Paper ML04

Accelerating Translational Research by Leveraging AI and Machine Learning

Chris Asakiewicz[A], Frank Corvino[B], Marko Zivkovic[B], Rong Liu[A]

[A] Stevens Institute of Technology, School of Business, Castle Point on Hudson, Hoboken, NJ 07030, United States
[B] Genesis Research, 5 Marine View Plaza, Suite 312, Hoboken, NJ 07030, United States

Abstract
The web has revolutionized the conduct of global, cross-disciplinary research. In the life sciences, interdisciplinary approaches to problem solving and collaboration are becoming increasingly important in facilitating knowledge discovery and integration(1). The big data revolution, accompanied by the development and deployment of wearable medical devices and mobile health applications (2), has enabled the biomedical community to apply artificial intelligence (AI) and machine learning (ML) algorithms to vast amounts of data.

AI and ML is fundamentally changing healthcare, creating new research opportunities in predictive analytics, precision medicine, virtual diagnosis, patient monitoring, and drug discovery and delivery(3). This paper will discuss how AI and ML can be leveraged to derive meaningful insight from RWD and RWE(4) and in the process enable strategies that close critical gaps in knowledge, give physicians and patients broader access to therapies, and help payers realize the actual value of those therapies in improving health and reducing costs. One example of this will be a discussion of a research project using NLP techniques for data identification and extraction from pooled research resources.

Also discussed will be the future impact of AI and Machine Learning in accelerating translational research by helping to derive meaningful insight from RWD and RWE and in the process enable strategies that close critical gaps in knowledge, give physicians and patients broader access to therapies, and help payers realize the actual value of those therapies in improving health and reducing costs.

Introduction
What exactly is Translational Research? For many, the term refers to the ‘bench-to-bedside’ enterprise of harnessing knowledge from the basic sciences to produce new drugs, diagnostics and treatment options for patients. Translational Research fosters the multidirectional and multidisciplinary integration of basic research, patient-oriented research, and population-based research, with the long-term aim of improving the health of the public(5).

As can be seen in Figure 1, over the years as technology and healthcare has advanced, Translational Research has evolved. Translational Research 1.0 focused on taking discoveries from basic research, demonstrating their safety and effectiveness in clinical trials and then applying then, as part of clinical practice, in the delivery of healthcare. Translational Research 2.0 enhanced the process of leveraging healthcare data from all ecosystem stakeholders in order to facilitate both the discovery and translation, as well as the health outcomes associated with new products and therapies. The next stage in the evolution will be Translational Research 3.0 which leverages AI (Artificial Intelligence) and ML (Machine Learning) to derive meaningful insight from RWD (Real World Data) and RWE (Real World Evidence) and in the process enable strategies that close critical gaps in knowledge, give physicians and patients broader access to therapies, and help payers realize the actual value of those therapies in improving health and reducing costs.

However, with any evolution there are roadblocks to overcome.

Overcoming the roadblocks to translational research involves taking a fresh look at how knowledge is discovered across the ecosystem and how it is integrated at the organizational intersections.

Understanding and leveraging a core framework of standards, tools and processes will assist ecosystem participants in overcoming the barriers to translational research and enhance their ability to facilitate better evidence-based decisions across the ecosystem. Web technology will facilitate knowledge discovery and integration, as well as enable reproducibility, aid in collaboration, and accelerate and transform medical and healthcare research.

