The Adaptability of using AI for Drug Risk Assessments within Regulatory Science: A Case Study of DeepDILI

Skylar Connor[1], Ting Li[1], Ruth Roberts[2,3], Zhichao Liu[1], Weida Tong[1]

1. National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR, 72079, USA
2. ApconiX Ltd, Alderley Park, Alderley Edge, SK10 4TG, UK
3. University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

U.S. FOOD & DRUG ADMINISTRATION
National Center for Toxicological Research

Introduction

Artificial intelligence (AI) has played a crucial role in recent advancements within the biomedical field, especially in areas like drug risk evaluation and assessment; however, from a regulatory science perspective, AI has yet to have the impact it merits. As the field of regulatory science advances, in silico and in vitro approaches have been extensively evaluated as alternatives to some animal studies in a drive to identify and mitigate safety concerns earlier in the drug development process. Although several AI tools are available, few are being used in regulatory areas (e.g., drug efficacy and safety evaluation) and to support the FDA review process. As two important aspects of regulatory significance – especially for the application of AI – the application domain and context of use play central roles in enhancing AI solutions for Drug Risk Assessments within the regulatory arena. Within the field it is commonly known that AI models improve every time more data are added to the training set; however, this approach has not been extensively challenged within the scope of drug risk assessment models for regulatory use. In this study we propose a way to explore the adaptability of an AI solution for Toxicity and Drug Risk Assessments within the regulatory arena. To confirm whether a drug risk assessment model improves as more data are added to the training set, we set up a comprehensive study to mimic the real-world scenario of annually adding novel drugs to the market, using a model we previously developed known as DeepDILI. In using this approach, two important questions can be addressed:

First, did the models performance improve or decline as more data were added? In order to evaluate the performance of DeepDILI we conducted a comprehensive study to better understand how a model's performance evolved when more data were available. The DeepDILI model was developed using drugs approved before 1997; for the adaptability assessment of DeepDILI we incrementally added drugs approved after 1997 to the model in a chronological fashion to test whether the model improved as more data were added.

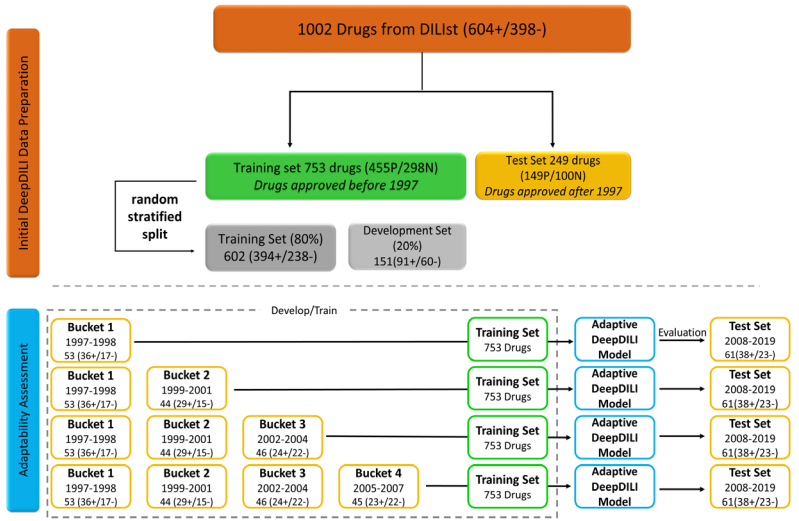
Second, has the context of use changed as the model adapted? There must be a firm grasp on the applicability of a new model, meaning the developer must understand both where and how their model is best utilized. Establishing an AI model's context of use (the where) and application domain (the how) are essential to ensure the effective and efficient use of the model within the regulatory arena.

