Translational Data Fusion Architectures for Efficacy Signal Propagation: AI-Enhanced Systems for Clinical-Preclinical Data Integration in Drug Development

Dr. Li Guo, Professor of Computer Science, Nanyang Technological University (NTU), Singapore

1. Introduction to Integrating Clinical and Preclinical Data

Integrating clinical and preclinical information with emerging evidence during drug development is crucial for increasing the probability of making the correct decisions. A number of different types of information can be combined. This can be done at the level of connecting clinical trial results to real-world evidence, walking through the enriched virtual trial concept. That means that data from laboratory studies, such as those advancing target validation, can, through methods such as integrated mathematical modeling, be useful to check assumptions made to underpin the unmet medical need, translate into a clinical setting under the term of mechanistically enriched virtual trials, and aid decisions on whether an idea is worth testing in a clinical trial. This translational information can come from many different sources, such as modeling, epidemiological data, and registries. Also, during late-phase research, more data, such as real-world evidence, can be incorporated. The advent of mobile technologies results in the integration of new data sources also in earlier stages of drug development.

There is a growing appreciation for the added value of data integration and the potential of combining clinical and real-world evidence. The added value of these integrated data is that we can use these systems-based approaches in drug development to develop earlier, faster, and cheaper decision-making, thereby increasing the probability of getting better treatments to patients faster than the other development pathways. Indeed, an analysis revealed that a benefit-risk assessment combining clinical trial data with retrospective real-world evidence could save up to six months in the pre-investigational new drug stage during discovery. The contribution of integrated data to the rational use of drugs is increasingly recognized. Against that background, an urgent

need exists to develop and formalize the methodologies to qualitatively and quantitatively translate across the translational chasm. Prior attempts to review the potential ways to perform such research have identified intellectual debate, probably due to a tracking bias across scientific communities. Critically, all members from scientists to knowledge brokers to pharmaceutical companies, who would benefit from this information, are inadequately networked under current systems. That stands in their way to perform the recently declared 'status: quo vadis?' as proposed in a much elaborated report.

1.1. Significance of Data Fusion in Drug Development

Quantifying the integration and application of data across the drug development lifecycle using AI-enhanced systems

1.1. Significance of data fusion in drug development

Data fusion, i.e., the effective integration and exploitation of both clinical and preclinical data or any other type of multi-omics data environments, has the potential to design better drugs and achieve better patient outcomes. To this end, the combination of diverse and multi-scale observational and necessary predictive data sources with results from controlled and defined experimental perturbation studies is expected to enrich the understanding of drug efficacy and toxicity endpoints. More generally, the convergence of information from various omics and disease models across disease pathophysiology stages can reduce the time and cost associated with the identification of lead molecules and also improve the outcome and reduce the time and risk of drug development through unbiased decisions based on the analysis of the entire drug discovery and development project workflows.

Examples of multimodal analyses across different platforms leading to the identification of novel target candidates and drug combination therapies are becoming increasingly popular. A good example of such a strategy is the use of large-scale high-content screening results for the integration of systems biology-derived sub-network composition along with unbiased disease knowledge converted into a target risk score. In one such study, the combination of in silico systems biology analysis with individual molecular assay-generated data has been applied to reconstruct biochemical pathways from high-content screen readouts and further evaluate sub-network composition and

hypothesis prioritization into the design. Additionally, in vivo work in non-human primate spinal cord injury and humanized spinal cord injury studies revealed that citicoline loads this signature of neural injury transporter in spinal cord injury and has potential meaning in multiple disease cascades with CNS involvement. A recently published proof-of-concept study has also demonstrated how causality approaches targeted for disease signature causality discovery can be used to define predictive and early molecular toxicological readouts using both publicly available data and proprietary biomarkers to identify and validate potential biomarkers for predicting potential readouts for test articles or drugs equipped with high specificity for PMI. Moreover, the development of a portable platform for the identification of the initial potential PMI has the potential to help reduce late-stage toxicity findings in clinical trials in humans. These studies clearly highlight the ensuing advantage of information fusion across different biological platforms.

2. Fundamentals of Machine Learning in Data Integration

In general, machine learning applications consist of several subsequent steps. Data preprocessing involves removing missing values or converting categorical variables to a numerical representation. Furthermore, preprocessing aims to normalize or standardize the data to distribute the values of different features. Model selection comprises the choice of an appropriate model fitted to the problem characteristics. Last but not least, feature engineering can be applied in order to extract relevant information contained in the data and to deliver new, more valuable characteristics, or to 'amplify' the influence of selected data samples.

