6.S897/HST.956 Machine Learning for Healthcare

Lecture 9: Translating Technology Into the Clinic

Instructors: David Sontag, Peter Szolovits

1 Introduction

There is a lot of "hype" around the possibilities of AI interfacing with our everyday lives, and the topic of AI as it relates to healthcare is one that is especially gaining popularity in this era. However, when looking at any new technological advancement, it's important to keep something called The "Hype" cycle in mind. The figure below graphically represents the life cycle stages a technology usually goes through from conception to adoption, and it's not necessarily a linear path by any means:

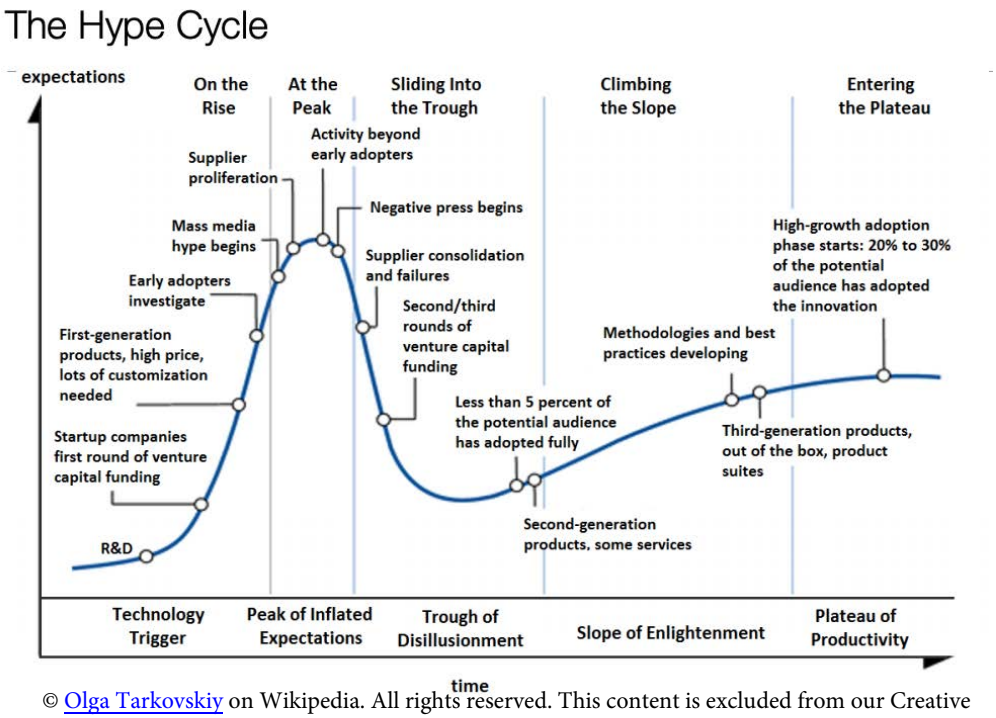

Figure 1: The Hype Cycle, detailing the life stages of technological innovation.

We see above that technology in its conception begins with RD. Startups are the "trigger", and products are adopted by early adopters. A peak is reached at a time of extreme promise, after which a steep decline occurs, called the Trough of Disillusionment. One of the best examples of this occurred in the 1980's, a phenomenon called "AI Summer", during which lots of AI technology was gaining a lot of traction among a wider audience beyond the early adopters. However, around the mid 1980's, we see the emergence of an "AI Winter", a period of time characterized by pessimism within the AI community and a decline of reduced funding and interest in AI research. But, as is the case with most adopted innovation, this technology was slowly adopted in the years that followed into products without all of the "hype".

A similar incident occurred in the early 2000's with the widespread adoption of the Internet. E-Commerce was the next "big thing", which gained a lot of hype initially before interest declined rapidly. However, we see today Amazon as the leading giant of E-Commerce today. The AI in healthcare industry follows and is following a similar pattern, as we discuss today.

2 Existing Healthcare Technology

We can study the effects of the "hype" cycle on two different types of ventures undertaken in the field of Healthcare and AI, one successful and one a failure, and learn how the rise and fall of this technology is influenced by the various factors in its life-cycle.

2.1 A Cautionary Tale

One of the best examples of the difficulties of an endeavor in the healthcare industry is the story of IBM's Watson, which the company itself coined as "a revolution in cancer care". This is a good example of the pitfalls of media marketing over-promising results with no clear support or backing.