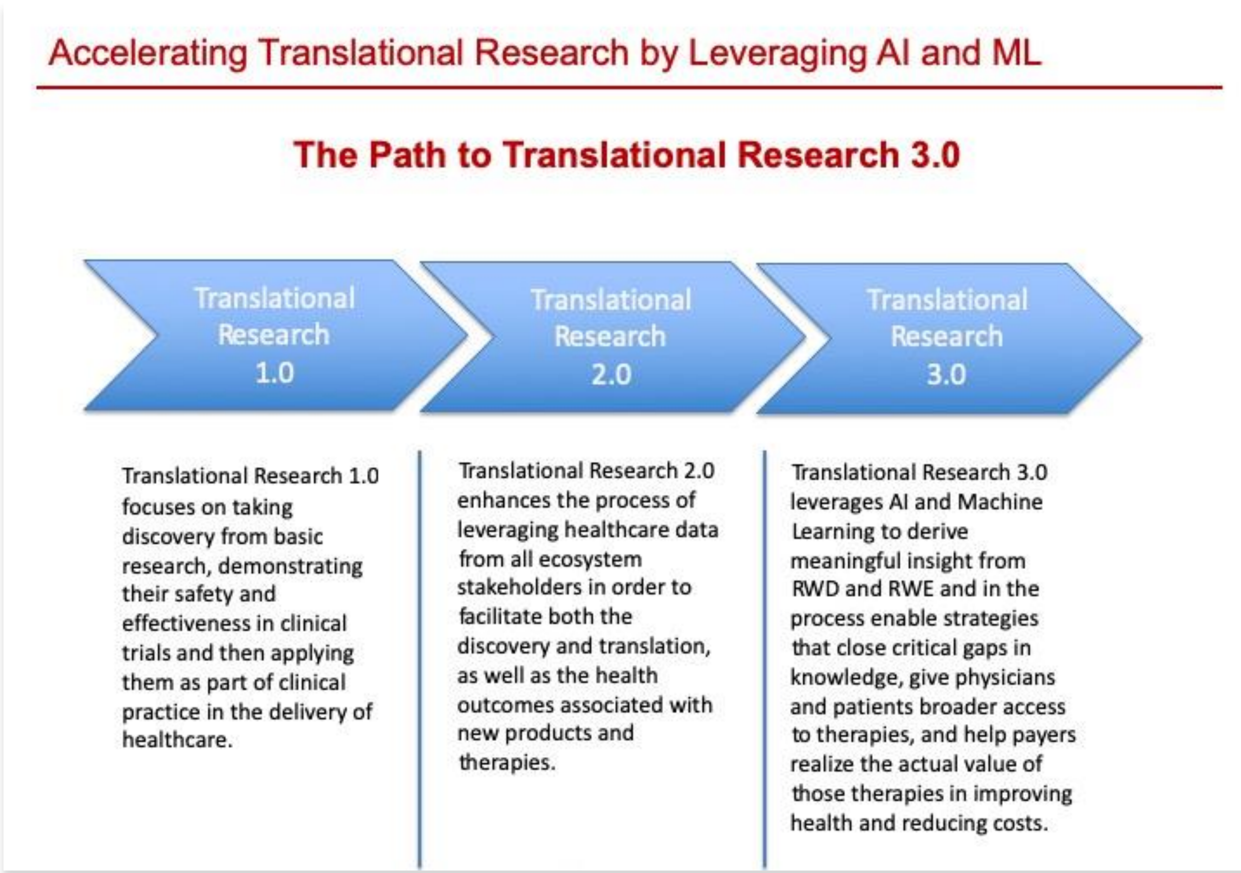

Figure 1

Accelerating medical and health research involves the integration of knowledge across ecosystem boundaries and the leveraging of data, analytics and processes to facilitate collaboration and discovery. Knowledge discovery can be enhanced through the use of web-based tools that facilitate the process of knowledge acquisition, representation, inference and explanation. Knowledge mining can facilitate the synthesis and visualization of ecosystem knowledge and when coupled with web-based tools, aid in the collaborative discovery process.

Healthcare and Life Sciences
The availability of large quantities of clinical, biological, and healthcare data promises unprecedented advances for the diagnosis and treatment of disease as well as the improvement of patient health; however, the appropriate use of this data mandates changes in the traditional paradigms of data management, analysis, and interpretation.

