Takeda using Advanced Machine Learning to find Insights about NASH & TRD

by Sandor Szalma, Ph.D.

Takeda Pharmaceuticals is laying the groundwork for the future of the pharmaceutical industry. The success of a recent project on deep learning methodologies against de-identified healthcare claims data is shedding light on the potential for computing power to change the trajectory of healthcare.

Advancements in predictive modeling and analytics have opened the door for a new research lifecycle, helping data scientists better understand real world disease etiology. Methodologies applying machine learning across research and development can identify novel digital biomarkers, recognize adherence patterns, and predict drug performance.

Concurrently, available real-world healthcare data is exploding in volume and variety. Thanks to the increased accessibility of – and engagement with – digital platforms, mobile apps, and wearable devices, information regarding patient activities and health status is more readily available to pharmaceutical organizations for faster processing.

The convergence of these factors is driving the ongoing transformation of the healthcare ecosystem toward more personalized treatment and patient-centric care models. For the pharmaceutical industry, this demands a steep learning curve to understand disease progression and therapies to keep up with care expectations from patients.

Two years ago, Takeda Pharmaceuticals leadership established the R&D Data Science Institute, combining quantitative sciences such as biostatistics, computational biology, epidemiology, biomedical informatics, and machine learning into an integrated unit. The Institute's mission is to transform the practice of pharmaceutical research and development through the application of digital technology and rigorous data science; and one of its first achievements was to establish a strong governance around R&D data and implement a R&D Data Hub integrating disparate high-content datasets such as observational studies, population-level biobanks, clinical trials, and real-world data. Once this data hub was ready we could turn our attention to applying advanced modeling techniques, such as machine learning, to derive novel insights and help drive data-driven decisions.

Intractable Therapeutic Areas

To test the viability of capabilities of machine learning techniques, we derived an experiment focused on two highly complex therapeutic areas: Nonalcoholic steatohepatitis (NASH) and Treatment-resistant depression (TRD). The objective of the study was to explore how to use deep models to teach us about the selected diseases' etiology, progression, and performance of therapies using large claim datasets.

The two diagnoses were selected, specifically, for their complexity. There is little understanding about the etiology and progression of both NASH and TRD. And in the case of NASH, the complexity is compounded by its newness as a specific diagnosis. Established as an ICD-10 diagnosis in claims databases in 2015, the classification carries little historical data.

The industry's existing understanding of the diagnostic pathway to NASH is limited to what we know about nonalcoholic fatty liver disease (NAFLD). What causes some NAFLD patients to develop NASH is a mystery; and without a better understanding, it is difficult to identify patients with early stages of NASH to engage in clinical trials. In many ways researchers are starting from scratch in trying to understand its prevalence, the different stages of NASH that are impacting the population, and the effectiveness of therapies at different stages.

Elucidating NASH disease progression and the effectiveness of therapies are the types of insights we are hoping deep machine learning will be able to extract and was therefore the foundation of our inquiries for this study. Specifically, we aimed to gather insights into what differentiates those who live with a NAFLD diagnosis with no documented progression versus those that progress from NAFLD to NASH.

The hope is, that by applying advanced machine learning to large datasets, such as claims and clinical EHR data, differentiating characteristics with respect to comorbidities and the diagnostic journey will become apparent. Not only will it enable our ability to improve the treatment journey for patients with NASH, this information will impact decisions on drug development investments, planning protocols for clinical trials, defining the economic value of a treatment, applying for reimbursements, and more.

In the case of treatment-resistant depression, our approach was slightly different. Our questions were more focused on the therapies themselves. Patients of depression are often prescribed multiple therapies, usually in no specific order or combination, and there is very little understanding into why some patients respond while others do not.

In other TRD research we have identified patterns in how well patients can or cannot manage their depression, but we know very little around the effects of different drug combinations or sequencing of transitions between therapies. Therefore, for this study we decided to look at two cohorts of depressed patients on SSRI therapies – one who did not switch therapies and another who did – to gain more understanding around the impact of treatment switching.

This was our initial broad hypothesis. In both cases, it took multiple iterations to narrow the questions to be posed down to specific questions that we were able to answer with the given dataset. (The claims was limited to diagnostic codes, medical procedures, prescriptions, etc.) We consistently found ourselves posing a question only to realize we either didn't have enough data or were missing a necessary metric that could only be found in deeper clinical databases such as electronic health records.

This was one of the most important learnings, I believe. Without the right questions aligned with the right data, you will uncover insights that will probably not be useful. For example, in the TRD study we identified some interesting side effects present in the cohort that was switching SSRI therapies, but as designed, the results were inconclusive for use in our development processes.