Watson for Oncology was a cloud-based supercomputer designed to digest massive amounts of healthcare data – from doctors notes to medical studies to clinical guidelines. The company claimed that given a patients record and history, the technology would be able to "solve the problem of treatment of cancer diagnosis" in providing recommendations for cancer treatment. However, the technology had a great deal of difficulty trying to solve this problem:

- The technology wasn't able to live up to the lofty expectations that IBM created. It struggled with the basic step of learning about different forms of cancer, which is by no means a simple task. The marketing department had over promised the potential outcomes of this technology, and attempting to perform such a large task as this one was seen as a disappointment when it did not succeed. The lack of data as well as the difficulty of the task proved to be a great hurdle for IBM.
- The advice was criticized by foreign doctors to be biased to its data's American patient population and American methods of healthcare.
- The treatment recommendations were largely based on human input, and the doctors were even recommended to **give their own recommendations**
- **There were no publications: IBM had not exposed the product to critical review by outside scientists or conducted clinical trials to assess its effectiveness**

IBM had an alliance with MD Anderson, but this also proved to cause many problems for Watson, including difficulty teaching the machine to read medical records as well as deploying the system in clinical practice, which lead to many complications in dealing with legal and compliance issues. The biggest obstacle, though, was of data: even if Watson was able to make immense advancements in the making treatment recommendations, how is it possible to obtain enough data to make significant advancements in the field of healthcare?

2.2 A More Successful Tale

A more promising tale of a healthtech endeavor is one of Computerized Physician Order Entry (CPOE). This was less driven by marketing and more driven by medical data, which essentially had a higher guarantee of success in terms of accuracy and adoption. CPOE reduced error rates by 55%, and rates of serious medication errors fell by 88%. It actually had higher rates of accuracy in terms of its usage and provided many benefits to physicians:

- Prompts that warned against drug interactions, allergy, and overdose

- Information to help physicians keep up with newer drugs as they are introduced to the market
- Drug specific information to eliminate confusion between drugs that sound alike
- Improved communication between physicians and pharmacists
- Reduced long-term healthcare costs

With the ever-changing face of AI in healthcare, there are also numerous other potential future benefits, such as using machine learning to learn more about drug-drug interactions as well as identifying occurences in patient notes and reports. Due to its big success in terms of accuracy, CPOE was recommended by the National Academy of Medicine by 1999 for universal adoption.

However, there were some adverse effects due to the widespread adoption of this technology. In both the short-term and long-term-CPOE sites, pharmacists began to spend more time on distributive tasks and less on clinical tasks, which places new burdens and challenges on the pharmacists. As noted in the presentation, a study found that physicians were spending over twice the amount of time on EHR and desk work than on direct clinical face time after CPOE adoption. This is one of the negative effects of rapid helathcare technology adoption, and it is something to keep in mind when moving forward with such innovation.

3 Diffusion of New Medical Technologies

There are different rates of adoption for different technologies for a variety of external reasons. Selected technologies in the graphic below generally had much different diffusion curves:. We see in this graphic that

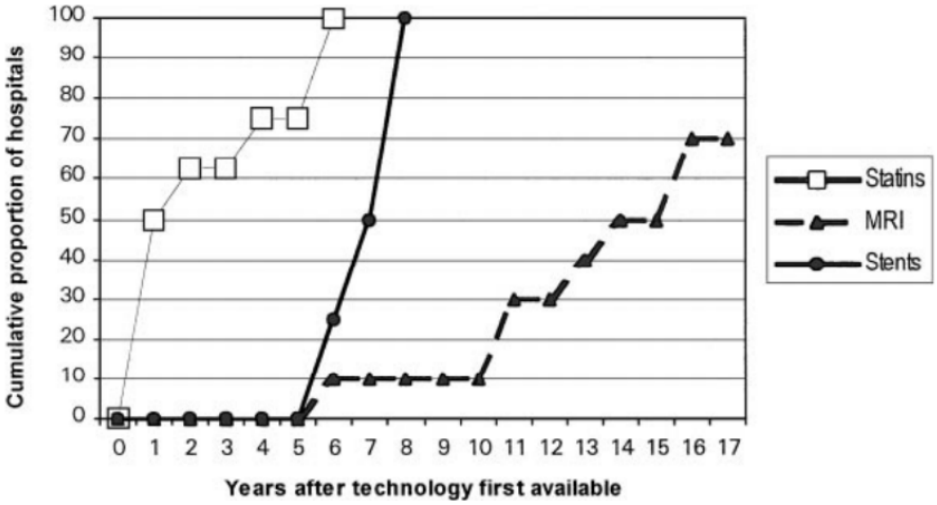

Figure 2: Diffusion Curves of Selected Technologies

Statins, a class of drugs often prescribed by doctors to help lower cholesterol levels in the blood, diffuse soon after launch, possibly due to its medical success and commercial marketing. Coronory stents, which are tubes inserted into a arteries to keep passageways open in the case of blockages, were also adopted quickly, but at a slower pace than statins. However, MRI scanners had a substantially slow rate of diffusion, initially purchased six years after the technology became available and still had a slow rise after that. The primary reason for this was cost, which also impacted the diffusion of stents. However, enthusiastic individuals were key in the diffusion of stents. This graph shows that even though various technological successes might enter

the market in the field of healthcare, there are a variety of factors that ultimately decide how quickly it might be adopted into the market, or whether it is to be adopted at all. There is a certain drug that is recommended for victims of heart attack and cornary artery disease that reduces the probability of a heart attack by 35% but has not seen high diffusion rates to due various other environmental factors.