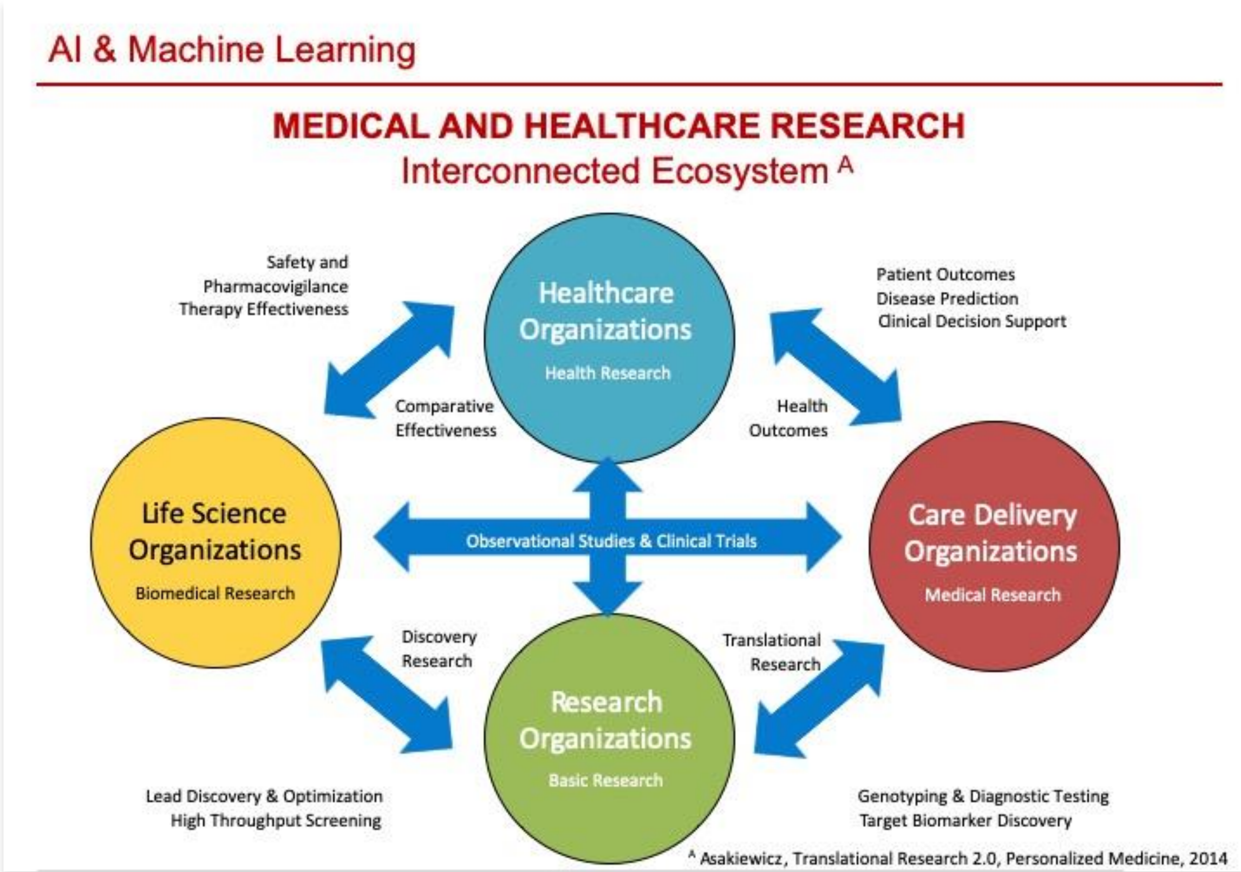

Figure 2

In the life sciences, interdisciplinary approaches to problem solving and collaboration are becoming increasingly important in facilitating knowledge discovery and integration. Ai and ML technologies promise to have a profound impact – enabling reproducibility, aiding in discovery, and accelerating and transforming medical and healthcare research across the healthcare ecosystem. However, knowledge integration and discovery require a consistent foundation upon which to operate. A foundation should be capable of addressing some of the critical issues associated with how research is conducted within the ecosystem today and how it should be conducted for the future. Figure 2 highlights a framework for enhancing collaborative knowledge discovery across the medical and healthcare research ecosystem. This framework that could serve as a foundation upon which ecosystem stakeholders can enhance the way data, information and knowledge is created, shared and used to accelerate the translation of knowledge from one area of the ecosystem to another.