A deeper look at deep learning

In addition to the novel therapeutic questions this was an opportunity to conduct a technical pilot using Deloitte's ConvergeHEALTH Deep Miner™ platform. The platform is part of Deloitte's larger portfolio of life sciences and health care technology solutions and is designed to accelerate predictive modeling by utilizing machine learning and neural network algorithms with real-world healthcare data. Therefore, this was a test of the platforms' deep learning capabilities using real-world datasets harmonized through OMOP CDM, and validating our results against correlations extracted from simpler, linear modeling methodologies.

We found that automating the case control generation was a critical piece of the process, something we did in both the NASH and TRD models. While using queries to build our cohorts, we discovered that the machine learning platform was so precise that it turned out both meaningful correlations and extraneous ones that were artifacts of the variations across populations established when constructing definitions to label cases and controls.

To adjust, we opted to train the system to recognize similar sub-populations within the cohorts based on the available patient data (diagnostic codes, treatments, drugs, medical lab results, etc.) in order to eliminate the potential for the unintentional bias that can result from selecting random controls. We used a machine learning tool KNN (K Nearest Neighbor) to select specific matched controls to each case from within our selected cohort definitions. This allowed us to remove bias such as differences in age or gender between cases and controls that naturally occurred in building rules based cohorts and ultimately were providing results about the differences in these populations rather than the features of interest in the disease progression.

Unexpectedly, the exercise for creating the propensity matched case-control training data turned out to be fruitful in multiple ways. It highlighted interesting information about how NASH is being diagnosed, and highlighted variations between the two cohorts in the TRD study (e.g. age, level of illness) in treatment categories—although none offered any insight into why they did or did not respond to the therapies.

On more than one occasion we questioned the accuracy of the deep learning models. It is an easy assumption to make, that the deep data models would deliver better results, but there is no guarantee. In fact, we initially achieved such high accuracy that it was suspicious (around 98%), and then we would find ourselves far off because we had included a clue in the case data such as a synonym diagnosis code or procedure in the diagnostic pathway that had not been included in the disease definitions and never occurred in the control records. The algorithms we built had to be precise in order to balance the risk, handle the variables in case control, and eliminate selection bias in propensity matching.

One way we validated the accuracy of the deep learning algorithms was with linear machine learning models. The advantage of applying linear models turned out to be one of our key takeaways. Not only did they provide a baseline for testing, but they allowed us to significantly improve the accuracy and performance of the deep models.

The transparency of the linear models allowed us to identify which factors had the greatest impact on predicting an outcome of interest. For instance, the linear analysis could highlight the top 50 determinants from the hundreds of thousands of possible factors that were driving the prediction. Without narrowing it down, the deep algorithms would consume computing power and produce meaningless false positive correlations. By strategically reducing the space of variables in the feature vectors to input only the driving variables found in linear models we significantly sped up the non-linear computations and improved the prediction accuracy in the outcomes in deep learning models. In summary, the linear models allowed us to know where to look for useful signals, and the deep learning models identified patterns indicating a specific classification or prediction from those variables through sequences such as time or order of events. The level of accuracy we achieved will lay the groundwork for future research into which clusters of sequences are meaningful for predicting a specific outcome and for experts in the disease to help deduce why.

> **"ON MORE THAN ONE OCCASION WE QUESTIONED THE ACCURACY OF THE DEEP LEARNING MODELS. IT IS AN EASY ASSUMPTION TO MAKE, THAT THE DEEP DATA MODELS WOULD DELIVER BETTER RESULTS, BUT THERE IS NO GUARANTEE."**

What's Next?

Consider if life science organizations could better predict how the combination or order of therapies impacts treatment, such as why some patients respond to a second or third line therapy and others require a specific combination. Health care professionals would be better positioned to develop personalized therapy plans that get to the root of the problem quicker or prescribe the right drug regimens faster. This could greatly reduce experimentation of multiple drugs and the time spent experiencing limited-to-no benefit from treatments for an individual patient, and therefore in the end, it may also reduce the total healthcare cost.

The success of our experiment was in that it confirmed that the technology is ready for prime time, as long as the user has the right setup, the right data, and the right questions. For us, that consisted of having a system such as the ConvergeHEALTH Deep Miner pipelines and models, in an appropriate framework, sitting on top of a harmonized data lake environment. Furthermore it consisted of strong engagement from and collaboration with the clinical and epidemiology experts to guide the deep learning analysis process to understand when insights are already known, when to ask a key follow-up question, see what is novel in results vs. literature, and when to question potentially spurious correlations as artefacts of the experimental design. It is a matter of restructuring questions in an agile way to create a scientific approach towards identification of a cause and effect, not just correlations, using the combination of linear and non-linear models.