4 Biomedical Research

Research itself proves to be a complicated area of healthcare advancements. In this section we examine various problems associated with biomedical research and ways that these could be prevented to lead to higher rates of advancements of research in this area.

4.1 The Problems with Biomedical Research

One of the major problems with biomedical resarch is that a large amount of bias exists in published reports. Generally, only successful studies get published, which leads to many problems related to reproducability. Testing by groups unknown to each other only yield some positive results (there are not just slight differences; statistical significance may vary greatly), and there is a well-known inability to replicate a large portion of these studies due to these reasons.

4.2 Quality Assurance: Suggestions for Research

These problems and more suggest that better precautions must be taken in approaching biomedical research. There are various suggestions that might ensure quality in terms of research. One such suggestion is to require the replication of studies in more than one data set to avoid bias and allow for models that may be able to generalize better. This might also fight outliers that might lead to the results that lead to rapid publications.

Another suggestion is to make publicly available the studies that may have failed. Registration of every study might also us to learn whether the successful studies do yield significant results or whether they are just statistical outliers. Various studies use such different methods of data mining, preprocessing, and applications of various learning algorithms, so even knowledge of this might be helpful in designing future studies. The sole publication of interesting results might not lend itself to valuable information that may be lost in the failure of various other forms of biomedical research.

Other suggestions for biomedical research include practicing detailed analysis and caution, where a hand-off technique might be implemented such that different groups from developers could evaluate the results of a study. However, too much caution might not allow for fast development, which is another important factor to consider. Other suggestions include:

- Regression testing as programs are improved, so that old code can be run
- Automated search for inconsistencies
- Retrospective review, judged by clinicians, as circumstances might differ between real-time and retrospective views (i.e. mechanisms of data collection)
- Prospective review, judged by clinicians; evaluate the answer as well as the effect on health outcomes (i.e. when the decision support system is run, do clinicians change their behaviors, and do patients get a better outcome?
- Compare to the answers of unaided doctors

5 FDA Approval

FDA clearance for class III medical devices have large amounts of regulatory control. They are regulated across five major domains: product quality, patient safety, clinical responsibility, cybersecurity responsibility, and proactive culture. Approved technologies are generally imaging interpretation applications. Various approved technologies can be found in the corresponding class presentation.

6 Guest Speaker

We hear from guest speaker Adam Wright, PhD. He is a research scientist at Brigham and Women's Hospital, and today he talks speaks to us about the development of a medication related decision support system innovated by his group. The technology performs actions such as: reminders for patients that are overdue for certain procedures; make recommendations to doctors regarding patients; as well as some predictive models specifically dealing with hospital readmission and hospital care (such as ML tools for falls and risk avoidance in terms of environmental – often hospital bed – support for patients).

6.1 Q & A

Q [instructor]: How does the risk analysis system work?
A: It is a utility based model for cost. It does a full utility analysis in terms of the cost of falling and its probability to figure out where on the ROC curve it may fit in order to determine the best course of action in terms of patient care. However, it is often very difficult to weight these things accordingly. Different doctors make different decisions, so decision support system is hard enough if the humans can't even decide what to do.

Q [instructor]: I was concerned a few years ago when visiting a hospital that the staff interacting with the computer, a resident, did not even know how to interface with it. Furthermore, those interacting with these technologies are often on the junior level, and a junior resident must be bold to go against the recommendations of a doctor. How to deal with that?
A: Residents are scared, yes, and there are a lot of other challenges as well. Alerts sometimes happen to often and people start to dismiss them. It is our responsibility to improve these interactions as well.

Q [instructor]: How far are we from using ML to do this?
A: The biggest challenge is accuracy. Our models can't take into account things like end-of-life scenarios, or extreme measures. There is also a lack of reliable documented coded usable features. Second, the flow of patient care is also extremely hard to determine, as in who the primary provider of the patient actually is, so who might actually use this feature. However, the technical ability to help with these models is much better than before.

Q [instructor]: 20 years ago there were attempts to create formal rules for this. A Tufts doctor was alerted on every abnormal characteristic of the patient. In other hospitals, there might be more sophistication for the alert system: if there is no response, escalate the response to a higher up. Are there things in place such as this/should there be?
A: There are differences between in-patient and out-patient automated messages. There is also great incentive to do this, as pages will always keep going unless a response is given. However, out-patient is difficult as doctors are out for a lot of the week.