The Road to Translational Research 3.0
Artificial intelligence (AI) is bringing a paradigm shift to healthcare, powered by increasing availability of healthcare data and rapid progress of analytics techniques. AI can be applied to various types of healthcare data, both structured and unstructured. With the increasing availability of healthcare data and

the rapid development of big data analytic methods, powerful AI techniques can unlock clinically relevant information hidden in massive amounts of data, which in turn can assist clinical decision making. Popular AI techniques include machine learning methods for structured data, such as the classical support vector machine and neural network, and the modern deep learning, as well as natural language processing for unstructured data. Major disease areas that use AI tools include cancer, neurology and cardiology.

AI can use sophisticated algorithms to ‘learn’ features from a large volume of healthcare data, and then use the obtained insights to assist clinical practice. It can also be equipped with learning and self-correcting abilities to improve its accuracy based on feedback. An AI system can assist physicians by providing up-to-date medical information from journals, textbooks and clinical practices to inform proper patient care. In addition, an AI system can help to reduce diagnostic and therapeutic errors that are inevitable in the human clinical practice. Moreover, an AI system extracts useful information from a large patient population to assist making real-time inferences for health risk alert and health outcome prediction.

In developing its Real World Evidence (RWE) program, the FDA believes it is helpful to distinguish between the sources of RWD and the evidence derived from that data(6). Evaluating RWE in the context of regulatory decision-making depends not only on the evaluation of the methodologies used to generate the evidence but also on the reliability and relevance of the underlying RWD. For the purposes of the framework, FDA defines RWD and RWE as follows:

- Real-World Data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.
- Real-World Evidence (RWE) is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

Examples of RWD include data derived from electronic health records (EHRs); medical claims and billing data; data from product and disease registries; patient-generated data, including from in-home-use settings; and data gathered from other sources that can inform on health status, such as mobile devices. RWD sources (e.g., registries, collections of EHRs, administrative and medical claims databases) can be used for data collection and, in certain cases, to develop analysis infrastructure to support many types of study designs to develop RWE, including, but not limited to, randomized trials (e.g., large simple trials, pragmatic clinical trials) and observational studies (prospective or retrospective).

Current RWD and RWE-informed efforts include:

- Identifying new drug targets and better biomarkers
- Finding new uses for existing and late-stage products
- Identifying undiagnosed, under-diagnosed and misdiagnosed illness
- Identifying sites with higher enrollment probability and capability to aid clinical trial recruitment
- Predicting patients who have higher risk for adverse events
- Predicting medication compliance (or lack thereof) in defined patient cohorts (which can inform mitigation efforts)

Using AI and machine-learning techniques to recognize trends and patterns in social media and healthcare sensor data (to inform clinical study development and define study objectives, patient-relevant endpoints and study procedures)

What does this mean for pharmaceutical and biotech companies. The technology advancements of AI, ML and NLP, as seen in the above efforts, offer pharmaceutical and biotech companies the power to increase meaningful RWE output, decrease time to insights, and make the most of the vast data sources now available. An RWE technology platform that delivers smart data processing, analysis and outcomes offers an unparalleled opportunity to capitalize on these computing advancements. When used as part of an overall comprehensive RWE strategy, AI, ML, and NLP innovations can enhance drug development, improve patient treatment and access, and drive valuable new business opportunities.

Although the AI and ML technologies are attracting substantial attentions in medical research, the real-life implementation is still facing obstacles(7). The first hurdle comes from the regulations. Current regulations lack of standards to assess the safety and efficacy of AI systems. To overcome the difficulty, the US FDA made the first attempt to provide guidance for assessing AI systems (e.g., guidance clarifies the rules for the adaptive design in clinical trials).