Several modeling techniques can handle data of increased dimensionality, complexity, or size. Artificial neural networks and particularly deep learning models, such as convolutional neural networks or recurrent neural networks, enable the use of a large number of interconnected variables for modeling. Their training often demands increased computational effort in order to optimize the extracted model, making the appropriate hardware environment a further requirement, in addition to the algorithmic approach. The main purpose of applying machine learning in prediction and decision-making processes consists of employing computational models of learning to improve the prediction and decision-making process, in order to have an impact on healthcare. Relevant examples include predictive models of risk, recurrence, or prognosis. In

addition, they may be exploited to identify sub-populate associations, pathways, or processes that are driving the phenotype of interest. Many successful examples have been reported, involving different domains and health issues. Additional examples are given throughout this work.

2.1. Supervised Learning Algorithms for Data Fusion

Supervised learning relies on labeled datasets to learn the associations between a set of predictor variables and a target variable. The training data is used to learn the relationships between variables, and then a supervised learning model is developed and validated. The model can then be used to predict the unknown target variable in a new dataset. A test set can also be used to further evaluate the generalization power of the developed model.

Compared to unsupervised learning algorithms for data fusion, the main advantage of using supervised algorithms is that they consider the target variable when learning associations between inputs across datasets. Precision in tuning the level of bias and variance in multicohort datasets enables improved accuracy in predicting drug responses, and thus increases translational success. Supervised learning algorithms for data fusion can be generally categorized into regression, nearest neighbors, Bayesian techniques, decision trees, and support vector machines. Regression analysis is widely used to discriminate biological groups based on molecular features and learn associations between preclinical and clinical datasets for different translation applications. Decision trees and support vector machines further partition the outcomes into increasingly homogeneous groups; the classification trees are parallel to linear decision boundaries learned by support vector machines, but over most applications do not show improvements in predictive classification for fusing data with genetics, due to relatively internal heterogeneity. In addition to the selection of appropriate supervised learning algorithms for data fusion, the performance of the algorithms strongly depends on their fine-tuning. Regression algorithms have many adjustable parameters that are critical to performance, including the regularization strength, penalty type, kernel parameters, and the algorithm dimensionality. Underfitting potential does not optimally adjust these in the learning algorithm in a multicohort context, and in order to determine which parameters are optimal will require large sets of combined preclinical and clinical data. The main challenge of applying supervised learning algorithms is deciding if and

when to level on these hyperparameters, which directly requires sufficiently sized and contextually relevant subsets. In general, it cannot be predicted in advance which algorithm and parameters will be best for which learning task. Most clinical studies do not have a sufficient number of patients and drug responses targeted for approval and admission to validate the superiority of using novel supervised learning algorithms. Therefore, evidence from simulations of independent patient cohorts can provide evidence to overcome this limitation. Several studies were conducted to apply supervised machine learning algorithms in integrating clinical and preclinical data in drug development to aid better translation in the preclinical stage and provide stronger evidence of drug efficacy in clinical settings.

In essence, supervised learning models have strong potential to fit a rich class of joint ranges given they have enough data. Overall, supervised learning models can learn different relevant information from multiple levels of biological organization for integrating and benchmarking. More studies in quantifying evidence will shift the translation paradigm from preclinical to clinical and reverse-translational by determining the optimal, most sensitive, and most specific algorithms for the patient population.

3. Challenges and Opportunities in Leveraging AI for Data Fusion

The following section discusses the challenges and opportunities of harnessing the power of artificial intelligence (AI) to perform automatic data fusion. There are many obstacles to achieving this goal, beginning with the fact that data are inherently heterogeneous. Different data sources have diverse data collector conventions, values, and usage; when multiple data sources are combined, the risk of bias is amplified. Each step of the data pipeline, including data cleaning, transformation, and integration, can introduce bias. Moreover, data privacy, security, and governance may also be obstacles because valuable clinical data are sensitive and the regulations and penalties associated with their misuse can be severe. The complexity and variability of integrating data affect research efficiency and reproducibility. In addition, the inherent bias of the data may confound the AI prediction of outcomes and jeopardize experimental outcomes. The value of using AI to integrate various data sources lies in the power to uncover hidden, non-obvious patterns. However, these same non-obvious outputs may have unknown explanations, which creates additional challenges in their use by domain scientists.

Despite these challenges, AI algorithms are increasingly being deployed to predict the outcome of clinical trials or the potential success of new cancer drugs. Moving forward, it is important that industry, academia, and institutions come together to establish best practices and leverage this technology, while at the same time avoiding the introduction of unwanted bias. One way to begin to mitigate these challenges is to analyze use cases in which innovative organizations have been successful in tackling these particular issues. In these following use cases, examples demonstrate how AI can be part of the solution to facilitate the integration of preclinical and clinical data.