Q [instructor]: There are people in clinical labs in MassGen. However there may be potential problems: the lab systems are ancient; they may or may not have the relevant experience; there are legacy systems in place, which might lead to horrendous problems for building, and even risk prediction. How can we incorporate this technology into old technology?

A: There are many architectures, and there are many pros and cons. We don't favor using creaky old tech or new but we often try to build in the rules within the system, right into the workflow. If the model is too complex, we run the model in warehouse at midnight, extract the data, write the risk score back into the patient records, which does not depend on real-time data. However, the problem with this is that a lot of tasks depend on real-time data, so this only works for retrospective data. Messages are sent in an encoded format, and infrastructure is set up outside. Messages are sent from the EHR to model Webservice for EHR. We also try to embed the app into EHR. However, there are heavy fees and regulation to limite the rate people can charge app developers for AI Access, as recently released by the Fed.

ASIDE: I am optimistic. Even 5 years ago, systems were essentially locked out, and people would have to charge billions of dollars to build the appropriate software for you. However, now there are systematic ways to embed code into existing systems or well-documented ways to feed data out, process it, and feed it back into the system.

Q [student]: Have you thought of putting this on something like app orchard?
A: We haven't decided if we want to sell it yet – we want to first give away apps for free.

Q [instructor]: What are biggest challenges?
A: The biggest battle is that there is less complete data. Most EHR's don't implement write-back capabilities. Growth curve is ok, but medication ordering is challenging.

Q [instructor]: Are you worried about cognitive overload (i.e. overalerting)? What about liability? If the system's recommendations are used and they are wrong, whose fault is it?
A: Yes we are worried about both. The liability problem: we can correct data in the database, and we can add bypass checks in the data, for potentially wrong or dangerous orders. There are concerns about protection of data in light of scandals (i.e. Cambridge Analytica). We can have patients authorize these medical records, but what liability does the vendor have for this? Do you block access? How do you sort this out?

Q [instructor]: How does a third party vendor deploy a ML model onto the system? On epic or app orchard? Is there a way to go around directly to the partner?
A: Epic: there is always a relationship between the vendor of app and healthcare org, we can work together directly that they want to use, share in 1) predictive modeling markup language, PMM model, imported natively 2) web service for answer. There are limitations of what we can share about epic or epic model, facilitated by joining this program for documentation etc? People using this program sometimes thinnk they can run their app always, but they must make deals with all parties in involved; Other EHR – centralized model, to develop app.

Q [instructor]: What if we don't have to talk to epic at all?
A: Epic won't stop you, it will limit the ability tot tell about how to map between the dat model to internal data model. Joining programs such as app orchard and others enable revenue sharing, etc.

Q [instructor]: On the side of quality, what if you develop an app and deploy. What if it just fits to the patient population in Boston and is not applicable elsewhere?
A: Generalization is a huge problem. Generally, many models transfer pretty well. Things such as poor human physiology might lead to the need for a better workflow, generate the model retreospectively and see if it is accurate and then retain.

Q [instructor]: Have you imported such models?
A: Yes 5 or 6 models.

Q [TA]: A bit more about whether the models are actually working: if the score agrees, then trust, but if

not, then what? Is the risk score deployed?
A: There are many levels of evaluation. What do people do when they are shown this score and recommendation? Based on this score, do people actually do what they say? If not, these are also clinical trials, so there may be an intent to remove something. We try to measure the rates of before and after; we do randomized trials on alerts vs no alerts and compare clinical outcomes. Unless you can show clinical outcomes, dont bother to deploy.

Q [TA]: Are there internal checks?
A: Yes we are also behind

Q [student]: Since the models are mainly recommendations, are there any metrics on how often the model's recommendation matches the doctor's actions?
A: We retrospectively test the model, see if the recommendations match the doctors actions. We can run it in silent mode and see whether doctor does what it suggests; we can turn it on and see if the doctor take this action or not. In Randomized mode: does doctor take action when shown the alert or not.

Q [student]: When it comes to a particular patient, how do you know how reliable this model is?
A: We can have calibration for certain patients depending on demographics, etc. We analyze inference from a model but value judgments, beliefs and knowledge about patient circumstances that are human so thats why the suggestions are delivered to a doctor or a nurse so that they can use experience and relationship and expertise to make a suggestion.

Q [student]: Will we eventually not need a human?
A: These are few cases, not the norm. We still need human doctors.

Q [student]: When you build a model and it becomes stale, how do you update it?
A: There is high desire to revisit. They are updated, often retrained 6 months to a yr, to keep up with new knowledge and technology to keep reliability and validity. They can still make mistakes due to new medicines, changed codes, etc Comment: some things are done in real-time - what bugs are growing in a microbiology report, always changing so systems are always updating in real-time.

That's all we have for today! Thank you Adam! [CLAPS]

References

[BSOM+14] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[WSW14] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019