The second hurdle is data exchange. In order to work well, AI systems need to be trained (continuously) by data from clinical studies. However, once an AI system gets deployed after initial training with historical data, continuation of the data supply becomes a crucial issue for further development and improvement of the system. Current healthcare environment does not provide incentives for sharing data on the system. Nevertheless, a healthcare revolution is under way to stimulate data sharing in the USA. The reform starts with changing the health service payment scheme. Many payers, mostly insurance companies, have shifted from rewarding the physicians by shifting the treatment volume to the treatment outcome. Furthermore, the payers also reimburse for a medication or a treatment procedure by its efficiency. Under this new environment, all the parties in the healthcare system, the physicians, the pharmaceutical companies and the patients, have greater incentives to compile and exchange information.

Opportunities – Data Science (AI, ML, and NLP)

Artificial intelligence (AI) is bringing about a paradigm shift to healthcare. Powerful AI techniques can assist in clinical decision making by unlocking clinically relevant information hidden in massive amounts of data. Popular AI techniques include machine learning methods for structured data, such as the classical support vector machines and deep neural networks, as well as natural language processing for unstructured data. An AI system can assist physicians by providing up-to-date medical information from journals, textbooks and clinical practices to inform proper patient care, and help to reduce diagnostic and therapeutic errors that are inevitable in the human clinical practice. AI and Machine Learning can be leveraged to derive meaningful insight from RWD and RWE and in the process enable strategies that close critical gaps in knowledge, give physicians and patients broader access to therapies, and help payers realize the actual value of those therapies in improving health and reducing costs.

Large volumes of textual data pose considerable challenges for manual qualitative analysis. We explore semi-automatic coding of textual data by leveraging Natural Language Processing (NLP). We compare the performance of human-developed NLP rules to those inferred by machine learning (ML) algorithms. The results suggest that NLP with ML may be useful to support researchers coding qualitative data. Researchers often apply qualitative research methods to analyze the work practices of groups. For example, researchers might examine transcripts of a group's discussions to understand how it solved some task. Because group artifacts are typically textual, they can require considerable manual effort to analyze, as researchers read and reread them to locate evidence to support or refute their theories, tagging specific passages in the text as evidence for the various concepts of interest. This analysis process is referred to as content analysis, more specifically, as coding. We discuss the use of natural language processing (NLP) technology for coding qualitative data for research as an example of how NLP can be leveraged in the analysis of RWE.

Industry-Academic Partnership

Research and education at the School of Business at Stevens Institute of Technology reflects Stevens' legacy of leadership in developing practical solutions to the most complex problems facing industry and society. The Stevens Alliance for Innovation and Leadership (S.A.I.L.) brings its wealth of thought leadership in technology, innovation, data and analytics to bear for industry professionals seeking competitive advantages and practical insights in a world of constant change. One example of which is S.A.I.L.'s partnership with Genesis Research in exploring the use of AI and Machine Learning (ML) with Real World Evidence, specifically by using NLP techniques for data identification and extraction from pooled research resources.

Using NLP Techniques for Data Identification and Extraction from Pooled Research Resources
Forming a strong study foundation by understanding the current state of a research field is necessary for successful attainment of an evidence-based answer to a narrowly defined research question. A literature review is an essential and effective method used to identify, collate, and consolidate existing research on a specific topic. Literature reviews are, in general, a labor-intensive process, with the data extraction part of the review process being highly time-intensive and requiring considerable human resources. According to a recent analysis of 195 reviews registered in the International Prospective Register of Systematic Reviews (PROSPERO), the mean time to complete a review was 67.3 weeks. Automating parts of the literature review process would greatly shorten the time needed for completion of a review, reduce the burden on human resources, increase overall efficiency of capturing the totality of evidence, and help save more time for result interpretation and quality assessment.

We have developed a deep learning framework to facilitate data extraction after the articles in a literature review are identified. This framework utilizes NLP techniques to automatically pre-process articles, extract salient features through a convolutional neural network, and identify text in target topics. In particular, to alleviate the burden of manually labeling training samples, this framework leverages few-shot learning techniques to recognize the differences in features representing target topics, with only a small number of annotated samples. We are experimenting this framework with a structured literature review of around 500 articles (out of thousands screened) about a rare, chronic, genetic (autosomal recessive) and potentially life-threatening liver disorder of impaired copper transport.

Literature Review
A critical analysis of a segment of a published body of knowledge through summary, classification, and comparison of prior research studies, reviews of literature, and theoretical articles(8).