3.1. Ethical Considerations in AI-Enhanced Data Integration

3.1. Ethical Considerations in AI-Enhanced Data Integration

Patients participating in clinical trials, the clinics in which they receive treatment, and preclinical data are sensitive and private within the European Union, which has some of the strictest data privacy laws in the world. As clinical data are anonymized and moved through bifurcated systems, there also needs to be assurance of how knowledge gleaned from the anonymized data can be fed back to inform treatment within a specific clinic. Trust frameworks established between clinics and patients would be completely violated if trial data were not managed and kept fully private and secure.

Algorithms, if not carefully created and managed, can inherit the implicit biases we possess. If these data-driven tools are used to support shared clinical decision-making or guide treatment planning, algorithmic bias can lead to disparities in the actual care received across multiple health centers, or could provide increased health inequality based on the vulnerability of certain populations. A transparent process for understanding AI predictions can be understood within the pipeline that data travel. Accountability is also important, both for the interpretation of knowledge secured from data as a result of AI processing, and in terms of real-world outcomes that may result from decision-making based on incorporated AI data. There is also no doubt that employing AI data fusion algorithms for drug development entails ethical considerations. Stakeholders driving AI for drug development need to put in place specific values and rules of conduct to give guidance to those developing the AI algorithms and overseeing how they are used, and holding them accountable for their actions. As a patient-centric initiative, eTRANSAFE will involve patients and caregivers to review the plan, ensure it is patient-centric, help embed ethics and inclusion, and

advise on any unintended consequences of our work. Results on those interactions should be described as the project progresses. An ongoing dialogue between ethicists, technology developers, regulators, and those who will use the digital therapeutics in the clinic will be critical to ensure the results of AI-based knowledge extraction in data fusion inform translational research in a responsible and ethical manner.

4. Case Studies and Applications in Translational Research

Translational research case studies: The three case studies in this section showcase how data integration has been applied to clinical and preclinical data from a range of different sources and methodologies, leading to breakthroughs in drug development. The first case study describes how the combined use of microarray data from human in vitro systems and rat in vivo experiments enabled a new and potentially important biological hypothesis to be defined. It discusses the advantages of combining in vitro and preclinical in vivo data to improve robustness and fit-for-purpose utility. The second case study describes an interdisciplinary pilot study where in vivo data from rats exposed to inhaled biopharmaceuticals was combined with a study on human epithelial cells grown at the air-liquid interface, which was exposed to the same materials. This case study demonstrates the value of involving toxicologists and clinicians in early project stages to avoid fundamental mistakes that jeopardize the success of an integration project. The third case study describes a custom microarray approach to successfully integrate several phenotypic data layers using normal human lung tissue and the impact of this collaborative work.

The data integration and systems biology community has long advocated interdisciplinary research to foster innovative data-integration approaches: three case studies presented in this paper have taken starkly different clinical samples and followed unique data generation or integration strategies. The diversity of the data produced allowed for the broad experience of the different research groups to be harnessed, illustrated the potential of experiential integration, and highlighted major issues in stigma within the field. The three case studies outlined in this paper all have in common: a consideration of biological relevance; a combined data analysis leading to potentially valuable hypotheses; and the involvement of multiple stakeholders from the outset. The case studies we present provide transferable guidance for data production, integration, and analysis. The opening statement could easily be interchanged with

'omics' type data and underscores just how much we can gain from the integrated, rather than individualized, application of systems biology methodologies in the future.

4.1. Successful Implementation of AI in Drug Development

There are a number of instances where AI has been successfully implemented at different stages of the drug development process. In drug discovery, AI is used to predict absorption, distribution, metabolism, excretion, and toxicity profiles of compounds, automatically extract drug-dose response curves from raw publications, and translate raw genetic information and patient health data to potential patient diagnoses and treatments. In the clinic, AI helps optimize trial planning to improve clinical trial efficiency, support clinical image analysis, and predict medical equipment downtime. In this special issue, we focus on lessons learned for integrating clinical and preclinical data, an important and challenging bottleneck in drug development that ranges from the preclinical phase through to managing approved drugs.

