NCTR Division of Bioinformatics and Biostatistics (DBB) Review
Subcommittee Site Visit Review, May 19-20, 2022
Submitted September 15, 2022

Overview

Review Subcommittee and Expertise.

The review took place via Zoom on May 19-20, 2022. The names and affiliations of the site visit reviewers are provided in Appendix 1. Three members (Drs. Ganey, Ramos and Tropsha) of the NCTR Scientific Advisory Board (SAB) participated in the review. Drs. Tropsha and Ganey served as subcommittee chair and co-chair, respectively. Drs. Pariser and Zhang participated as subject matter experts in content areas of interest to the Division of Bioinformatics and Biostatistics (DBB, referred to hereafter as "the Division"). The Subcommittee members received a written overview of their charge in a memorandum dated November 7, 2021, from Tucker A. Patterson., Ph.D., NCTR Deputy Director for Research and Donna L. Mendrick, Ph.D., Designated Federal Official and Associate Director for Regulatory Activities. The charge memo is attached in Appendix 2. Site visit reviewers were provided with project overviews divided into 5 focus areas on May 3, 2022; they listened to presentations by the leaders and scientists from DBB followed by a Q&A session during the meeting and received some additional materials in response to a post-meeting request on May 26, 2022.

Primary and secondary reviewers were assigned to each focus area in advance of the meeting. Reviewer assignments for each focus area are shown in the table below in the order these areas were presented during the review meeting. The reviews of each theme area have been written primarily by the assigned experts and harmonized by the chair and co-chair of the Committee who also co-wrote the introductory paragraphs.

Focus Area	**Reviewer 1**	**Reviewer 2**
Focus Area 1: Regulatory Applications and Support: An Overview	Anne Pariser	Patti Ganey
Focus Area 2: Alternative Methods and Knowledge Bases	Patti Ganey	Alex Tropsha
Focus Area 3: Precision Medicine and Therapeutics	Ken Ramos	Anne Pariser
Focus Area 4: Artificial Intelligence (AI) and Machine Learning (ML)	Alex Tropsha	Hongmei Zhang
Focus Area 5: Real-World Data (RWD) and Real-World Evidence (RWE)	Hongmei Zhang	Ken Ramos

Agenda, Reviewed Materials and Process

The agenda for the two-day site visit (May 19-20, 2022) is shown in Appendix 3. An SAB – DBB booklet was provided to reviewers prior to the site visit. The booklet included the Division Overview, followed by a Summary of the Divisional Research and Support; both sections emphasized changes since the last site visit in 2015. The Summary introduced five Focus Areas that were detailed in the subsequent sections of the Booklet. Finally, the booklet included Biosketches of principal investigators (PIs). Additional documents provided prior to the meeting included the previous Review of the NCTR Division of Bioinformatics and Biostatistics conducted on November 5-6, 2015, and dated 05/10/2016 and the response to this review authored by Dr. Weida Tong, Director of the Division, dated 10/03/2016. Subcommittee members were instructed to review these materials prior to the site visit. Prior to the review meeting, the Subcommittee was also provided with copies of presentation slides and posters.

All members of the Subcommittee were present during both days of the meeting. Prerecorded presentations were shown in the order outlined in the Agenda; each presentation was followed by a live Q&A session where both the presenter and additional members of the leadership of the Division answered questions from the reviewers. A few days after the review meeting, the Subcommittee asked the Division Director, Dr. Weida Tong, via Dr. Mendrick, for additional information by email; specifically, the request was for a summary of (adequacy of) computing capabilities afforded to his group in terms of CPU and GPU support and programming if/as needed. Also, a summary of the transformation of the Division since the last review was requested in terms of both initiation and termination of various projects and personnel size and additional comments on the number of unfilled positions and current efforts to fill these. Dr. Tong, via Dr. Mendrick, quickly answered these questions via email; the answer included pointers to the specific sections in the Booklet addressing these additional questions as well as a more detailed summary of personnel changes since the previous review. Also included in the response was a separate document outlining computational capabilities of the Division and a summary of research projects completed by the Division members since the last review. These documents are included in Appendix 4.

Division of Bioinformatics and Biostatistics Overview

DBB was established in May of 2012, making the timing of this review coincide with the 10th anniversary of the Division. The Division was reviewed previously in 2015, three years into its existence; at that time the Division consisted of three branches, the Bioinformatics Branch (led by Dr. Tong); the Biostatistics Branch (led by Dr. Chen); and the Scientific Computing Branch (led by Mr. Bearden). Following the previous review, the Division established a new branch, Research-to-Review (R2R), led by Dr. Joshua Xu: this branch was established in 2017 from the existing personnel to reflect an important supportive function of the Division to the agency.