Forming a strong study foundation by understanding the current state of a research field is necessary for successful attainment of an evidence-based answer to a narrowly defined research question. A literature review is an essential and effective method used to identify, collate, and consolidate existing research on a specific topic.

- High quality systematic reviews require two independent researchers to screen studies and extract data(9)
- There is a high prevalence error in data extraction for literature reviews and meta-analyses(10-11)

Literature reviews are, in general, a labor-intensive process, with the data extraction part of the review process being highly time-intensive and requiring considerable human resources. According to a recent analysis of 195 reviews registered in the International Prospective Register of Systematic Reviews (PROSPERO), the mean time to complete a review was 67.3 weeks.

Automating parts of the literature review process would greatly shorten the time needed for completion of a review, reduce the burden on human resources, increase overall efficiency of capturing the totality of evidence, and help save more time for result interpretation and quality assessment.

To date there are **no comprehensive, suitable and intuitive solutions** for automated data extraction of healthcare-related information from pooled research resources.

Background of Few Shot Learning
Deep learning models usually require sufficient high-quality data to achieve good performance. However, it is expensive, and sometimes unrealistic, to obtain sufficient data. Few-shot learning discovers patterns in such challenging scenarios where data is limited. Each class has a **prototype representation**, which is learned from labeled samples. Train a neural network to **differentiate** the prototype representations of classes

Given **a test sample** and **support sets** (i.e. labeled samples)
- Use a **convolutional neural network** (CNN) to extract features (i.e. words or phrases)
- Calculate the **distance** between the test sample and each support set

- Compute the **weight** of each sample in support sets based on the distance
- Produce the prediction as the sum of weights in each class

Objective: Identify sentences containing aspects of interest (i.e. **classes**) in research literature

Perspective	Study Design	Data Source	…
Country	Study Period	Sample Size	…

Preprocessing: each sentence in a research article is tokenized into words, and words are mapped to pre-trained Glove word vectors in order to capture meaning. Classes are **not mutually exclusive**. A sentence may describe multiple aspects.

One-vs-the-Rest approach: a binary model with two classes: Positive class: one target class (e.g. study design) and Negative class: all the rest classes. So our model is **2-way k-shot learning**, where ***k*** will be tuned, ***k*** = 1, 2, … Each training sample is a 3-element tuple: (support set 1, support set 2, test sample).

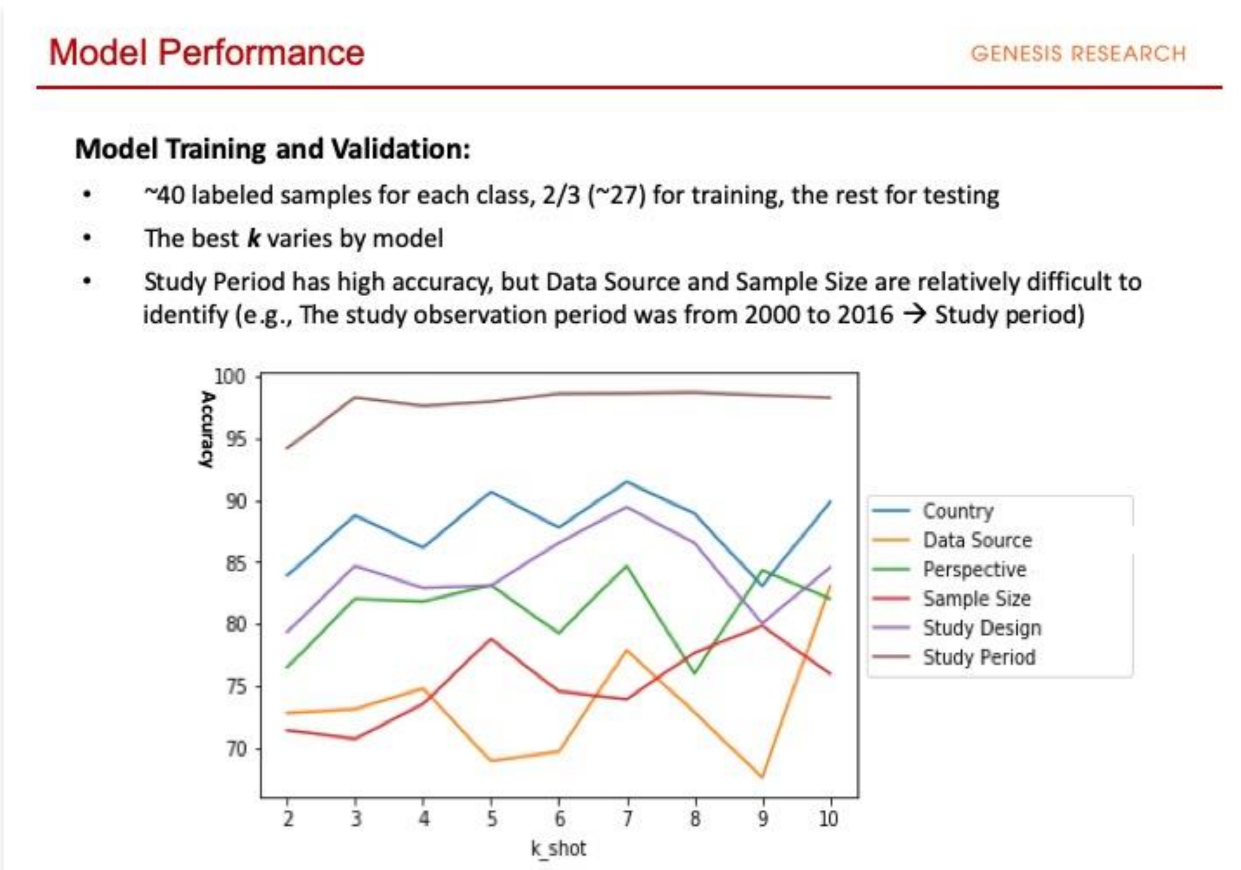

Figure 3

Model Performance is highlighted in Figure 3. Overall performance is reasonably good considering small sample size. The model focused on capturing semantics instead of relying of keyword matching. However, there are a few challenges:

- **Overfitting** due to small sample set

- A sentence with **unseen patterns** can be easily misidentified, due to equal distance to both support sets (e.g. Poster presentations s630 journal of hepatology 2018 vol. → mislabeled as Study Design, Data Source)
- A sentence alone may not have the **right context** to make prediction (e.g. Thirty four patients (32%) were transplanted at mean age of years. → labeled as Sample size; however, we need to identify the total WD sample size)

Future Work

Future efforts will emphasize improvements on the model (e.g., continuously testing the model with a wide variety of articles) as well as focusing on more regularization to overcome overfitting, dealing with unseen patterns, as well as, incorporating broader context of each sentence into learning. Finally focus on transfer learning by applying the model to different literature domains.

AI and ML for Enhanced Data Analysis and Decision Making

In addition to the above example, at Stevens our faculty and students are actively pursuing research projects in the area of leveraging AI and ML for enhanced data analysis and decision making in the following areas relevant to healthcare, namely:

> **Smart Health Care -** AI is able to make sense of vast amounts of medical history and genetic data and provide insights that can be used to diagnose patients, determine the appropriate course of treatment and monitor that treatment.
>
> **Diagnosis and Treatment -** Beyond simply retrofitting the AI technology that already exists, developments are also underway to fine tune it to meet niche health care demands.
>
> **Detection and Prevention -** Prevention is the best medicine and in this vein, wearable medical devices equipped with AI to detect and predict potentially life-threatening events.
>
> **Smart Health Care -** AI is able to make sense of vast amounts of medical history and genetic data and provide insights that can be used to diagnose patients, determine the appropriate course of treatment and monitor that treatment.

Conclusions

In the life sciences, interdisciplinary approaches to problem solving and collaboration are becoming increasingly important in facilitating knowledge discovery and integration. The big data revolution, accompanied by the development and deployment of wearable medical devices and mobile health applications, has enabled the biomedical community to apply artificial intelligence (AI) and machine learning (ML) algorithms to vast amounts of data.