We are well positioned to continue to build on this work in multiple dimensions. The system we built is not dependent on the claims dataset, but rather, a self-learning set of tools that can be easily re-applied to a similar problem with a wider lens of data such as medical records, molecular markers, lab tests, or digital phenotyping. More so, we developed strong methodologies for matching patients and controls, linear and non-linear machine learning, and around team collaboration that can be scaled up to accommodate wider and more complex problems.

Our dataset was made up of more than 200 million subjects, yet they did not all suffer from depression or NASH. Wider and deeper datasets are necessary to target the fine-grained, specific questions necessary in drug discovery and development contexts. These could include lab results, imaging results, EMRs, and even alternative data.

For our next steps, we are developing a strategic plan to obtain the deep data for all the diseases we want to research so that we can gather the insights most important to our work. We have the visibility and plans to improve the model building process to use whole data sets to improve classification logic through new types of feature vectors with derived features in clustered information that can be fed into linear models.

In hand with this, we want to improve our queries to be able to identify more meaningful information, such driving patterns or significant sequences. We also reviewing ways to automate and visualize the otherwise 'black box' results from deep learning to have explain ability for prediction performance.

Until now, the number of variables involved have made this an impossible task for humans therefore keeping these insights out of reach.

Deep learning is opening the door for new discoveries. The greatest impact is going to be in areas with significant complexity, like NASH and TRD. The ability to recognize a disease trajectory and identify patients who are more likely to respond, and therefore improving the treatment journey, can have a major commercial impact for life science and health care organizations.

Rare diseases also fall into this category of complexity. Right now, what is frequently happening in pharma is the development of new drug categories based on emerging biological understanding simultaneously with the emergence of new disease definition. Deep learning is capable of more quickly linking the two.

With all this said, the success of machine modeling still relies on the knowledge of subject matter experts. The machine may spit out the most important correlations but it does not understand what it is looking at; for example, some of the strongest correlations in patients that developed NASH was

A network of support

As is the case with any successful study, the technologies behind the scene play just as an important of a role as those front and center. Takeda's study utilized on a number of supportive and accelerator solutions to improve the processes and outcomes of the pilot.

Standardizing Vocabularies

There are significant initiatives across the life sciences industry to standardize data with common vocabularies and terminologies. The investments into these accelerators are key to moving the machine-learning community forward.

Our team heavily leveraged open source standards, such as OHDSI OMAP CDM vocabularies, to harmonize datasets from different claims vendors. Having a common formatting and vocabulary framework made it possible to cross-compare the results, and also provided important insights into the fine grained differences between different real-world data sources.

Synthetic Data

Synthetic data plays a significant role in the design and testing of machine learning models. Using synthetic data, researchers can create a baseline for their experiment, and test their methodologies and techniques at various scales.

Takeda used SYNPUF data to run models before working with the licensed commercial datasets. We were able to test and improve our algorithms at a smaller scale to ensure we had the accuracy we wanted. There was also an added benefit that it enabled researchers without access to the controlled Takeda environment to work on the methodologies using synthetic data in the CDM format.

Big Data

The application of cloud technologies like Deloitte ConvergeHEALTH Miner running on Amazon Web Services made it easier to rapidly handle big data while controlling infrastructure costs during experiments.

conducting liver biopsies and analysis of alpha fetoprotein, which clearly recapitulated the diagnostic trajectory.

One of the benefits of the collaboration between Takeda and Deloitte was having access to specialists in both machine learning and in the disease states and conditions that were under examination. The Takeda team is dedicated to understanding the biological, clinical, and epidemiological background of these diseases and the patients. The Deloitte ConvergeHEALTH team and the Deep Miner platform glued the pieces together. ■

Sándor Szalma is head of biomedical informatics in Takeda R&D Data Science Institute. He is responsible for computational biology, machine learning and informatics approaches for forward and reverse translation and scientific and competitive intelligence efforts in oncology, neuroscience and gastroenterology. Most recently, he was head of Translational Informatics and External Innovation, R&D IT in Janssen Research & Development, LLC. He serves as a member of the governance board of Open Targets. Previously, he was member of the industry advisory committee of ELIXIR, member of the board of the Pistoia Alliance, member of the Translational Medicine Advisory Committee of the PhRMA Foundation and led the Data & Knowledge Management Strategic Governance Group of Innovative Medicine Initiative. His past positions included president of MeTa Informatics, general manager of QuantumBio and senior director of Computational Biology and Bioinformatics at Accelrys, Inc. He was co-founder of Acheuron Pharmaceuticals, Inc. He lectured at UCSD Extension and was adjunct professor at Rutgers University in the Computational Biology and Molecular Biophysics program. He is the author of 45 scientific publications and book chapters and two patents.
He received his doctoral degree in physical organic chemistry from A. Szent-Györgyi Medical University in Szeged, Hungary.