6.S897/HST.956 Machine Learning for Healthcare

Lecture 21: Automating Clinical Workflows

Instructors: David Sontag, Peter Szolovits

1 Improving Medical Care: Overall

How to integrate the models we have been discussing in class into real world medical settings? So far, we've really only discussed how particular models and tools can be used for various medical contexts, but we haven't looked at how these advances can integrate into the monstrous existing medical workflow. Unfortunately, new technologies cannot simply be inserted and then be accepted by all. The medical system is bloated, behind in technology capabilities, filled with stubborn doctors, and lacking well structured data for all contexts. Being able to integrate into the existing workflows of the medical field is essential for any of these advances to have a meaningul and lasting impace.

1.1 Expert Systems

With the expert systems idea, the basic idea is study the practices and methods of the leaders and experts of a given field, and then have everyone else follow their lead. This is because the best doctors are much much better then average doctors, so using these experts can provide decision support raise the average performance of all doctors. So for example, if we are trying to find the best methods for treating liver cancer, we'd find the ten best liver oncologists in the world and have then have protocols and systems follow their work.

1.2 Protocol Systems

What if instead of trying to raise the average of all doctors, we just try to make the bad doctors decent (even at the expense of making great doctors a little worse)? In the past, we measured success and failure simply by pain and suffering, but now, monetary cost is a much more significant factor. Since bad doctors cost the most money to the system (patients have to make more repeat visits, wrong diagnoses cause secondary problems, etc.), it makes sense to focus on the bad doctors and make them average by having all doctors follow set protocols, regardless if they are a good or bad doctor. While it may actually make great doctors worse then they were before, it also narrows the variance between doctor quality, which is the most preferable outcome.

1.3 Which is better?

Hospital consensus is that the Protocol idea is ideal, as it lowers quality variance and thus keeps costs as low as realistically possible. Another issue with the Expert Systems idea is that it is hard to generalize the ideas/methods of a few doctors into every case. If none of the top doctors have experience dealing with patients in the rural United States, we can't expect those Expert Systems to be inherently useful for those rural patients. Furthermore, while the Expert Systems idea aims to improve the average quality of doctors, it doesn't actually specifically target bad doctors, which are the biggest strain on resources (money).

2 Narrowing Performance Distribution

2.1 Guidelines and Protocols

With hospitals and providers wanting to provide generalized care to reduce quality variance, standard guidelines were required to be made from learned bodies prescribing appropriate methods to diagnose and treat patients. These methods to diagnose and treat often come from and are dependent meta-level analysis of clinical trial results (including all usual issues coming from the fact that there is often a lack of necessary trial for many conditions).

From the analysis, top-down guidelines are provided for doctors to follow, usually in steps of severity or in flow charts. For example, the first guideline in the Take-Home Messages to Reduce Risk of Atherosclerotic Cardiovascular Disease (ASCVD) through Cholesterol Management" guide is "In all individuals, emphasize heart-healthy lifestyle across the life-course", an extremely broad guideline that's appropriate really for any patient. However, as we get deeper into guidelines we get into much more specific scenarios, often for much more serious issues around the condition, such as step 5 of the take-home messages being "In patients 40 to 75 years of age with diabetes mellitus and LDL-C 70 mg/dL (1.8 mmol/L), start moderate intensity statin therapy without calculating 10-year ASCVD risk". Obviously that is a mouthful, but it highlights how specific treatments can quickly be because of how the same disease can manifest in different environments, so having a set of guidelines that can generalize to almost all situations is essential.

Importantly, the prescribed guidelines must be able to utilize both quality of evidence and strength of recommendation in order to generalize to as many permutations of the ailment as possible. In particular, the color coded strength of recommendation indicators are particularly important they help inform how much a particular treatments benefits outweigh the risk, in order to inform doctors how acceptable their recommendations may be. As shown in the guideline flow chart, we see the risks of various actions indicated by their color. A low risk action like switching to a moderate intensity statin if high-intensity statins are not tolerated makes complete sense, and actually alleviates risks posed by the statins. But other actions, like adding ezetimibe in certain situations (orange box), highlight risks that come with actions, and thus forces doctors to decide if those risks are entirely necessary (which they very well may be).

2.2 So where can you find these guidelines?

Unfortunately, the main source of these guidelines, AHRQ's National Guideline Clearinghouse, was shut down in July 2018 by the current administration. However, there is a new source that contains around 2,000 of these guidelines at www.guidelines.com (covering Counseling, Diagnosis, Evaluation, Treatment, etc.).

2.3 Top-Down or Bottom-Up?

Top-Down set ups are what we've most seen above, where we collect data on different medical issues and treatments, and then create a high level set of guidelines for how doctors should treat specific conditions. Alternatively, what if we want to build systems and protocols based on what's already working: Nursing Care Plans Clinical Pathways!

In the real world, doctors interact with patients for hardly any time. Most of the care is done by nurses and their care plans. Because there are so many nurses and care plans that have been tried, nurses have already found out the general plans and procedures that work best for particular hospital visits. That's great if your a nurse with 20 years of care plan experience, but if we want to bring together all of these care plans and generalize, as we could assume they would cover many more types of cases than we could possibly account for with top-down guidelines?

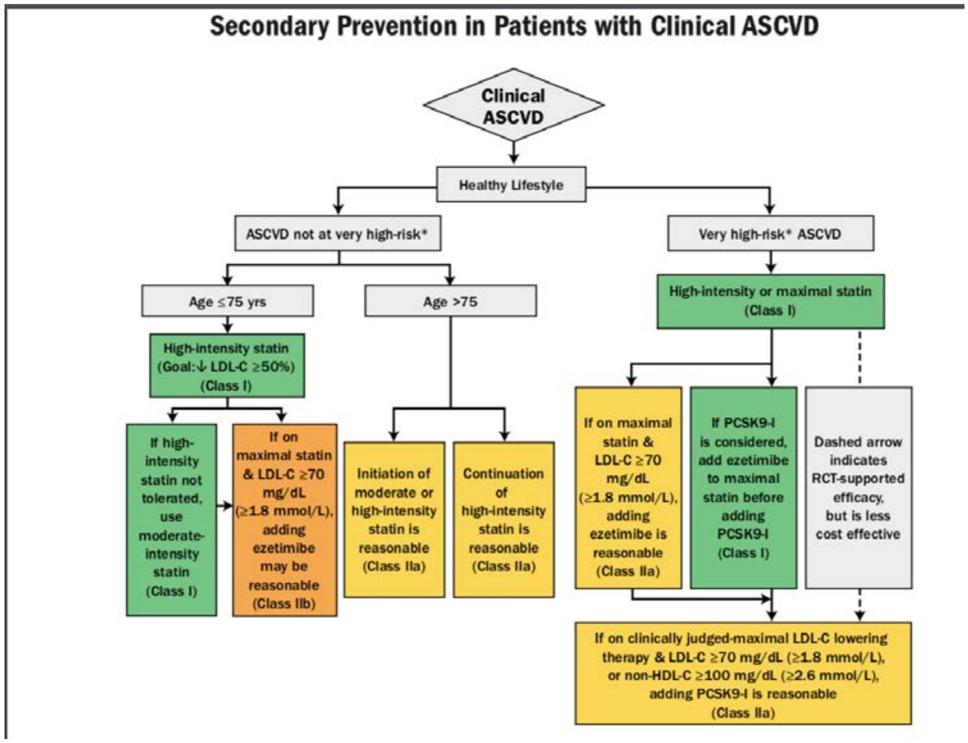

Figure 1: Secondary Prevention in Patients with Clinicial ASCVD Guideline Flowchart

3 Real World Example: Clinical Pathways COWpath (Zhang et al 2014)

The purpose of this example paper [ZPL14] was to learn, visualize and understand real world clinical pathways from electronic health records. Because these pathways are often not in writing, or set in stone, being able to identify the pathways is crucial in understanding patient care, and potentially fixing any current problems with how go about these pathways.

3.1 Overall structure

The overall structure in finding and learning these pathways was as follows:

- Obtain treatment data from electronic health records
- Pre-process the text data and set up the problem model
- Identify and cluster patient subgroups that can then be explored
- Mine for common treatment patterns and pathways in those subgroups
- Have medical experts evaluate those pathways to identify the practice-based clinical pathways

3.2 Representation

In order to correctly mine the treatment patterns and pathways, it was important correctly represent various events. A event originates in a visit, with sets of procedures, medications, diagnoses all being applicable. All

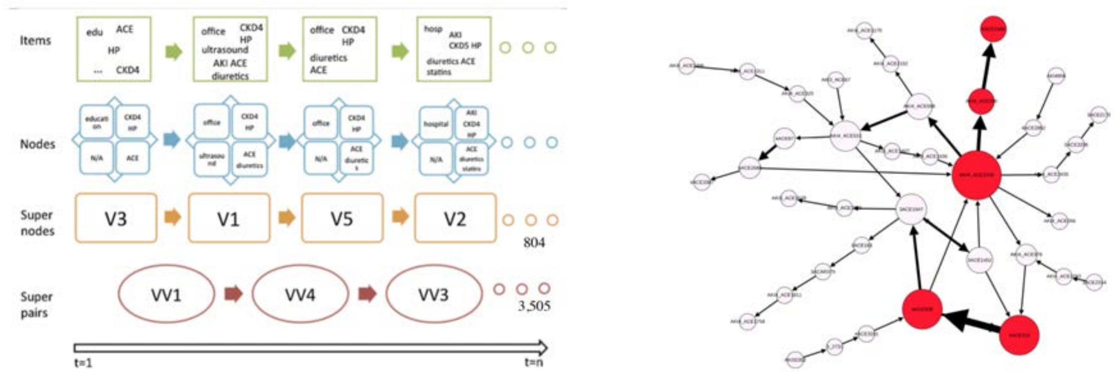

Figure 2: Representation Steps (left) and subgroup cluster event transitions (right)

of these events are then abstracted into super nodes that capture unique combinations of the events from a visit. When those super nodes are time ordered, a visit sequence is produced to describe those events over time. Finally, super-pairs can be created between pairs of super nodes which in turn allows for the creation of Markov chains and analysis (as shown in the figure).

3.3 Clustering and Analysis

Once the researchers have a representation, they can start clustering based on the length of common sub sequences between pairs of visit sequences, turning a hierarchic clustering into distinct subgroups to be studied.

These subgroups are then studied individually, and in particular, the transition matrices between super pairs are calculated. These transitions are visualized to show all the potential pathways of these subgroups based on the real world health data. Once they are parsed further, distinct clinical pathways can be identified and then reviewed by professionals, allowing us to formally learn these clinical pathways at a high level.

4 Decision support from local data

4.1 Adaptive Order Menus: Klann et al (2013)

Klann et al's 2013 paper *Decision support from local data: Creating adaptive order menus from past clinician behavior* describes the utilization of treatment and comorbidity data to suggest actions to clinicians [KSDS13]. The paper uses 3 years of data from the Regenstrief Clinic, focusing on the issues of: back pain in the emergency department (n=9228), inpatient pregnancy (n=4843), hypertension in the Urgent Visit Clinic (n=1821), and altered mental state in the ICU (n=1546). For each domain, they considered the 40 most frequent orders (at a low granularity - e.g. drug, but not dosage) and 10 most frequently-occuring comorbidities.

4.1.1 Modeling Clinician Behavior

To model clinician behavior, a *wisdom of the crowd* technique was implemented, hinging on the idea that the average behavior of many physicians is better than that of any individual. However, although it did generally reduce variance, this model was problematic because it took away context and transitive associations from the recommendations.

Thus, they decided upon a Bayesian network models of diagnoses, possible orders, and evidence from completed orders to learn these decision support rules and deal with more complicated cases. This was implemented via Tetrad's "greedy equivalence search" algorithm. Examples of the resulting networks can

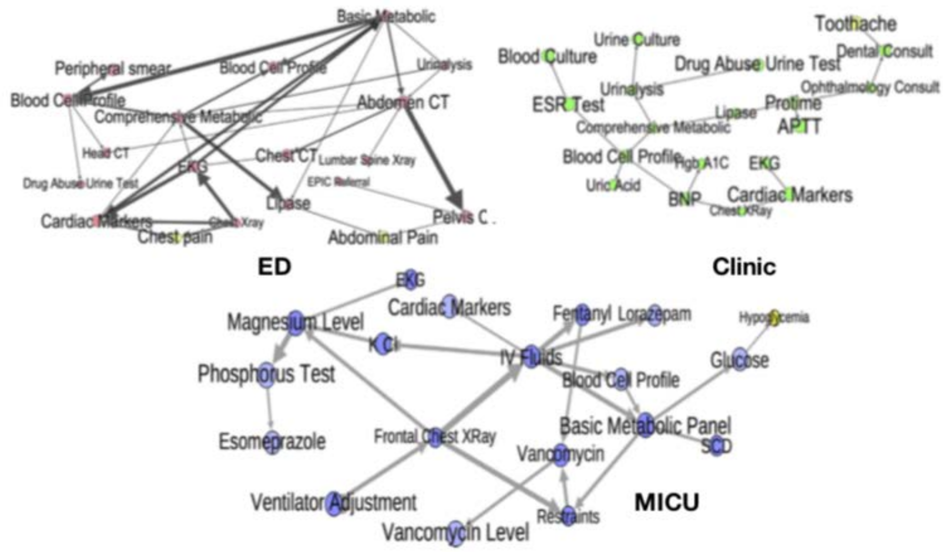

Figure 3: This figure depicts Bayesian networks for various clinical problems.

be found in Figure 3. In these graphs, nodes represent orders and problems with size proportional to AUC, and edges represent strengths of relationships between nodes.

4.1.2 Iterative Treatment Suggestion (ITS)

Based upon the Bayesian networks, we can perform Bayesian inference using the fixed nodes of orders that have already been completed. Using these, we compute the probabilities of possible orders (that have not already been done) and present these in descending probability order to a clinician as possible next steps. This iterative treatment suggestion continues until the user concludes a session.

The goal of ITS was to predict next orders given the context of diagnoses and orders placed. To evaluate ITS, the AUC of each recommended action was calculated, and its position on the recommendation list was also considered. As a reference point, recommendations were compared to assocaiton rule mining. On average, the BN weighted average position (by frequency of order) ranged from 0.23 positions higher [medical ICU] to 4.04 positions higher [back pain in ED]. Generally, and especially for at least the top 10 recommendations, the model performed fairly well.

4.2 Clinical Decision Support System Malfunctions: Wright et al (2016)

Some problems did arise with the implementation of clinical decision support systems (CDSSs). Wright et al describe and analyze patterns in these systems in *Analysis of clinical decision support system malfunctions: a case series and survey* [AWB16]. Using data from Brigham and Women's Hospital, they encountered the following problems:

1. amiodarone identifier - A software update changed the internal identifier for this drug. As a result, a thyroid function alert stopped functioning. This alert fired after patients were on the drug for at least a year, so levels of the alert dropped subtly as the error propogated. Thus, upon fixing the logic, an abrupt increase increase in alert firing rate ensued.

2. lead screening - A discontinuity was observed in testing for 2-year-old children, but not in 1, 3, or 4-year-olds. This error was attributed to inadvertent edits to the rule logic, but logging data was lost, so the root cause was never determined.

3. EHR software update - Upon an update of EHR software, numerous alerts seemed to fire for no reason.
4. external system malfunctions - An external drug classifcation system error caused some patients to be suggested drugs that they were already taking.

As a result, the authors suggested that CDSS malfunctions are common and can exist subtly, so we need to implement methods to prevent and detect them.

5 Change-Point Detection to Monitor Rule Firings

Knowing the importance of monitoring CDSS malfunctions, we now come upon the question of whether this monitoring can be done automatically. In terms of detecting malfunctions, generally, we are just looking for abrupt changes in the general trend of readings or alert firings. Thus, we implement a *dynamic linear model with seasonality* to detect change. To account for the fact that "normal" values change over the course of the week, we also implement a seasonality factor.

5.1 Dynamic Linear Model (DLM) with Seasonality

A DLM models a sequence of real-valued observations (y_t) with a sequence of real-valued hidden state vectors (x_t) of dimension d. This can be expressed as:

$$y_t = Fx_t + v, v \sim N(0, V)$$

$$x_t = Gx_{t-1} + w, w \sim N(0, W)$$

where G is a transition matrix modeling how the hidden-state changes over time, and F is an emission matrix that reflects the relation between the hidden-state and the observed values. With v and w, we also introduce stochasticity and zero-mean Gaussian noise.

5.1.1 Seasonality

To account for seasonality shifts, we decompose x_t into multiple parts:

$$x_t = (u_t, l_t, s^{([t]_p)}, s^{[t-1]_p}, ...s^{[t-p+2]_p})^T$$

where the baseline (u_t) represents the mean, the slope (l_t) represents the trend of the mean, and the seasonal component (s_t) represents the change in the mean for each day of the week. In order to map each time to its phase (day of the week), we take the modulus:

$$[t]_p = (t + p - 1) \pmod{(p+1)}$$

5.2 Multi-Process Dynamic Linear Model

To extend upon this idea, we can have multiple DLMs represent different behaviors, both normal and abnormal. Thus, given a series of observations, we want to determine which model is generating the sequence of observed values so far. Using these models, we can detect change by when we predict that a different model begins driving observations.

To formalize, we let $M_t^{(i)}$ be a random variable indicating whether model i is driving the time series at time t and generating the observation y_t. We also generate a vector M_t of $M_t^{(i)}$ for all i. Thus, the probability that model i drives the series of observations before observation y_t is $p(M_t^{(i)} = 1|Y_{t-1})$, and the probability after y_t is $p(M_t^{(i)} = 1|Y_t)$. These probabilities are used to detect change.

In terms of models, we have 3 basic models:

1. MS (stable): normal observations

2. MAO (additive outlier): specific incident occurred
3. MLS (level shift): signifies potential error in a rule

Finally, we generate a *change point score* $p(M_t^{(MLS)} = 1|Y_{t+1})$, which represents the probability of any model being in control. Figure **??** shows the application of this method to a time series of data.

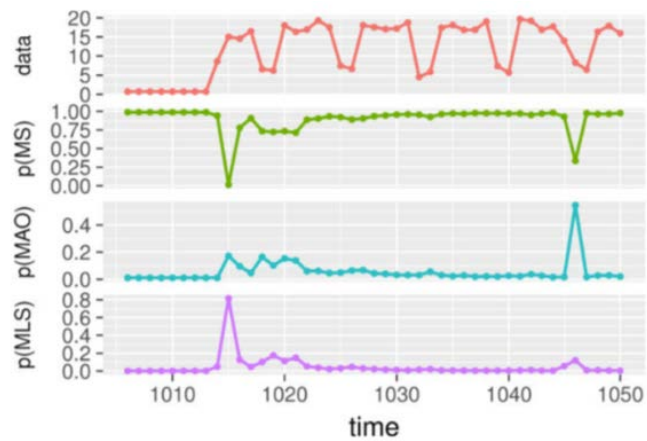

Figure 4: This figure depicts an application of the MPDLM method to a time series. Given the observations in the top graph, we can determine the posterior probabilities of the 3 different models.

5.3 Challenges & Evaluation

One of the large challenges of creating a DLM model is estimating the model parameters. This challenge arises due to a lack of labeled data, as well as a lot of potential error from trying to estimate priors for different behaviors (generally from MS) and the hyperparameters for the amount of noise.

The model was evaluated on both real data and various simulations. In terms of evaluation metrics, a delay vs. false positive rate AUC was calculated for various methods.

6 Other Workflow Issues

In addition to the previously posed points, there are other points in the workflow that are pain points or have shown potential for improvement.

6.1 Alerting

One big issue with the workflow is actually alerting clinicians or enacting change based on alerts or messages. Doctors are oftentimes swamped with a lot of tasks and information, so there is a fear of missing important messages or non-response to alerts.

As a response, a Beth Israel study in 1994 studied the impact of escalating alerts upon unread messages in the Patient Portal. For patients taking nephrotoxic or renally-excreted drugs, an alert to monitor rising creatinine levels led to a 21.6 hour reduction in reaction time. This intervention less than halved the risk of renal impairment, and on the clinician side, 44% of doctors found them helpful versus 28% of them finding them annoying. On the whole, 65% wanted the alerts continued, suggesting that this may be a viable improvement in the workflow.

6.2 The Communication Space

The communication space represents the largest part of the health system's information space, and it contains a substantial proportion of the information that we know. Various studies have shown that half of clinicians' information requests are to colleagues (Covell et al, 1985), and half of their time is spent in face-to-face communication (Safran et al, 1998). Most data is acquired and presented through communicative methods, yet it is largely ignored in our informatics thinking. This may occur due to a variety of problems, which studies have estimated as the most common cause of in-hospital disability/death (Wilson et al, 1995), as well as a contributory factor to 50% of adverse events in primary care (Bhasale et al, 1998).

Overall, clinicians seem to be faced with way too many messages, causing interruptions, inefficient fragmentation of time, as well as errors from forgetting information. As a solution, we can implement techniques such as new channels (ex. v-mail), message types (ex. alerts), policies (ex. no organization-wide emails), communication services (ex. role-based call-forwarding), and automatic agents (ex. web-bots for information retrieval). Additionally, we can consider both synchronous and asynchronous forms of communication.

6.3 Error Prevention

With the implementation of these new workflow techniques, we need to ensure that there is a framework that prevents clinicians from dropping the ball or failing to make the correct decisions. This can take the form of coordination between different parties, resource allocation, as well as the design of rational human-institution-technology systems.

We can consider the workflow as a discrete-event simulator, where we can add in coordination and notifications between different parties, as well as checks for the completion of actions before others are accomplished. A systematic diagram of this "engine" can be seen below:

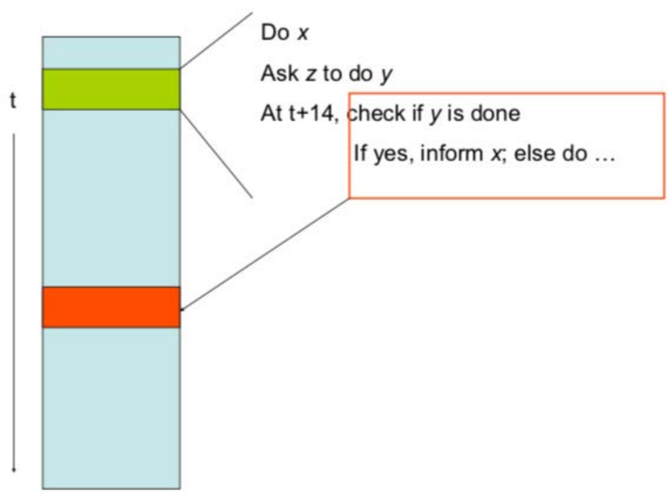

Figure 5: This figure depicts a workflow "discrete-event simulator" with various actions and agents, as well as checks to prevent errors from propogating.

6.4 Data Source Integration

The Guardian Angel Manifesto (ga.org, 1994) suggested that we can improve the healthcare system by engaging people in their health and providing them with personal methods to manage their own health. Thus, ideally, we can create a computational process of a person's life from their conception to autopsy.

Google Health (2008-2011) represented an attempt to create a concrete representation of a personal health record. To implement this, Google partnered with various healthcare providers, such as insurance agencies (ex. Anvita Health, Blue Cross Blue Shield), hospitals (ex. Beth Israel Deaconess, The Cleveland Clinic), and pharmacies (ex. CVS, Drugs.com). However, this program struggled and eventually was stopped for a few major reasons, including:

1. On the data entry side, data could only be automatically imported from Google Health partners, so many users had very sparse records or faced the challenge of manually and tediously importing data.
2. On the clinician side, there was no incentive for clincians to look at or utilize records, especially because they already had existing health records with more complete information. The lack of large-scale integration with healthcare providers and electronic health records strongly hindered the success of this product.

However, there is still hope for a system like this to succeed in the future, given the advancement of technology, lessons from products like Google Health, and a desire from the public for this type of product to succeed.

References

[AWB16] Dustin McEvoy Skye Aaron Angela Ai Jan Marie Anderson Salman Hussain Rachel Ramoni Juile Fiskio Dean F Sittig Adam Wright, Thu-Trang T Hickman and David W Bates. Analysis of clinical decision support system malfunctions: a case series and survey. *Journal of the American Medical Informatics Association*, 23(6):1068–1076, 03 2016.

[KSDS13] Jeffrey Klann, Peter Szolovits, Stephen Downs, and Gunther Schadow. Decision support from local data: Creating adaptive order menus from past clinician behavior. *Journal of biomedical informatics*, 48, 12 2013.

[ZPL14] Yiye Zhang, Rema Padman, and James E Levin. Paving the COWpath: data-driven design of pediatric order sets. *Journal of the American Medical Informatics Association*, 21(e2):e304–e311, 03 2014.