AI and ML is fundamentally changing healthcare, creating new research opportunities in predictive analytics, precision medicine, virtual diagnosis, patient monitoring, and drug discovery and delivery. AI and Machine Learning is accelerating translational research by helping to derive meaningful insight from RWD and RWE and in the process enable strategies that close critical gaps in knowledge and giving physicians and patients broader access to therapies.

References

1. Asakiewicz, C., Translational Research 2.0 – A Framework for Accelerating Collaborative Discovery, May 2014, Journal of Personalized Medicine, Volume 11, Number 3, pp 351-358.
2. Dimitrov, D., Medical Internet of Things and Big Data in Healthcare, Healthcare Informatics Research, (2016)3:156-163, http://dx.doi.org/10.4258/hir.2016.22.3.156.
3. Shah, P., et.al., Artificial Intelligence and Machine Learning in Clinical Development: A Translational Perspective, npj Digital Medicine (2019)2:69; https://doi.org/10.1038/s41746-019-0148-3.

4. Artificial Intelligence and Machine Learning to Accelerate Translational Research: Proceedings of a Workshop—in Brief, National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Government-University-Industry Research Roundtable, Washington (DC): National Academies Press (US); 2018 Jul. available at: https://www.nap.edu/catalog/25197/artificial-intelligence-and-machine-learning-to-accelerate-translational-research-proceedings
5. Rubio DM, Schoenbaum EE, Lee LS, et al. Defining translational research: implications for training. *Acad Med*. 2010;85(3):470–475. doi:10.1097/ACM.0b013e3181ccd618
6. FDA RWD and RWE 2019 Guidance at: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
7. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–243. Published 2017 Jun 21. doi:10.1136/svn-2017-000101
8. Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.
9. Borah R, Brown AW, Capers PL, et al Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry BMJ Open 2017;7:e012545. doi: 10.1136/bmjopen-2016-012545
10. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005 Jul;58(7):741-2. Epub 2005 Apr 18. PubMed PMID: 15939227.
11. Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA. 2007 Jul 25;298(4):430-7. PubMed PMID: 17652297.

Contributors

Chris Asakiewicz, PhD, is currently an Industry Professor and Program Director at Stevens Institute of Technology's School of Business. He is also the Director of the School's Industry/Academic Partnership for Innovation and Leadership. Chris is a senior management professional with a strategic background in Business Technology Management, Information Technology, Consulting and Education with a 21-year history as a Vice President of Global Business Technology at Pfizer. Chris has done pioneering work in the areas of Knowledge Mining and its application within Life Sciences, most notably as a means of accelerating Translational Research. (Email: casakiew@stevens.edu)

Frank Corvino, PhD, is a Managing Partner and the Chief Operating Officer at Genesis Research where he leads the Real Word Evidence and Advanced Analytics group. Dr. Corvino specializes in real-word evidence generation, analysis, and communication and, together with his team, he developed and disseminated value evidence for scientific communications, market access, business development evaluations, regulatory support and C-suite strategic and innovative solutions across a portfolio of therapeutic. (Email: frank@genesisrg.com)

Marko Zivkovic, PhD, is the Chief Scientific Officer at Genesis Research. He has over 10 years of experience in modeling, epidemiology and database analytics. Dr. Zivkovic manages the Digital group at Genesis and plays a pivotal role in the development of technology systems that support and leverage data analytics. Dr. Zivkovic has a particular interest in development of predictive models and machine learning utilization designed to facilitate the patient prioritization and clinical insights critical to healthcare management. (Email: Marko@GenesisRG.com)

Rong Liu, PhD, is an Associate Professor of Information Systems at School of Business, Stevens Institute of Technology. Before joining Stevens, she was a Research Staff Member at IBM T. J. Watson Research Center. She received her Ph. D. in Business Administration from The Pennsylvania State University. Her research interests include deep learning, Fintech innovation, Blockchain, and business process management. (Email: rliu20@stevens.edu)