At the time of the present review, the Division consisted of four branches: (1) Bioinformatics led by Dr. Huixiao Hong; (2) Biostatistics led by Dr. Dong Wang; (3) Research-to-Review (R2R) as mentioned above; and (4) Scientific Computing led by Mr. Edward "Ted" Bearden. The "immediate office" led by the Division Director, Dr. Weida Tong, was also established. The Division personnel are spread across four branches and the immediate office; it currently includes government staff (research scientists and support staff), postdoctoral fellows, graduate students, and 10 vacancies. The vision of the Division is described as "to be an indispensable resource to

FDA in the areas of bioinformatics and biostatistics", and its mission is to assist FDA in the review process, strengthen linkages with centers, and evolve its capabilities in tune with Agency needs. Division staff are divided roughly 50:50 between working on "research" or "support", which formally relates to different career paths for the personnel: the former is predominantly in the Scientific Computing Branch and the immediate office of the Branch Director; and the latter are distributed across the other three branches.

The Division has undergone substantial changes since the last review. Two senior leaders, Roger Perkins (Senior Advisor and Branch Chief) and Jim Chen (Branch Chief) have retired; a new branch (R2R) was established, and new leaders have been appointed for each branch except for Scientific Computing. The total size (including current vacancies) grew by about 10 positions. In an effort to maintain the active workforce size, the Division increased its support of the ORISE program to recruit more postdocs and is working to establish an institutional agreement with Arkansas State University to recruit graduate students to work on projects of interest to the Division. Establishing the R2R Branch enabled the Division, in fulfilling its mission, to expand its collaborations with multiple FDA centers (CDER, CTP, and CDRH) and ORA. An important recent change highlighted in the DBB presentation was the establishment of the Artificial Intelligence Research Force (AIRForce) in the immediate office of the Division director in 2021; this provides an example of the agile response of DBB to the changing landscape of modern computational toxicology. AIRForce has launched the Artificial Intelligence for Toxicology (AI4Tox) program including four initiatives: AnimalGAN, to develop novel accurate predictors of animal toxicity; SafetAI, to assist in safety review of drug candidates; BERTox, to employ natural language processing tools to accelerate processing of FDA documents and literature to improve toxicity assessment; and PathologAI, to improve the analysis of histopathological data.

General comments for the DBB.

For the purposes of the SAB review, DBB presented the summary of its activities in five focus areas (reviewed in detail below) rather than stratifying the presentation by branches; the activities of the Scientific Computing Branch (reviewed in 2015) were excluded from the review. Overall, the Report has richly reflected on multiple research projects aligned with the overall goals of the Division to develop and apply data analytics, bioinformatics and biostatistical methodologies, and computational modeling to regulatory science research in diverse fields that are critical to the FDA mission. The Division continues to play a major role in advancing collaborative activities across the agency including 10 collaborative projects with six FDA centers. Most of the collaborations are with CDER and the Offices of Translational Science, New Drugs, Computational Science, Pharmaceutical Quality, Medical Policy Initiatives, Generic Drugs, Product Evaluation and Quality, and Clinical Pharmacology. In 2015-2021, members of the Division published a total of about 200 papers (25-35 per year) in journals with an average IF of 7; these papers have been richly cited (14 citations per paper). The Division currently conducts 19 research protocols. The total funding received by the Division both from intramural and extramural sources has been steady; albeit there have been fluctuations over the years. Notably, the intramural funding in 2021 is at its lowest in the last 6 years, whereas extramural funding was one of the highest. Since the last review, scientists within the Division have trained more than 110 students (graduate and

undergraduate), postdocs, and professionals. For its collaborations with other Centers and noticeable publications, the Division has received multiple awards every year (12 total for 2015-2021). In summary, the SAB notes that overall, DBB's achievements have been outstanding in all major areas of its activities including basic research, collaborations, training, and support of other Centers. Still, there are areas of improvement that are reflected upon below in the overall recommendations as well as in the comments for each focus area.

Recommendations for the DBB.

