Leveraging Artificial Intelligence (AI) and Machine Learning (ML) to Support Generic Drug Development and Regulatory Efficiency

Meng Hu

Division of Quantitative Methods and Modeling (DQMM)
Office of Research and Standards
Office of Generic Drugs
CDER | U.S. FDA

September 15, 2022

Disclaimer

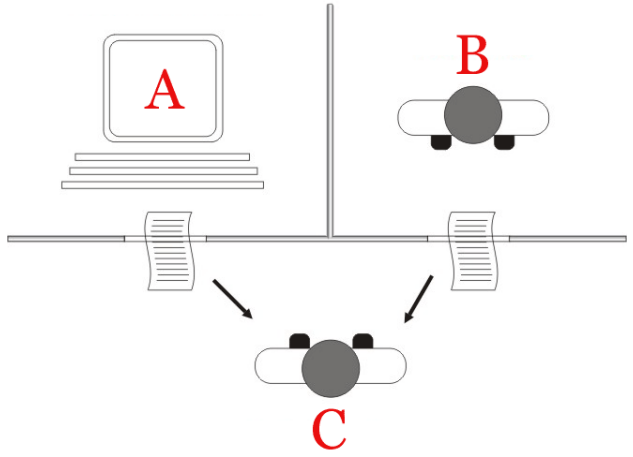
- This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

Outline

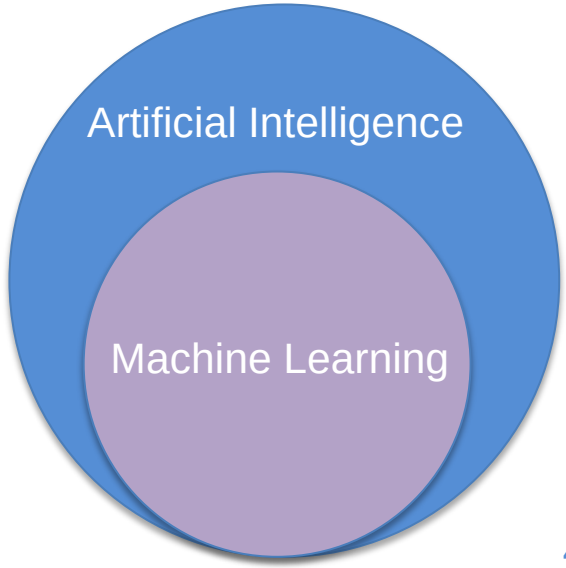
- AI is everywhere
- AI offers opportunities to facilitate generic drug development and regulatory assessment
- DQMM's efforts to answer the opportunity call
- Highlighted case examples
- Takeaways

AI is everywhere

- What is AI?
  - *Turing Test* proposed by **Alan Turing** in 1950



  - *"It is the science and engineering of making intelligent machines, especially intelligent computer programs."* by **John McCarthy** in 2004



AI is everywhere – continued

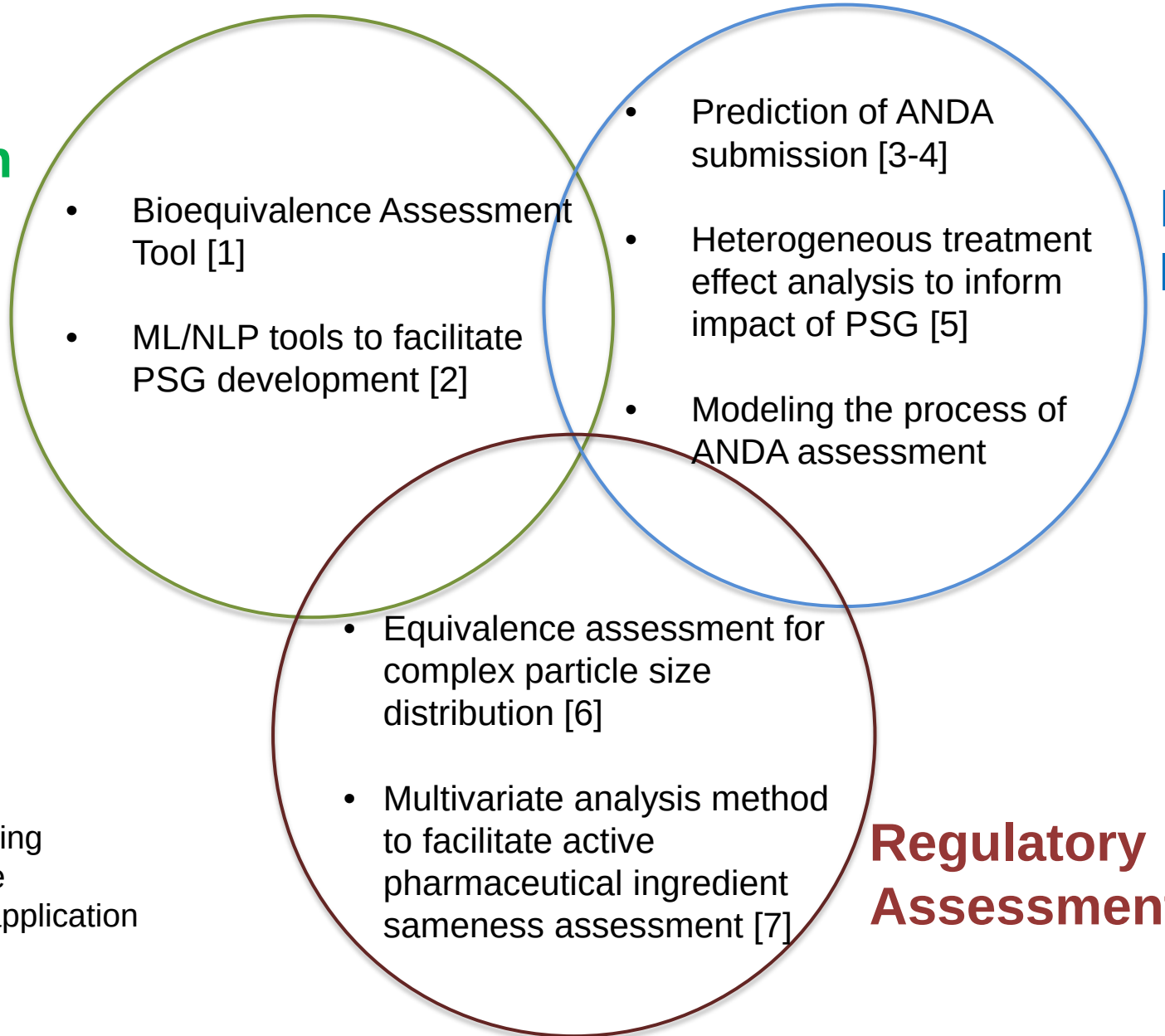
- The thriving AI community should thank the advances in information technologies and powerful chips
  - –Big data
  - –Data analytics tools (e.g., ML and natural language processing (NLP))
  - –HPC (High-performance computing) / Cloud computing
- AI is transforming many areas of everyday life
  - –Smartphone (e.g., face recognition)
  - –Autopilot
  - –Chatbot
  - –Personalized recommendation
  - –ATM (automated teller machine)

AI offers opportunities to facilitate generic drug development and regulatory assessment

- Development of automation tools:
  - Enhanced efficiency (e.g., saving time)
  - Improved consistency (e.g., reducing human error)
  - High-quality deliverables

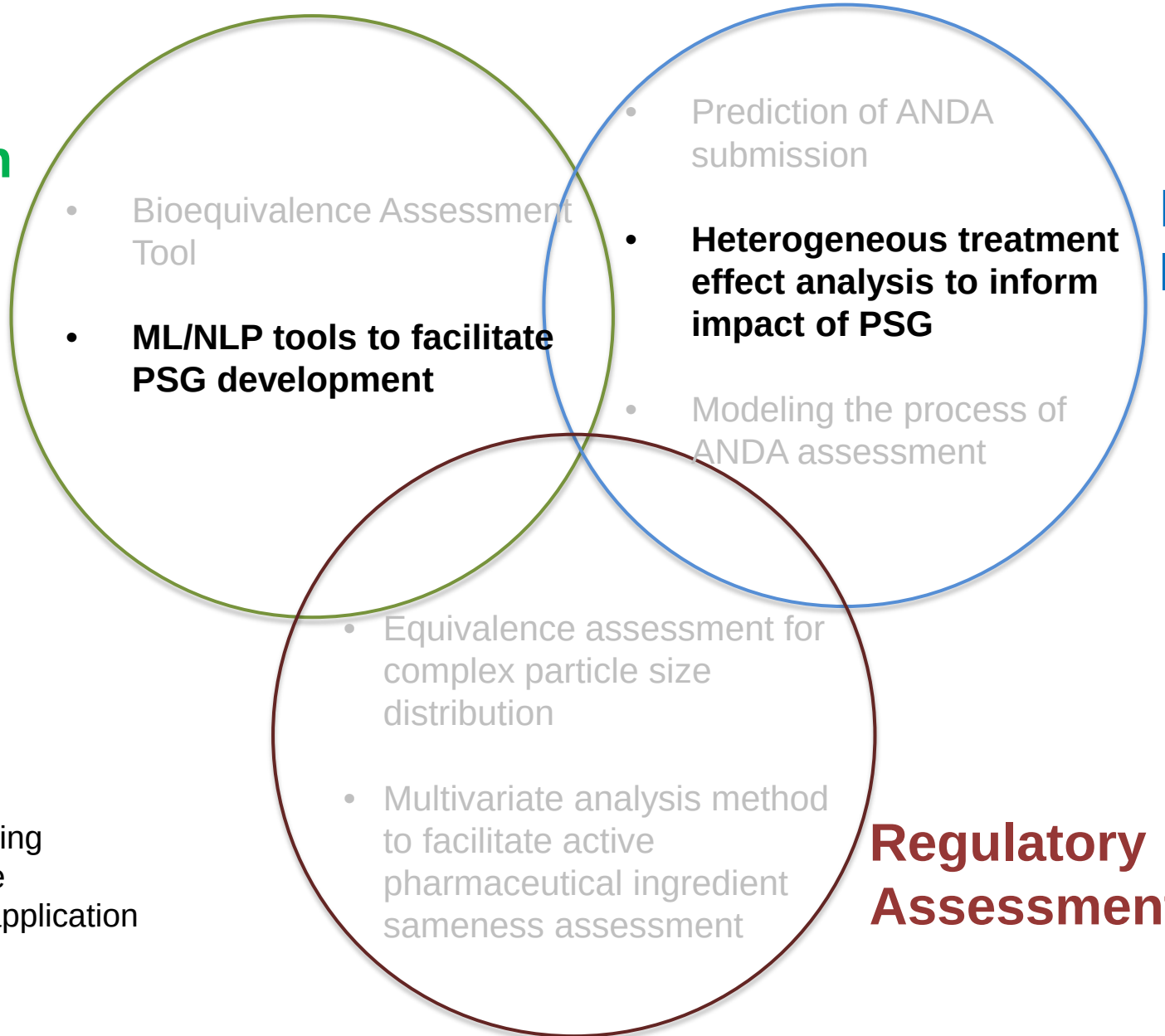
- Utilization of advanced data analytics methods:
  - Promoting business intelligence
  - Supporting regulatory assessment

DQMM's efforts to answer the opportunity call



ML: machine learning
NLP: natural language processing
PSG: product-specific guidance
ANDA: abbreviated new drug application

DQMM's efforts to answer the opportunity call



ML: machine learning
NLP: natural language processing
PSG: product-specific guidance
ANDA: abbreviated new drug application

Highlighted Case Study (I)

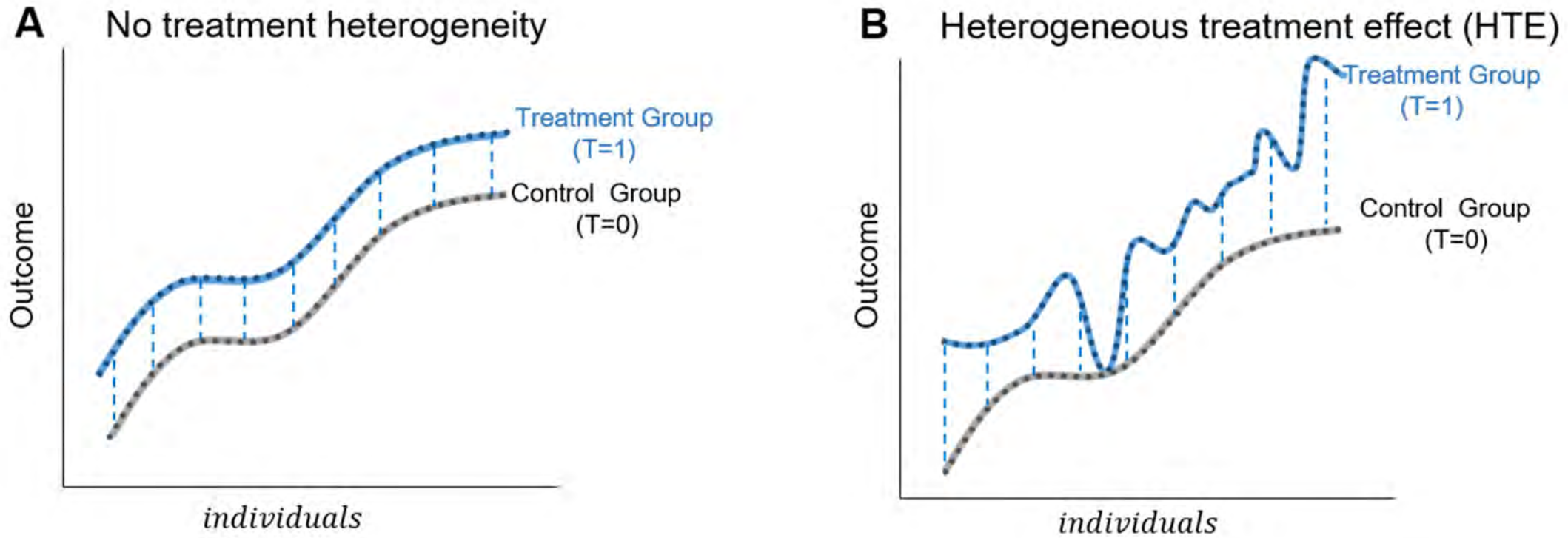
Heterogeneous treatment effect analysis based on machine-learning methodology

Project motivation

- Heterogeneous treatment effect (HTE) analysis has drawn growing attention in a variety of fields from economics to medicine.
- In the Big Data era, dramatically increased data volume and complexity pose significant challenges for HTE analysis, especially using conventional methods.
- Recently, ML methodologies have been employed in HTE analysis and show their merits in handling complex data, given their nature of no assumption on data.
- To introduce this advancement to the community (of pharmacometrics), we conducted a systematic performance evaluation for the ML method against the conventional method by simulating various complex scenarios.

What is HTE analysis?

- Compared to the *response analysis* that predicts the outcome itself, HTE analysis focuses on examining varying treatment effects for individuals or subgroups in a population, e.g., for personalized medicine.



One unique challenge is that the treatment effect is often not explicitly observed on given data, as each subject can often only be exposed to one of the treatments, which is also known as the **fundamental problem of causal inference**.

Current Methods

- Conventional method
  - **Two-step model**
- ML-based methods
  - Tree-based
  - Forest-based (**Causal forest**)

	Pros	Cons
Two-step model	• Intuitive • Easy to implement	• Not necessarily an accurate model; • Often based on ordinary regression models; • Sufficient domain knowledge needed to introduce interactions between variables
Causal forest	• Capable of handling complex practical problems; • Developed to overcome issues of single-tree method	Issues inherited from ML methodologies, such as: • Blackbox • Dependence on quality of data

Simulations

- Interactions between treatment and covariates of subjects often lead to HTE among the study population.

$$Y_i(T) = f(X_i) + g(X_i)T + \varepsilon$$

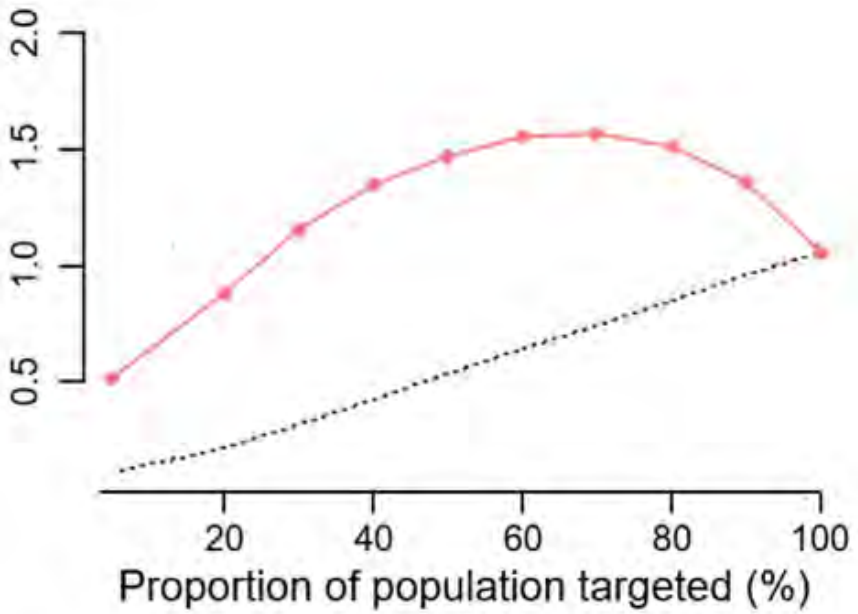
- The below four Models (I - IV) simulating scenarios with different levels of HTE complexity were used to fully characterize model ability in identifying effect heterogeneity [5].

TABLE 1 Summary of the four simulation mathematical models generated with increasing heterogeneous treatment effect complexity

Model	Description of relationships between heterogeneity covariates	Outcome model
I	No heterogeneous treatment effect	$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_k x_{ik} + \delta T + \varepsilon$
II	Linear	$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_i x_{ik} + \left(\gamma_0 + \sum_{k=1}^{p} \gamma_k x_{ik}\right) T + \varepsilon$
III	Nonlinear + interactive	$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i + \left(\gamma_0 + \gamma_1 x_{i1}^3 + \gamma_{23} \cos(x_{i2}) x_{i3}\right) T + \varepsilon$
IV	High-dimensional covariates	Model II

Performance evaluation

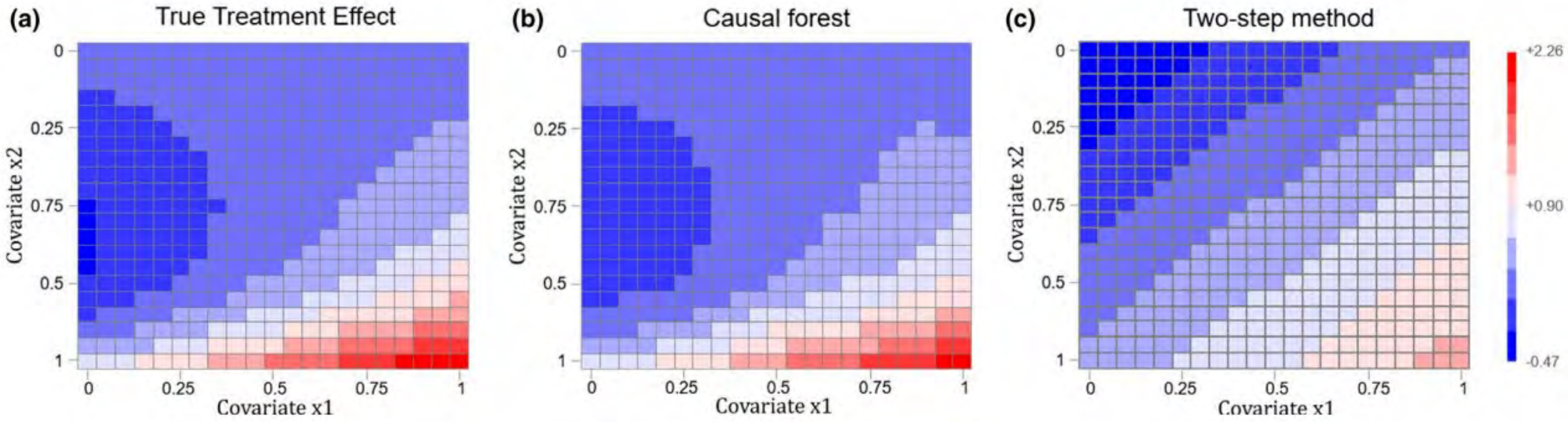
- Root mean square error (RMSE)
- Incremental gains curve ("Qini curve")



Results

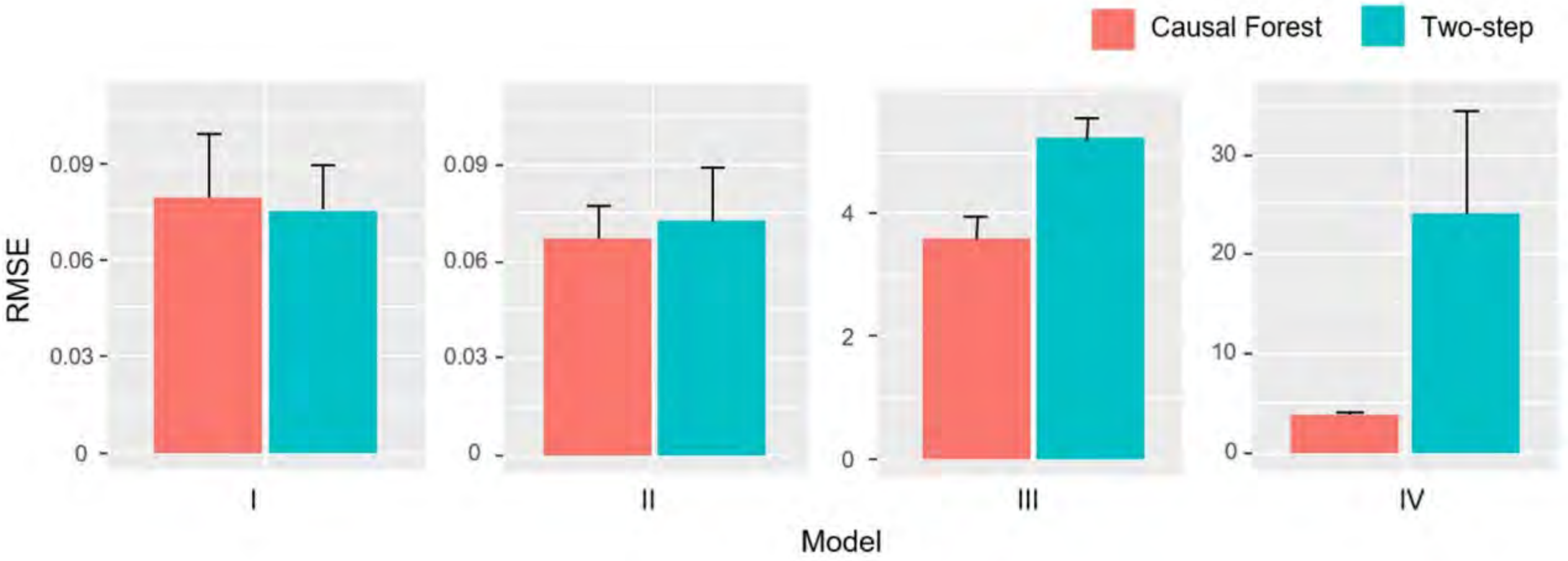
- Case example – a nonlinear model :

$$Y = \beta_0 + \sum_{k=1}^{10} \beta_k x_k + (\gamma_0 + \gamma_1 x_1{}^5 + \gamma_2 e^{x_2} + \gamma_{12} x_1 x_2)T + \varepsilon$$



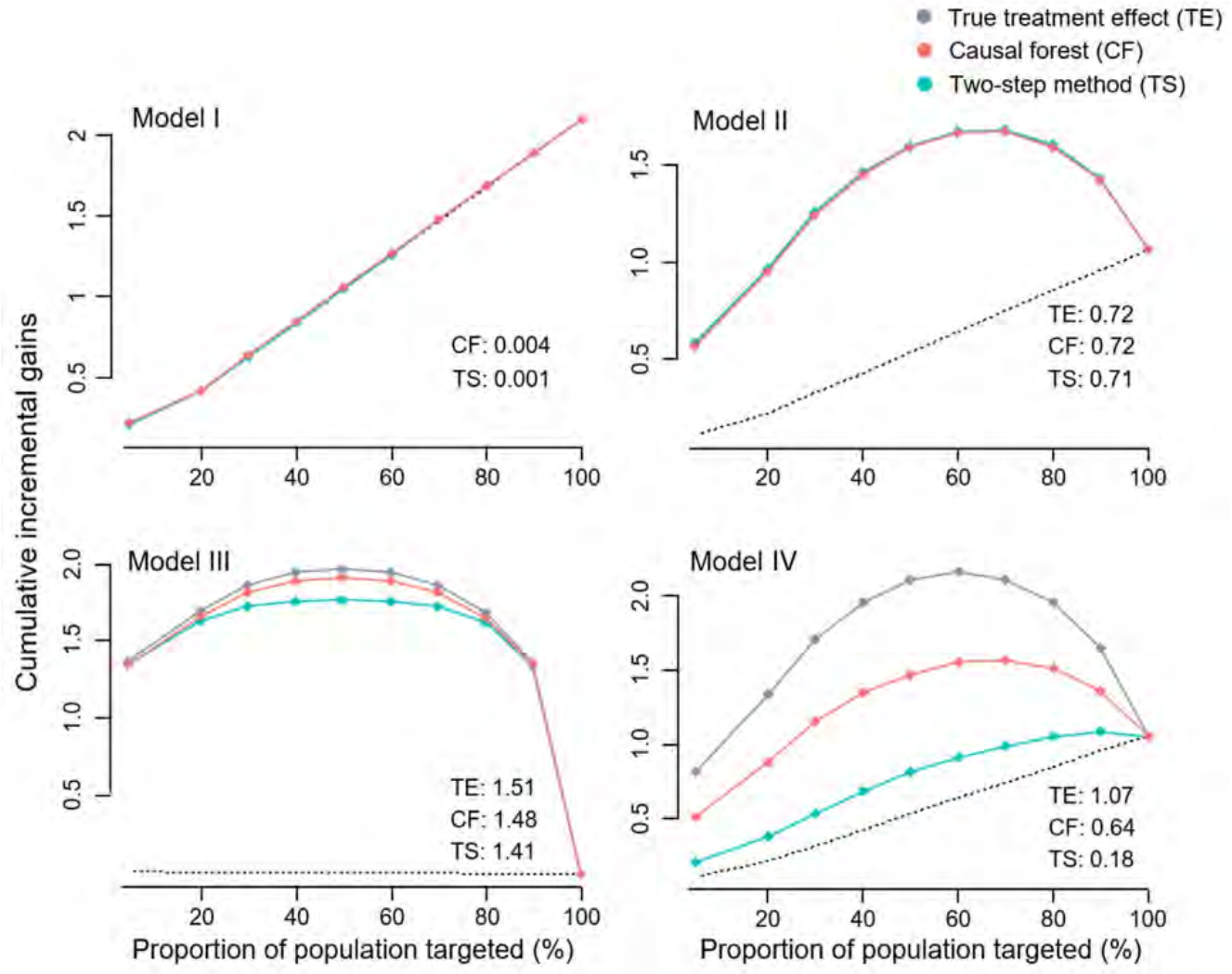
Results

- RMSEs from Models I-IV



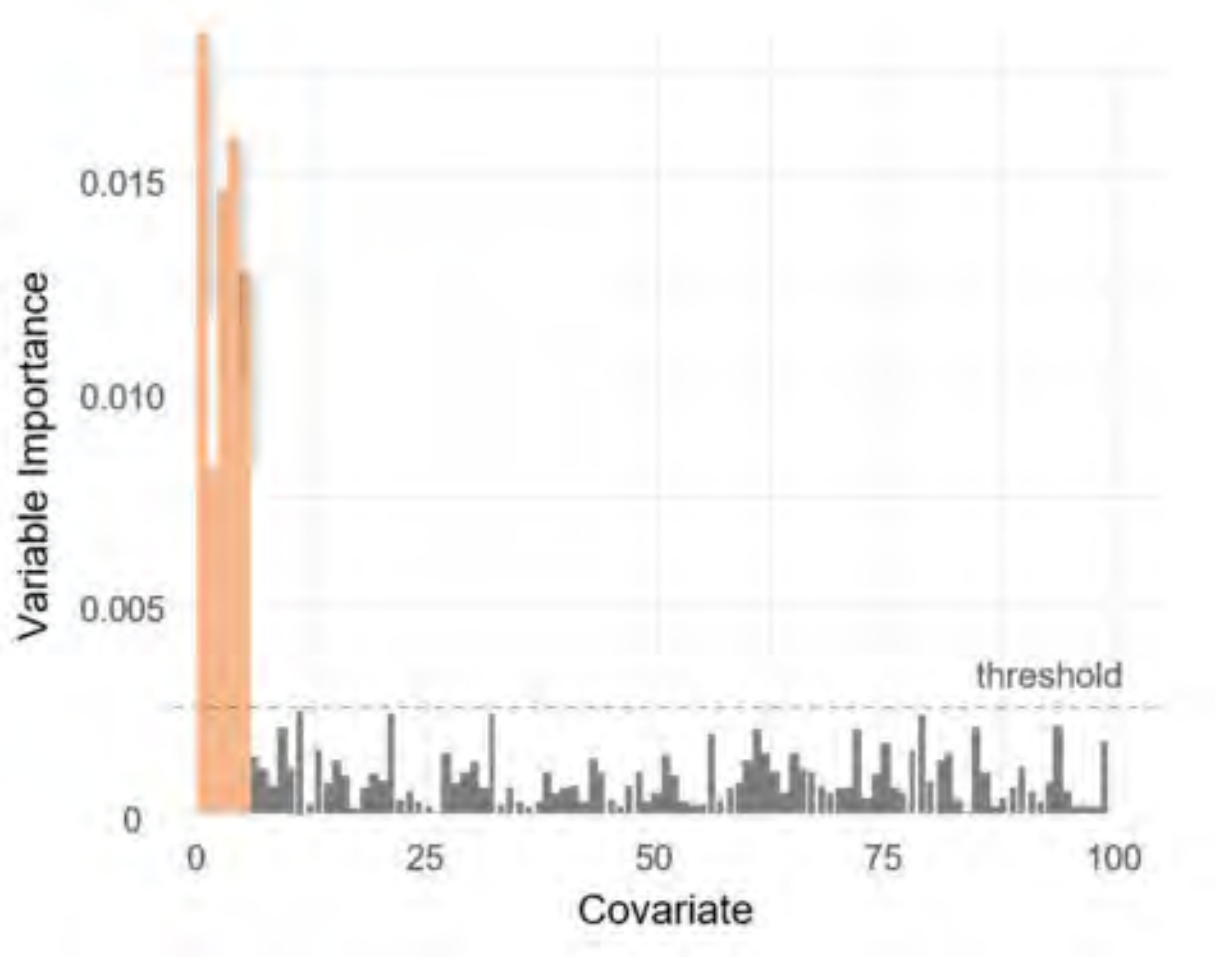
Results

- Qini curves from Models I-IV



Results

- Variable importance
  - Model IV (high dimension)



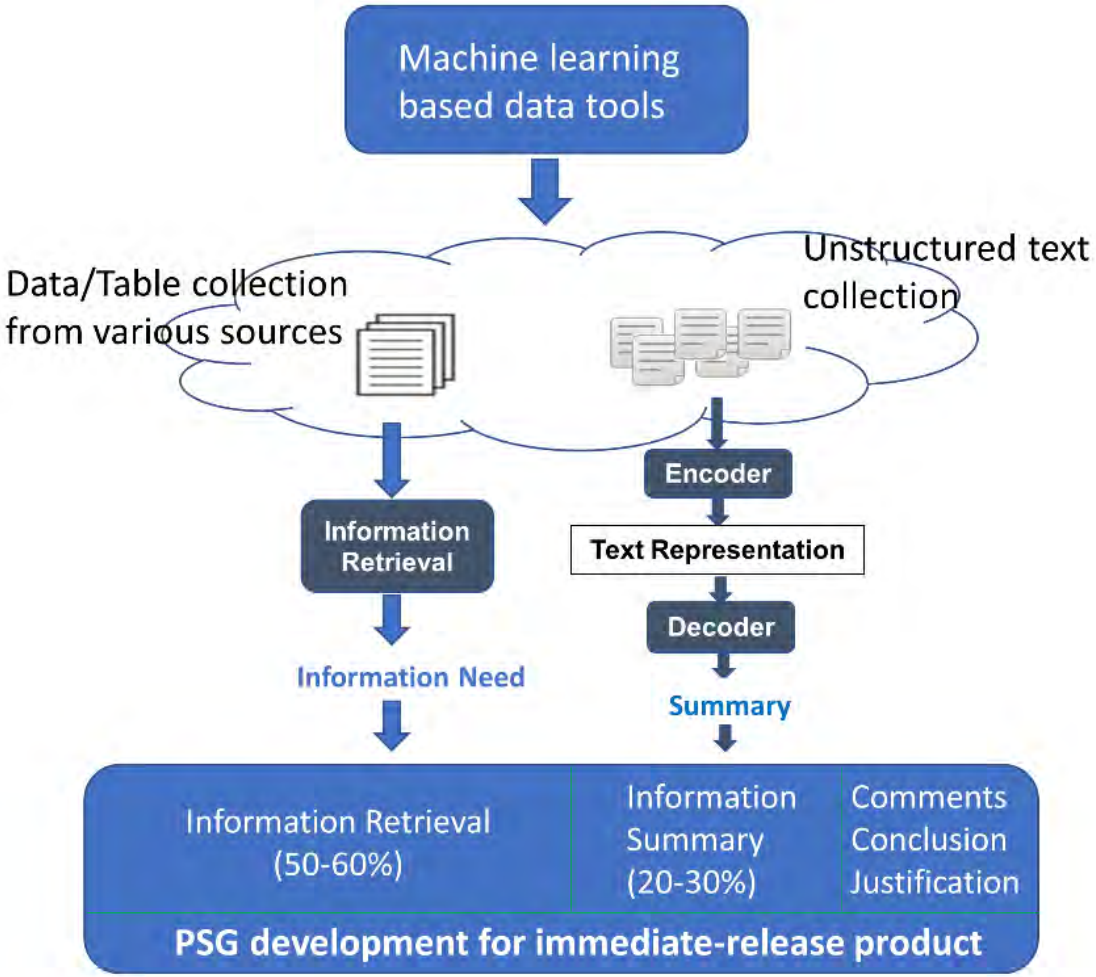
Discussion

- Causal forest, an ML-based method, is a promising tool in real-world applications for HTE analysis.
- It can potentially be used to improve business intelligence in the Agency.

Highlighted Case Study (II)

Text Analysis and Machine Learning to Facilitate Product-Specific Guidance Development

Project overview



Challenges:

- Evolving layouts of source documents (e.g., drug labeling and internal review documents).
- Need for information retrieval based on semantic understanding.
- Capturing information from unstructured text (e.g., reviewer's analysis/comments in review documents)
- Choosing a proper NLP model

www.fda.gov
https://www.fda.gov/drugs/generic-drugs/science-research

Sematic understanding-based information retrieval from drug labeling for "food effect"

EXAMPLE 1: NDA 205832

Absorption

Nintedanib reached maximum plasma concentrations approximately 2 to 4 hours after oral administration as a soft gelatin capsule under fed conditions. The absolute bioavailability of a 100 mg dose was 4.7% (90% CI:3.62 to 6.08) in healthy volunteers. Absorption and bioavailability are decreased by transporter effects and substantial first-pass metabolism.

After food intake, nintedanib exposure increased by approximately 20% compared to administration under fasted conditions (90% CI: 95.3% to 152.5%) and absorption was delayed (median t_{max} fasted: 2.00 hours; fed: 3.98 hours), irrespective of the food type.

https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/205832s000lbl.pdf

EXAMPLE 2: NDA 210491

Absorption

After a single dose in healthy subjects in the fed state, tezacaftor was absorbed with a median (range) time to maximum concentration (t_{max}) of approximately 4 hours (2 to 6 hours). The median (range) t_{max} of ivacaftor was approximately 6 hours (3 to 10 hours) in the fed state.

When a single dose of tezacaftor/ivacaftor was administered with fat-containing foods, tezacaftor exposure was similar and ivacaftor exposure was approximately 3 times higher than when taken in a fasting state.

https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/210491lbl.pdf

Note: keyword searching does not work for this task. For example, searching for "food" will lead to a high false positive rate.

Current progress

- The state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) model was used for this NLP application.
- An NLP pipeline was developed to extract drug product information (e.g., ADME information) from drug labeling with minimal human intervention.
- Published a paper on automatic ADME information retrieval from drug labeling (*Frontiers in Research Metrics and Analytics*) [2].

Takeaways

- AI technologies:
  - bring opportunities to advance development and regulatory assessment of generic drugs.
  - have been applied to facilitate BE assessment, PSG development, business intelligence and regulatory assessment, etc.
  - will play more important role in providing high-quality generic drugs for U.S. public as more challenges get addressed.


FDA
U.S. FOOD & DRUG
ADMINISTRATION

References

1. Hu, M. 2021. BE ASSESSMENT MATE (BEAM) - A Data Analytics Tool to Enhance Efficiency, Quality, and Consistency of Bioequivalence Assessment. Presented at: 2021 FDA Science Forum https://www.fda.gov/science-research/fda-science-forum/2021-fda-science-forum-agenda

2. Shi Y, Ren P, Zhang Y, Gong X, Hu M, Liang H. Information Extraction From FDA Drug Labeling to Enhance Product-Specific Guidance Assessment Using Natural Language Processing. *Front Res Metr Anal*. 2021 Jun 10;6:670006.

3. Gong X, Hu M, Zhao L. Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis. *Clin Transl Sci*. 2018 May;11(3):305-311.

4. