Introduction

Integrating artificial intelligence with bioinformatics promotes public health

Huixiao Hong[1] and William Slikker[2]

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA; [2]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA
Corresponding author: Huixiao Hong. Email: Huixiao.Hong@fda.hhs.gov

Experimental Biology and Medicine **2024; 248: 1905–1907. DOI: 10.1177/15353702231223575**

This thematic issue is a product of the 9th annual conference of the Arkansas Bioinformatics Consortium (AR-BIC), which was held on March 13 and 14, 2023, with the theme "Bioinformatics, Big Data, AI (Artificial Intelligence), and Public Health: An Integrated World." This conference gathered more than 180 scientists and trainees with diverse scientific interests discussing current research works and future perspectives on coupling bioinformatics and AI for promoting the public health. The conference intensively discussed bioinformatics and AI in pathogen surveillance and microbial genomics, showed various applications of AI and bioinformatics in healthcare, displayed emerging bioinformatics and AI approaches in advancing integration of multi-omics data in biology and medicine, and explored a variety of machine learning and deep learning methods for big data analysis and drug development. Moreover, the conference hosted two workshops, one on AI applications in natural language processing, and the other on utilization of deep learning in analysis of histopathological images. The articles included in this thematic issue are from participants of this conference and highlight some scientific efforts of AR-BIC to promote public health by integrating AI with bioinformatics.

Toxicity prediction has become a critical component in various endeavors, including drug development, environmental science, and chemical safety. With the growing concern over chemical and biological hazards, there is an increasing demand for accurate and efficient toxicity-prediction models. The review article by Guo *et al*.[1] provides a comprehensive overview of the state-of-the-art technology in machine learning and deep learning models for toxicity prediction. The core of this review delves into the methodologies and algorithms employed in toxicity prediction. It discusses classical machine learning models such as random forest, support vector machines, and ensemble methods, highlighting their strengths and limitations. In addition to model architecture, this article explores data sources and collection methods, as well as data augmentation and imbalanced data-handling techniques, crucial for developing reliable toxicity-prediction models. It concludes with a critical analysis of current challenges and future directions in the field and discusses issues related to data availability, model interpretability, and the need for standardized evaluation metrics. This comprehensive review serves as a valuable resource for researchers, practitioners, and decision-makers in various fields where toxicity prediction plays a pivotal role.

Medical image segmentation, particularly in the context of brain tumor analysis using magnetic resonance imaging (MRI) scans, has witnessed significant advancements due to the integration of machine learning and deep learning techniques. Accurate and efficient brain tumor segmentation plays a pivotal role in diagnosis, treatment planning, and disease monitoring. Khan *et al*.[2] give a thorough examination of the state-of-the-art machine learning and deep learning models applied to brain tumor MRI image segmentation. This review first outlines the critical importance of accurate brain tumor segmentation in clinical practice and research and highlights the impact of segmentation on early diagnosis, surgical planning, and the assessment of treatment efficacy. It then comprehensively explores the methodologies and algorithms used in brain tumor MRI image segmentation, including classical machine learning approaches such as support vector machines. It extensively discusses deep learning methods, including convolutional neural networks (CNNs), U-Net, and other state-of-the-art architectures, showcasing their remarkable performance in addressing the complexities of medical image segmentation. Evaluation metrics and benchmark datasets are also explored to provide insight into the comprehensive assessment of segmentation models, thus facilitating informed comparisons between different approaches. This article concludes with current challenges and prospects in the field, emphasizing interpretability of deep learning models and the need for standardized practices. It provides an invaluable resource for researchers, radiologists, and health-care professionals to advance the development of accurate and efficient tools for brain tumor analysis, ultimately contributing to improved patient care and outcomes.

The integration of data science in drug discovery safety has ushered in a transformative era in pharmaceutical

research. Roberts[3] writes a comprehensive review of the challenges and opportunities presented by data science techniques in the context of drug safety assessment. This article highlights the increasing reliance on data-driven approaches in the pharmaceutical industry, emphasizing their potential to expedite drug development, reduce costs, and improve patient safety. It discusses the significance of harnessing vast and diverse datasets, including chemical, biological, and clinical data, to enhance safety prediction and monitoring throughout the drug development lifecycle. This review also explores the challenges faced by data scientists and researchers in the field of drug discovery safety and underscores the need for innovative data integration techniques, scalable computational infrastructure, and standardized protocols to address these challenges effectively. In addition, it explores the myriad opportunities data science affords to pharmaceutical research and development. This article offers an informative resource for researchers, pharmaceutical professionals, and policymakers to foster a deeper understanding of the evolving field of data science in drug discovery safety.

The advent of Bidirectional Encoder Representations from Transformers (BERT)-like large language models (LLMs) has revolutionized natural language processing and, in turn, their application in domains like patient safety and pharmacovigilance (PSPV). Wang *et al.*[4] report an in-depth comparative study on BERT-like LLMs for causal inference in PSPV. Generic pretrained BERT LLMs, domain-specific pretrained LLMs, and domain-specific pretrained LLMs with safety knowledge–specific fine-tuning are compared using three publicly accessible PSPV datasets to assess the influence of data complexity and model architecture, the correlation between the BERT size and its impact, and the role of domain-specific training and fine-tuning. They find that data complexity and model size have little impact on the performance of the BERT-like LLMs, and domain-specific BERT-like LLMs outperform generic pretrained BERT models in causal inference. This study demonstrates the challenges and opportunities associated with utilizing BERT-like models in PSPV, particularly in the context of causal inference. Their findings highlight the importance of BERT-like models for the timely detection of adverse events, signal detection, and the potential for proactive risk management in drug safety, providing insights into the utility of these models in automating the extraction and analysis of valuable information from unstructured textual data sources.

In the domain of pharmaceutical research and healthcare, efficient and accurate analysis of drug labeling text is crucial for drug development, safety monitoring, and informed clinical decision-making. Wu *et al.*[5] introduce RxBERT (a BERT model pretrained on Food and Drug Administration [FDA] human prescription drug labeling documents), a novel approach that leverages advanced natural language modeling techniques to enhance drug labeling text mining and analysis. RxBERT is evaluated with multiple datasets, resulting in slightly better performance than other natural language processing models. RxBERT is a domain-specific language model fine-tuned on a large corpus of drug labeling texts. Their findings demonstrate the significance of RxBERT in advancing drug labeling text mining and analysis in the pharmaceutical and health-care industries.

The opioid crisis in the United States has raised significant concerns regarding the safety and adverse effects of opioid medications. The research article by Le *et al.*[6] present a comprehensive systematic analysis and data mining study of opioid-related adverse events reported in the FDA Adverse Event Reporting System (FAERS) database. This research explores the methodologies employed to systematically analyze the opioid-related adverse events in the FAERS database. With appropriate data preprocessing and quality control measures, the authors created a structured dataset suitable for in-depth analysis. They then utilize data mining and statistical techniques to identify trends, patterns, and associations among adverse events and opioid drugs in the FAERS database, yielding 3317 pairs of potential risk signals for the 13 opioid drugs. The uncovered specific adverse events are statistically significant and shed light on the prevalence and severity of opioid-related side effects. By systematically analyzing and data mining opioid-related adverse events from the FAERS database, this research contributes to a better understanding of the risks associated with opioid medications and informs evidence-based strategies to mitigate these risks, with the goal of addressing the opioid crisis and improving patient outcomes.

The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, continues to pose a formidable global health issue. In the absence of specific antiviral therapeutics, drug repurposing has emerged as a promising approach to identify existing drugs with potential efficacy against SARS-CoV-2. Central to this strategy is the targeting of the SARS-CoV-2 Main Protease (Mpro), a key enzyme essential for viral replication. Liu *et al.*[7] report the development and validation of a novel Random Forest machine learning model designed to predict the binding affinity of various small molecules to the SARS-CoV-2 Mpro. They constructed the predictive model using a comprehensive dataset of known Mpro-binding ligands, encompassing a diverse set of chemical compounds. Mold2 molecular descriptors,[8] including structural and physicochemical properties, were calculated for training and validating the Random Forest model, which demonstrated excellent predictive performance, as indicated by rigorous cross-validation and external validation against experimentally verified Mpro inhibitors. Furthermore, they employed this predictive model to screen a vast library of FDA-approved drugs for their potential to bind to the SARS-CoV-2 Mpro. Their screening efforts unveiled a collection of FDA-approved drugs with high predicted binding affinities, thus highlighting novel candidates for drug repurposing in COVID-19 treatment. By offering insights into the binding interactions of existing drugs with the SARS-CoV-2 Mpro, their model provides a foundation for future *in vitro* and *in vivo* studies to expedite the development of effective treatments for this global health concern.

Tumor mutational burden (TMB) has emerged as a critical biomarker for cancer prognosis and immunotherapy response. While whole-exome sequencing (WES) has traditionally been employed for TMB estimation, targeted panel sequencing has gained popularity for its cost-effectiveness and efficient coverage of clinically relevant genes. However,

the accuracy of TMB estimation using targeted panels remains a subject of investigation and debate. Therefore, Li *et al.*[9] report a comprehensive simulation-based analysis to evaluate the performance of TMB estimation by targeted panel sequencing. By leveraging extensive *in silico* simulations and the cancer samples and mutation profiles in The Cancer Genome Atlas (TCGA), they compared TMB estimation performance based on WES and targeted panel sequencing and found that panel size and genes are the major factors impacting TMB estimation. This research sheds lights on the optimal configurations for reliable TMB assessment using targeted panels, providing a foundation for improving the clinical utility of TMB as a predictive biomarker in cancer management.

This editorial reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

REFERENCES

1. Guo W, Liu J, Dong F, Song M, Li Z, Khan KH, Patterson TA, Hong H. Review of machine learning and deep learning models for toxicity prediction. *Exp Biol Med* 2023
2. Khan KH, Guo W, Liu J, Dong F, Li Z, Patterson TA, Hong H. Machine learning and deep learning for brain tumor MRI image segmentation. *Exp Biol Med* 2023
3. Roberts R. Data science in drug discovery safety: challenges and opportunities. *Exp Biol Med* 2023
4. Wang X, Xu X, Liu Z, Tong W. Bidirectional Encoder Representations from Transformers-like large language models in patient safety and pharmacovigilance (PSPV): a comprehensive assessment of causal inference implications. *Exp Biol Med* 2023
5. Wu L, Gray M, Dang O, Xu J, Fang H, Tong W. RxBERT: enhancing drug labeling text mining and analysis with AI language modeling. *Exp Biol Med* 2023
6. Le H, Hong H, Ge W, Francis H, Lyn-Cook B, Hwang YT, Rogers P, Tong W, Zou W. A systematic analysis and data mining of opioid-related adverse events submitted to the FAERS database. *Exp Biol Med* 2023
7. Liu J, Xu L, Guo W, Li Z, Khan KH, Patterson TA, Hong H. Developing a SARS-CoV-2 main protease binding prediction random forest model for drug repurposing for COVID-19 treatment. *Exp Biol Med* 2023
8. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* 2008;**48**:1337–44
9. Li D, Wang D, Johann JJ, Jr, Hong H, Xu J. Assessments of tumor mutational burden estimation by targeted panel sequencing, a comprehensive simulation analysis. *Exp Biol Med* 2023