- Overall, DBB is an extremely valuable resource for both FDA and the research community at large, providing important regulatory science tools including important databases, software, and expertise in bioinformatics and computational toxicology. As the importance of data science and AI are increasingly appreciated by both the agency and the external research community, and the demand for DBB expertise is likely to increase, it would be helpful for DBB to establish clearer approaches for project selection and prioritization. The current approach was not made clear to the SAB.
- It is obvious and commendable that DBB has achieved prominent, and growing recognition by multiple Centers within FDA. It would be helpful to outline in greater detail how DBB collaborates with other branches within NCTR.
- SAB supports the efforts of the DBB to establish stronger working relationships with ORISE and Arkansas State University as it enables the Division to effectively recruit additional minds and hands to work on important and interesting projects. It would be advisable to continue in this direction and consider making similar strategic arrangements with selected academic institutions across the US.
- It is clear that research developments by DBB have achieved prominent recognition within FDA as reflected by multiple awards received by DBB scientists. It is important and highly advisable to increase DBB visibility at the national and international levels, increase collaboration with external researchers outside of the Agency, and promote the distribution of databases and tools via specialized publication types (such as Application Notes supported by such journals as *Bioinformatics* and *Journal of Chemical Information and Modeling* and database issue supported by *Nucleic Acid Research*).
- The report provides information about cumulative successes of the Division and average productivity by DBB staff. For the purposes of fair productivity assessment and career advancement, it would be helpful to outline metrics by which DBB staff are assessed individually.
- The Division supports multiple projects important for the field of regulatory toxicology, including the development of multiple computational toxicity prediction models. These models are disseminated via research publications and presentations at scientific meetings. Given the overall DBB and NCTR missions, it would be important to outline specific steps toward making such models into accepted regulatory tools that are used routinely by the agency.
- Because of its demonstrated and well-supported value to the development of regulatory science at FDA, DBB is encouraged to continue its outreach to other Centers, Offices and Divisions within FDA. The work done to date is potentially extensible to other areas that

could open up new areas of collaboration that contribute to both DBB's and FDA's mission.

Comments for individual focus areas.

Focus Area 1: Regulatory Applications and Support.

Overall assessment

An important component of the stated mission of DBB is to "ensure that the Division's activities relate to FDA's review process, our linkages with product centers continue to be strengthened, and our capabilities evolve to meet the FDA's current and future needs." Consistent with their mission and goals, DBB has rapidly initiated and furthered several regulatory application support programs and has played a major role in advancing the adoption of emerging technologies within the FDA review process, training and education of FDA's review staff, and advancement of the ability of FDA to track, summarize and search its information to better inform its work. These efforts have clearly substantially changed the efficiency and transparency of the Agency's work, as well as the ability to capture regulatory scientific advances in near real-time across several FDA Centers. The value of this work to FDA is notable in the large number of FDA awards DBB continues to receive year after year, as well as its dozens of peer-reviewed publications with 100s of citations each year. Specifics on some of the programs and tools DBB has developed are summarized as follows:

Scientific projects.

First, DBB has continued its close collaboration with CDER to develop and support a number of programs of high relevance to CDER's knowledge management. DBB has a long-standing collaboration with CDER's Office of Computational Sciences (OCS) to integrate machine-learning and NLP (natural language processing) to assist information capture from FDA reviews and labeling. This includes the Smart Template System, a structured Microsoft template that uses a deep learning language model for extraction and word searches that was developed by DBB. The Smart Template allows for easy text searching from prior reviews, which are then stored in a searchable database, giving FDA reviewers access to existing regulatory science information with no additional data entry burden because the information is machine-pulled directly from reviews. Similarly, FDALabel is a tool that captures essential scientific information directly from product labeling as an information resource for the public and for FDA staff. FDALabel currently contains over 130,000 drug labels and is accessed more than 8,000 times per year by the public. DBB has additionally transformed the DASH database, which houses detailed regulatory information on new product approvals into a large capacity Oracle database application compatible with FDA's enterprise infrastructure and IT standards that enables a high degree of searchability to provide detailed regulatory information to Agency staff. The success of DASH has led to the development of two other similar regulatory science review tools for Breakthrough Therapy Designation, for promising products to treat serious diseases, and a Safety Policy and Research Team post-market safety relational database to facilitate systematic analyses of safety issues, actions and outcomes. Without these tools, large amounts of information would otherwise be inaccessible to reviewers and managers due to the information being archived in unstructured documents; however, through these programs, FDA staff are able to harness this information to contribute to informed decision making and improve the application process.

The second area of Regulatory Application and Support is the development of the Automated Laboratory Information System (ALIS) for ORA, which is a customizable system able to be

tailored to the specific needs of ORA's high-throughput field labs that test medical products, tobacco and food. DBB is working closely with ORA to develop management and tracking through ALIS for samples and lab analyses intended to improve data quality and integrity, communication across systems and production of analytical data packages in fully electronic formats. DBB has developed prototypes for ORA and is working closely with ORA staff to extend the system to cover other testing domains.

The third area is the development of advanced semantic indexing techniques for the Center for Tobacco Products (CTP). CTP review applications are highly complex, composed of lengthy documents with highly detailed information and diverse terminology. CTP reviewers spend large amounts of time hunting through unstructured text to answer routine questions. DBB is working closely with CTP to develop a novel tool that will use semantic indexing to search complex tobacco applications using an AI-based NLP model with deep search capabilities called ASSIST4Tobacco, which uses machine learning to discern meaningful word relationships with free text reviews. A prototype has been developed that has shown that this approach is feasible.

Fourth, DBB has completed recent projects with CDER's Office of Translational Science to mine Approval letters for important regulatory information that can be captured and analyzed to inform FDA's work, without the need for the extensive data entry that is currently required. The prototype for this project demonstrated the utility of the approach and follow-on projects are underway. DBB also developed a process to fully automate CDER's Risk Evaluation and Mitigation Strategy (REMS) data capture for the Congressionally mandated REMS public website to allow for machine-importation of key information. The REMS website is currently kept updated by a manual process, and DBB's work with CDER will ultimately establish a more cost- and time-efficient process for REMS maintenance. A similar approach was also used for Office of New Drugs (OND) Meeting Minutes from sponsor meetings. FDA holds over 2,500 sponsor meetings each year, with advice in the Minutes appearing in unstructured free text. DBB worked with CDER's OND to use Natural Language Processing and topic models to organize the information in the Minutes to enable easier searching by FDA staff to ensure consistency in responses and identify precedent cases.

Finally, DBB is conducting exploratory projects with FDA Centers including using artificial intelligence to identify research drivers for CDER regulatory science such as areas with unmet needs by applying AI to research documents by category and AI to search patient narratives.

Summary and additional observations:

Overall, DBB has performed a remarkable amount of work on behalf of FDA Centers in a relatively short period of time. Notably, DBB has developed a number of tools to integrate IT and machine-learning/assisted solutions to allow FDA to better search, capture and assess large volumes of information that previously could only be managed through laborious free-text searching and manual capture and maintenance. These tools have demonstrated a high degree of user acceptance and rapid uptake by FDA staff within the Centers showing the importance of DBB's work in strengthening the review process and building efficiency. Given the success of these programs, DBB is urged to proactively find ways to expand this work across more work areas within the Centers to continue to gain IT-enabled efficiencies and transparency within regulatory science and review support. Expansion of these tools and closer integration with other

FDA Offices and Divisions would only serve to make the Regulatory Application and Support Focus Area stronger and even more relevant to the FDA's and NCTR's missions.

Focus Area 2: Alternative Methods and Knowledge Bases

Overall assessment:

The goal of the work being performed in Focus Area 2 is "To advance regulatory science at FDA". This is accomplished by development of a variety of knowledge bases to be used by FDA reviewers, those submitting applications to FDA and, in some cases, the broader scientific community. In addition, alternative approaches are being developed using artificial intelligence (AI) and machine learning (ML). Overall, the work being performed is important and related to the missions of NCTR and FDA, and the quality of the science is high.

Scientific projects:

Four research areas were presented to the Subcommittee: 1) Development of knowledge bases (KBs) for managing curated data and developed models to facilitate alternative methods for safety evaluation and risk assessment, 2) Development of predictive models for drug-induced liver injury (DILI) to support the FDA review process, 3) ML and deep learning (DL) to facilitate alternative animal toxicity testing, and 4) AI as alternative approaches in nanotoxicology – ML or DL.

The first area, the efforts to develop KBs has been productive. Five KBs have been developed and are under refinement: Estrogenic activity database, Endocrine Disruptor KB (EDKB), Liver Toxicity KB (LTKB), NCTR liver cancer database and Tobacco constituents KB. LTKB is a publicly available, mature KB of drugs associated with DILI. DILI continues to be a human health issue as well as a stumbling block for development of new pharmaceuticals. LTKB has been a useful resource for a variety of audiences, including researchers. The Division is encouraged to continue their efforts to incorporate additional factors, including immune and genetic factors, into the database. A relatively recent effort is the development of a database of herbal dietary supplements (HDS) associated with cases of liver injury. DBB is using PharmaPendium, an Elsevier database of historical agency approval documents, to incorporate pharmacokinetic (PK) and adverse effect data to develop predictive models for DILI. This will assist FDA reviewers and scientists. Through a recently approved proposal with the Office of Women's Health, DBB will use AI-based technologies to explore immune system-related risk factors to help explain sex differences (women represent most cases) in HDS-induced liver injury. DBB is encouraged to expand their efforts beyond immune factors.

Also under development is a KB of chemicals and their associated opioid agonist/antagonist activity (OAK) using data generated in house and curated from the public domain as well as other descriptive data. This effort addresses the health crisis surrounding opioid abuse. The goal is for the OAK to provide insights to molecular mechanisms of pain management and treatment of opioid use disorder and also to enable development of *in silico* models for predicting opioid activity. Data points have been collected for over 2800 compounds, and ML algorithms are being used for models to predict activity. This is an important effort that needs further refinement and thought concerning how it could be used to facilitate treatment of pain and opioid use disorder.

DBB is initiating efforts to develop a database of molecules with androgenic activity as a resource for FDA scientists and reviewers. There was no mention of trying to identify motifs within chemical structures that confer activity or build computational models with this data (similar to proposed developments for the OAK database), so the team should be encouraged to pursue this avenue. With respect to the second research area, development of predictive models for DILI, DBB is using the Rule-of-2 and a DILI score to identify inherent hepatotoxic characteristics of drugs under review. They plan to incorporate PK data from PharmaPendium and employ a ML approach to develop a predictive model that will then be validated with failed drug candidates. They also plan to use PBPK modeling to predict some additional PK parameters. The approach is reasonable, although it was not made clear how these models would be used in the review process.

Another effort within this research area is underway to identify genetic and nongenetic factors associated with increased susceptibility to HDS-induced liver injury in women. A scoring model will include data from whole exome sequencing and clinical risk factors and will be validated with data from the Spanish DILI registry. The investigators are aware of the limitations of using this registry for generalization to US populations.

One project was presented related to the third research focus area, developing predictive models as potential alternative methods to animal toxicity testing. This project is in the early stages and, as a pilot, data from guideline animal studies of multigenerational reproductive toxicology were used. The model was reported to have achieved "reasonable predictive power" although accuracy was 60-65%, raising a question of what is considered reasonable. Accuracy can be improved with increasing confidence in a prediction confidence analysis. Future studies will involve developing several DL and ML models, expanding to other toxicity types, and exploring modeling approaches to improve predictive accuracy. It was not entirely clear how the model is intended to be used, but if it is to be used as an alternative to animal testing, the criteria for applicability need to be defined.

With respect to the fourth research focus - using AI for alternative approaches in toxicology - one project was presented comparing ML and DL. The hypothesis was that ML would produce better results with small- to moderate-sized datasets whereas DL would be better for large datasets. It was stated that ML models for predicting CH_4 and CO_2 adsorption capacities of metal organic frameworks were high performing. These datasets are large, and DL models were constructed to compare performance. DL improved performance only marginally. One question that was raised was how large versus small is defined for datasets. The DBB has plans to decrease sample size to evaluate when ML outperforms DL. Other issues raised by the Subcommittee that should be addressed by DBB related to the availability of datasets large enough to require DL, and the criteria to assess improvement using DL instead of ML. There is a natural interest in the computational research community as to when DL techniques should be used to develop better models and when their application is not warranted as the standard ML approaches will provide models of the same or better accuracy as DL-based models. This is an interesting line of investigation that should be pursued further.

Addressed in the written material provided to the Subcommittee were two projects in the early stages. One was the development of a software tool to identify the potential hazard related to endocrine-active materials that can be extracted or can leach from medical devices. The tool is

intended for use during premarket submission of new medical devices. A learning capability will be added to the tool, and the long-range plan is to submit this to the Medical Device Development Tools program for qualification as a nonclinical assessment model for stakeholders. This could be very useful.

The other proposed project is an evaluation of microphysiological systems for the capacity to predict human DILI. These systems would be compared to *in vivo*, toxicogenomic and other *in vitro* methodologies. This is an important effort consistent with the current push to move to new alternative methods (NAMs).

Summary and additional observations.

Overall, work presented under these focus areas has been very intense providing an important contribution to the DBB mission. Further research developments are encouraged especially concerning the expansion of curated knowledge bases and making these accessible to the research community for building/benchmarking respective predictive models for respective endpoints. The work on making models both accessible (continuous sharing via github) should be encouraged and even more so, it is important to emphasize the use of models as NAMs achieving acceptable regulatory tool status, for instance, via registration for the MDDT program.

