Campbell JL, Brkljača R, Taylor KM, Burgell RE, Yao CK, Ward SA, McSweeney CS, Muir JG, Gibson PR. A human pilot trial of ingestible electronic capsules capable of sensing different gases in the gut. Nature Electronics 2018(1):79–87.

Kartoun U, Corey K, Simon T, Zheng H, Aggarwal R, Ng K, Shaw S. The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLOS ONE, 2017(a);12(10):e0186301.

Kartoun U. A user, an interface, or none. ACM Interactions 2017(b);24(1):20–1.

Kartoun U. Text nailing: an efficient human-in-the-loop text-processing method 2017(c). ACM Interactions 2017;24(6):44–9.

Kartoun U, Lu F, Park Y, Ng K. Selective proximity-based interrogation of portable health monitor device to assist physician evaluation of patient habits. USPTO patent application (filed by IBM). 2017(d).

Kartoun U. Toward an accelerated adoption of data-driven findings in medicine: Research, skepticism, and the need to speed up public visibility of data-driven findings. Med Health Care Philos 2018(a).

Kartoun U, Ng K, Chiu A, LaScaleia M, Park Y, Honour M, Das A, Tang P. Method for identifying and extracting stimulus-response variables from electronic health records. USPTO patent application (filed by IBM). 2018(b).

Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes 2016;9(6):649–58.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216–9.

Parmar A, Iyer A, Vincent CS, Van Lysebetten D, Prior SH, Madder A, Taylor EJ, Singh I. Efficient total syntheses and biological activities of two teixobactin analogues, Chem Commun 2016.

Sagiv-Barfi I, Czerwinski DK, Levy S, Alam IS, Mayer AT, Gambhir SS, Levy R. Eradication of spontaneous malignancy by local immunotherapy. Sci Transl Med 2018;10(426). pii:eaan4488.

Sulieman L, Fabbri D, Wang F, Hu J, Malin BA. Predicting negative events: using post-discharge data to detect high-risk patients. AMIA Annu Symp Proc 2016:1169–78.

Tseng P, Napier B, Garbarini L, Kaplan DL, Omenetto FG. Functional, RF-Trilayer sensors for tooth-mounted, wireless monitoring of the oral cavity and food consumption. Adv Mater 2018;30(18):e1703257.

Wang Y, Iyengar V, Hu J, Kho D, Falconer E, Docherty JP, Yuen GY. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. Front Psychiatry 2017;8:114.

Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018.