These case studies illustrate how, with careful application, AI can assist in solving complex challenges to support small and large decision-making hierarchies. These AI tools have had impacts on the time taken for drug development, entry of improved drugs into the clinic, better validation of preclinical studies with predictive value to the clinic, improvements in production subsystems, and business case discussions. However, with this also comes some important lessons, not least that it is essential, in practice, to monitor and validate the AI implementation. Regular monitoring of the entire system and potential improvements would have been necessary in all these cases to guarantee ongoing benefit. There can be some challenges associated with the use of AI. For example, users may be hesitant to generate or request predictions, missing the chance to gain insights in a timely manner and potentially improving the quality of the decision or operation. For these approaches to work in practice such that domain experts tasked with decision-making use the systems, effective designs are needed to make AI interventions easy to use and desirable in comparison to alternative approaches. In addition, AI integrations need to be sufficiently compatible with existing data and systems to ensure real benefit is achievable. Each implementation described in these stories has involved a degree of adaptation to maximize gain from integrating tools and data, and purchasing of ancillary data to provide more comprehensive models. Tailoring the AI tool to the assets, data, and challenges identified supports successful integration

within a decision-making framework. Integration with real-world clinical data often results in complexities that can be foreseen, such as legal or business limitations that mean the AI intelligences are not directly used in decision-making. A variety of insights and success points emerged, from a data scientist-skilled team member leading AI work, to timely identification of in-progress risks helpful for intervention planning, to favorable general attitudes to AI from users. Overcoming such hurdles will depend on the details of each case, but in every instance, a number of actions are recommended to leverage AI for clear gains in drug discovery and development.

5. Future Directions and Emerging Trends in AI-Enhanced Data Integration

Future Directions and Emerging Trends

The promise of AI-enhanced data integration for drug development With ongoing technological advancements, the way that clinical and preclinical data are integrated is likely to change. Potential future directions we identify include combining and extending existing data integration methodologies such as advanced federated learning and semantic-based matching for improved predictive power, as well as taking advantage of new data endpoints and systems biology to improve the cross-correlation of various data types of interest.

Regulatory and framework changes affecting integration of AI techniques Currently, regulatory agencies have no formal frameworks for interpreting AI models during the process of drug development. These types of models implemented by AI methods are seen as "black boxes" for the pharmaceutical industry. This is rapidly changing, and regulatory agencies now have significant initiatives to develop pathways to accept AI models for various aspects of drug development.

Sharing AI data sources for improved predictive power AI models are also expensive to develop and implement, and as a result, there has been little effort in the biopharmaceutical industry to integrate these models into the drug development cycle. AI market consolidation among a variety of partners, as well as publicly shared resources, is opening the drug R&D space for smaller entities to engage in data science and AI. Additionally, public sharing of AI-embedded data resources also enables cooperation on the evolution and validation of neuroscientific AI. Given the expertise and need at various biopharmaceutical companies, publicly shared AI models could also

potentially serve as the basis for consolidating knowledge or landscape analyses for specific neurodegenerative diseases.

Patient-centered considerations Modern AI in drug discovery and development includes numerous datasets from human subjects, ranging from instrumental testing to patient self-reported outcomes. Many machine learning and AI methodologies aim to predict patient response or disease condition using as many personally connected data as possible. Taken together, the pursuit and practice of AI encourage a patient focus because they are computationally connected with a variety of biomedical data endpoints.

6. Conclusion

Integrating clinical and preclinical data is an important way to optimize the pharmaceutical industry. However, data integration is not straightforward since datasets vary considerably in structure and quality. We have discussed the role of machine learning and AI in addressing such challenges before providing an example of an AI-enhanced system in clinical trials optimisation. Furthermore, we provide an overview of various software solutions built for this purpose following the four-level AI library architecture for analytics, operations, and strategy. Ethical concerns about integrating clinical and preclinical data are revisited to address the integration of other biological datasets. In such efforts, benchmarking will be a necessary part considering various challenges, e.g., inter-laboratory replication and platform-induced bias. Exposure to different therapeutic regimens among patients, the role of drug regimens in patient microbiome composition, and dietary or genetic factors are all likely to affect human reactions to drugs. Ultimately, collaborations between researchers, technology developers and regulatory bodies will be needed to successfully integrate clinical and preclinical data. Integrating clinical and preclinical electronic health records and outcomes with high-throughput 'omics' data within an AI-enhanced drug development pipeline is of primary importance to the future of pharmaceutical development. Progress is hampered by the existence of multiple regulatory organisation requirements, the absence of regulations, and ethical concerns about data sharing. Additionally, it is difficult to perform automated analysis when data is imported, as it sometimes requires manual input. In this section, we review potential strategies for the development of novel therapeutic interventions into the future. The AI-optimised and automated

chemical rescreening paradigm suggests synthetic lethality pairs of repurposing opportunities in both BRAF-inhibition sensitive and resistant melanoma cells. A case study for the treatment support systems is discussed, which aims to improve a patient's quality of life on dialysis by optimising their haemoglobin and ferritin levels. Following this, we discussed AI-optimised clinical trial algorithms that determine the best combination of trials for a patient according to their unique genetic signature. We end by addressing several challenges in these AI applications, such as patient distrust about the predictions, explainability, and ethical considerations that will be elaborated upon in the next section.