Focus Area 3: Precision Medicine and Therapeutics

Overall assessment:

Precision Medicine and Therapeutics was identified as one of the primary focus areas of the Division. In keeping with this designation, many of the activities within the Division directly advance priorities in this fast-evolving space, including efforts to develop AI tools to improve toxicity assessments and the identification and validation of immutable vaccine targets in severe SARS-CoV-2 infection. The progress made to date in developing a set of AI models for a broad range of toxicity endpoints critical to drug safety evaluation was significant and likely to generate valuable resources for the agency to address pressing questions related to liver toxicity, carcinogenesis, mutagenicity, and cardiotoxicity. These efforts are well-poised to provide expert guidance and support to the agency. In keeping with the feedback provided during the last review, considerable progress was made in aligning the work conducted by scientists in the Division with the stated mission of NCTR. The progress made since the last review in advancing knowledge to evaluate differences in susceptibility to toxicity and disease, response to pharmacological treatment and adverse drug reactions and in framing these efforts within precision medicine and therapeutics. Lastly, efforts to improve pathology workflows taking advantage of AI-enabled digital pathology platforms are to be commended.

Scientific projects:

Long-standing efforts within the Division, and the first area of interest have focused on the international consortium of Microarray and Sequencing Quality Control (MAQC/SEQC). This consortium was established to address the reliability, reproducibility, and utility of genomics technologies in regulatory use and clinical application, and application of drug repositioning principles to assessing reuse of existing drugs for the treatment of rare diseases and COVID-19.

Now in its 4th phase, MAQC/SEQC2 is focused on the development of quality control metrics, reproducibility, and benchmarking bioinformatics approaches for clinical and regulatory use of whole genome sequencing data, targeted sequencing, and other next-generation sequencing technologies. The Division's contributions to the overall effort are laudable and worthy of special recognition as they will facilitate development of objective criteria and metrics for data assessments in the regulatory setting. The success of this program is evidenced by the number and quality of published works in high quality journals such as Nature Biotechnology and Genome Biology. Proposed plans to engage with PrecisionFDA to further evaluate analytical tools for indel calling following Oncopanel sequencing is of major relevance in precision oncology. However, the degree to which this effort is being coordinated in collaboration with academic stakeholders and other government agencies should be considered.

A second area of focus has been the development of statistical tools for regulating deep sequencing-based testing as it applies to the analytical challenges associated with the application of new approaches such as the PraxisTM Extended RAS Panel, the OncomineTM Dx Target Test, and the MSK-IMPACT assay. The goal is to provide regulatory scientists with resources to evaluate statistical performance characteristics. This project is a collaboration with scientists in CDER, CDRH and the Office of the Chief Scientist. The simulation extrapolation (SIMEX) method coupled with targeted simulation has been tested, but firm conclusions validated with multiple studies have not yet been provided, though a manuscript was published using lung cancer databases. Plans to focus on clinical utility assessment seem appropriate, though the extent to which these efforts will be carried out in partnership with clinicians was not specified.

A third area of focus is therapeutics, with the goal to develop AI methodologies to identify existing drugs as options for the treatment of current and future pandemics. Given the pressing national needs, this has been a major area of activity for the Division during the past two years. The Medical Countermeasures Initiative (MCMi) project, in collaboration with more than 20 scientists from other FDA Centers, and another two internal NCTR projects, created several AI frameworks capable of rapidly identifying potential therapeutic options from FDA-approved drugs or from public domain sources, testing of drugs for COVID-19 and algorithms to clarify measurement error rates. Plans to continue these efforts are in keeping with Agency and national priorities and are strongly supported by the review committee.

A fourth area of focus is drug repositioning for rare diseases taking advantage of AI-powered frameworks. This project has resulted in ten papers published in high quality peer-reviewed journals and one of them receiving the FDA Chief Scientist Publication Award for Basic, Translational, and Applied Science. In addition, the FDA-NCATS Translational Science Interagency Fellowship selected the project for continuous funding. Four remaining activities were listed in the report to further explore AI solutions for enhanced prediction, incorporation of a drug safety perspective, web-lab data use for verification and RWD for further evaluation. It is not clear if these will be activities listed to bring to closure this area of focus or if plans are in place to continue these efforts into the future. As noted in the previous review, it is also not clear how this overall effort fits into the mission of NCTR and FDA.

A fifth area of focus is the study of early signs of sex differences in adverse drug events. This appears to be a relatively modest investment of effort to use a bioinformatics approach to facilitate review and inform clinical trials.

Summary and additional observations

In summary, research programming in the precision medicine and therapeutics focus area is well-aligned with division priorities and well-poised to address important questions of high relevance to the agency. Dr. Weida Tong continues to provide outstanding leadership for his team and to freely share his expertise with the scientific community at large. Clear evidence was provided of the collaborative nature of the program.

Focus Area 4: Artificial Intelligence (AI) and Machine Learning (ML)

Overall assessment

The Division has had the focus on the development and use of ML algorithms for a long time, even before the Division was formally organized. Recent growth of chemical toxicology datasets and the development of DL approaches naturally stimulated additional work by the Division in these areas of research. As noted in the DBB booklet, AI offers the FDA opportunities and challenges in two distinct functions: (1) evaluation and regulation of AI-centric products, and (2) implementation of AI techniques to improve and enhance the efficiency of the agency's regulatory operations. DBB has naturally focused on the latter area, and many application projects presented and reviewed as parts of other focus areas have benefited from basic developments in focus area 4. Projects described under this focus area are highly innovative; they are briefly reviewed below.

Scientific projects:

The first area presented by DBB was on AI-Based NLP for FDA Labeling Documents. This new area of research relying on modern NLP protocols has been developed to assist FDA reviewers in their highly laborious and mostly manual tasks to review submissions and provide concise summaries of findings in these submissions. The use of automated text processing tools is more than justified and the DBB research team is certainly on the right track with this project. As acknowledged by the team, this effort requires novel specialized language models, so the ongoing efforts to develop RxBERT in support of the FDALabel project are likely to be rewarding and the resulting tools will be in high demand.

Similar effort relying on NLP is also ongoing in the second area of research on DeepReview, an NLP Powered Information Retrieval System. This project addresses the broader task of text mining of various submission documents received from sponsors. The proposed developments have a high potential to make the review process more robust as the new system will enable the review of new documents in the context of historical relevant documents. Notably, DeepReview has already garnered a great deal of interest from CDER so the broadening use of this tool across the agency is expected. Of a minor comment, dividing the NLP-based developments into two separate areas appears somewhat artificial so it would be advisable to integrate these two projects under the same NLP umbrella.

A third area covered during this SAB meeting was on the development of SafetAI, which is a CDER initiative to enhance the IND review process. This area is a continuation of the traditional expertise of the team in developing QSAR models for important toxicity endpoints but now with DL methods. The team is also proposing to enrich current models that have been built historically with chemical descriptors by developing a new modeling framework where instead of chemical descriptors, the investigators will use "model-level representation from conventional ML classifiers into a DL framework. This description is somewhat unclear; hopefully future studies and publications will help establish both the utility of these models and their advantages (of any) as compared with traditional ML techniques.

The fourth area represents an effort toward achieving eXplainable AI (XAI) in Regulatory Science. This is an interesting and important area of investigation, especially in regulatory science where transparency of models, especially statistical models, is always preferred. Deep learning approaches that are widely employed in AI methods, are notorious for the lack of transparency and clear interpretability so the effort toward explainable AI is understandable. While the importance of this direction of research is obvious, the description of the proposed effort does not seem to address the complexity of DL model interpretation in terms of feature significance. Preliminary observations presented by DBB suggest that simple models (such as linear regression) may offer the best balance between accuracy and explainability. It appears that it will be important to address the utility of DL models for various endpoints as compared with simpler models and then contemplate DL model explainability only for those cases when the use of DL models is justified.

In addition to four areas discussed above, the DBB team also shared their plans for several additional programs including the development of PathologAI Framework to Support Preclinical Digital Pathology, DeepOCR for Scene Text Detection, and AI for Causality Assessment of FAERS Reports, to employ modern text mining to elucidate causal relationships between a drug product and adverse event. All three projects are scientifically appealing as they address important challenges faced by the agency.

Summary and additional observations

Overall, this area of research continues to be one of the major components of the DBB research portfolio. Importantly, the Division continues to stay at the forefront of using modern data analytical methods as applied to problems in computational and regulatory toxicology. Several novel initiatives relying on advanced NLP and DL approaches are noteworthy. It will be important in the next few years to establish the relative value of these new approaches as compared to more conventional ML methods as well as achieve the broad use of novel document review tools across the agency. It will be also important to continue to promote databases and tools created by the DBB across research community at large.

Focus Area 5: Real-World Data (RWD) And Real-World Evidence (RWE)

Overall assessment

RWD and RWE play an important role in guiding priority health care decisions for the agency. This is a relatively new focus area for DBB. Projects have been carried out by the Biostatistics branch led by Dr. Dong Wang. Dr. Wang's research focuses on novel statistical and ML methods in bioinformatics and public health, including predictive toxicology, biomarker development based on next generation sequencing technology, integrative models for drug adverse effects, RWD and RWE. Dr. Wang received an FDA intramural research grant in 2021 from the Office Of Minority Health and Health Equity (OMHHE) focusing on AI and RWD-related methods applied to the study of ethnic and racial disparities in critical care delivery . Since he joined the Division in 2020, Dr. Wang has led several projects in the area of RWD and RWE. His significant contributions to this focus area are noted.

The Biostatistics branch currently has eight employees (one vacancy), including six research scientists, Drs. Wang, Baitang Ning, Paul Rogers, Wei Zhuang, Zhiyuan Lu, and Dongying Li. The role of the research scientists in this branch focuses on research rather than supporting other studies carried out at NCTR, as noted in the Division's booklet. This is a change compared to the 2015 review of this branch. It is unclear whether this is a shift of focus for this branch, and if so, what the underlying motivation was.

Although the number of publications generated by this branch was not clearly articulated, the progress of the OMHHE project awarded in 2021 and of the other projects is highly promising. Four specific projects (including 2 posters) were reported at the review: 1) Developing a Charlson comorbidity index for the American Indian community: The Strong Heart Study (led by Dr. Rogers), 2) Detection and mining of prescription opioid use-associated safety signals from electronic medical records (led by Dr. Zou, Bioinformatics Branch), 3) Integrative profiling of microRNA expression in association with drug-induced liver injury in rats (led by Dr. Dongying Li), and 4) Investigation of racial/ethnic disparities in critical treatment to heart failure patients with propensity score methods (led by Dr. Dong Wang). Since the OMHHE award was received only about half year ago and RWD/RWE is a relatively new focus area, it is understandable that these projects are still on-going.

Scientific projects

The first area presented at the review by Dr. Paul Rogers focused on developing a Charlson Comorbidity Index (CCI) for American Indians, mCCI-AI. The motivation of the project was the lack of such an index for American Indians. The construction of mCCI-AI is an extension of Charlson's method which was built upon a Cox proportional hazards model with random family effects addressed. Although the Subcommittee agreed that there was a potential need for such an index, compared to CCI, the advantage of mCCI-AI was not clearly emphasized and the quality of the newly developed index, mCCI-AI, was not discussed.

The second area presented by Dr. Wen Zou focused on prescription opioid-use-associated safety signals. The project had several stages starting from a systematic literature review, followed by data retrieval from different sources and data mining and deep learning. Findings from a systematic

review was presented at the 2021 IEEE International Conference. The projects are of interest and relevant to FDA missions. It will be interesting to see whether the findings from the retrieved data are consistent with findings in the literature.

The third area was delivered via a poster by Dr. Dongying Li. It is a bioinformatics project focusing on miRNA biomarker detection. This research program addresses an important area and has made significant progress in identifying important leads for future development. The ability of the team to identify specific targets for future development will be critical to ensure continued success of these efforts.

The fourth area utilized propensity scores to assess the impact of racial/ethnic factors on critical care delivery. Using propensity score to match has a potential of overfitting. It will be informative to see the robustness of the propensity scores inferred in the study in terms of matching/weighting and the final conclusion drawn for associations between ethnicity and critical care delivery.

Summary and additional observations

Overall, the projects incorporated advanced approaches and bioinformatics tools in data analyses. It would be helpful to the reviewers if the Branch Chief Dr. Wang or each presenter, could have linked each project with FDA's overall strategic plans and hierarchy of FDA's priorities.

Final notes.

It is noted that the budget provided by FDA to DBB was significantly reduced in year 2021 ($500K in 2021 compared to >$1milion in 2020) and the reason for this reduction was not discussed. However, compared to the year of 2015, the number of staff members appears to have increased (to ~60 from ~50). New focus areas have been added to the Division to fit FDA's new needs. Also, a large amount of data including high dimension high throughput omics data require AI-techniques to process. All these seem to support the necessity of recruiting more manpower to accomplish all objectives and challenges faced by DBB. Related to the need of manpower, we also noticed that there are a number of vacancies in DBB (currently, 5 vacancies not including the Scientific Computing Branch). The Division has taken actions to ease this condition, e.g., recruiting from local universities through various programs. On the other hand, a stronger retention plan from FDA as an institute will certainly help the Division maintain a strong and productive team.