It is crucial that **every time new data are incorporated, or a new model is developed, these two aspects of regulatory significance are assessed.**

Case Study: The Adaptability of DeepDILI as a Drug Risk Assessment Model

DeepDILI is a model that was developed to assist in the drug risk assessment of drug-induced liver injury (DILI). Through leveraging the advantages of AI and Deep Learning we have developed an in-silico approach that identifies drugs with DILI potential during the early stages of drug development. The initial DeepDILI was a locked model developed using the DILIst data set: the training set contained 753 drugs approved before 1997 and the test set contained 249 drugs approved after 1997. The adaptability of DeepDILI was assessed by adding new drugs to the training set to develop adaptive DeepDILI models. For this adaptability assessment we split the locked models test set (with the yellow box on top) into five adaptive groups approximately the same size (with the yellow boxes on the bottom). To mimic the real-world scenario of annually adding novel drugs to the market, we increased the number of new drugs in the training set stepwise and chronologically added each bucket of drugs. For convenience we are using one adaptive model structure to explain this adaptive model framework: starting with the locked models training set (in green) we added one bucket at a time using the drugs from 1997 to 2008, generating four adaptive models. Once trained each adaptive model was evaluated with the test set, which in this case is the 2008-2019 group. This process was reiterated five times; each time a different bucket served as the new test set and all remaining buckets were added to the training set as described above. Additionally, the performance of the four adaptive DeepDILI models was compared to the initial Locked DeepDILI model each time a new test bucket was used, the results of which are shown in Figure 2.

Figure 1: Adaptability assessment framework



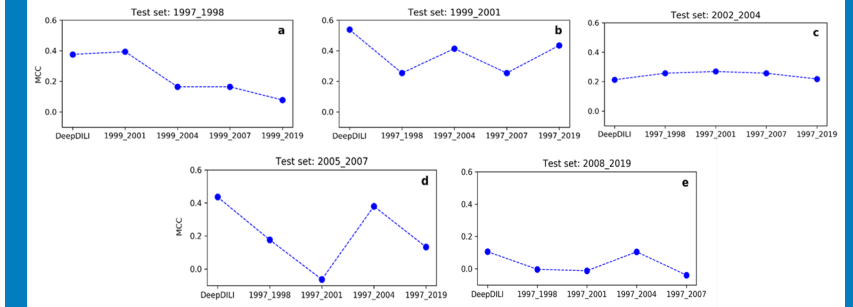


Figure 2: MCC trend of DeepDILI and the Adaptive DeepDILI models within sub test

To investigate whether the comprehensive study to evaluate the adaptability of the DeepDILI model is positively associated with the increasing number of drugs in the training set, we checked the MCC of each adaptive model for each test set. In **Figure 2a**, the MCC values of the adaptive DeepDILI models for test set one decreased as more drugs were added to the training set. In **Figures 2b and 2d**, the MCC values of the adaptive models for test sets two and four presented as a wave shape as more drugs were added to the training set. In **Figures 2c and 2e**, the MCC values of the adaptive models for test sets three and five exhibited a relatively flat trend as more drugs were added to the training set. All of these results indicate that adding more drugs to the training set does not substantially contribute to the performance of the adaptive models.

Conclusion

How did our model fulfill the two aspects of importance to regulatory significance?

1. **Did the performance improve or decline?**
   In the case of our DeepDILI model the results indicated that adding more drugs to the training set did not substantially contribute to the performance of the adaptive DeepDILI model. Although there was no positive relationship between the models' performance and the number of drugs in the training set, shown by Figure 2, we found that the most critical factor in accessing a model's performance is the data included in the test set.

2. **Did the context of use change?**
   DeepDILI was developed to assist in the drug risk assessment of drug-induced liver injury (DILI). Since we carefully picked the data included in the model to help firmly establish our model's context of use its regulatory significance did not falter.

Overall, based on these findings we have concluded that the proposed adaptability assessment framework has utility in the evaluation of a models' performance over time, which would greatly support the advancement of AI-based models in regulatory science.

Reference

Anklam, E., et al., Emerging technologies and their impact on regulatory science. *Exp Biol Med* (Maywood), 2022. 247(1): p. 1-75.

Li, T., et al., DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation. *Chem Res Toxicol*, 2021. 34(2): p. 550-565.

Dirven, H., et al., Performance of preclinical models in predicting drug-induced liver injury in humans: a systematic review. *Sci Rep*, 2021. 11(1): p. 6403.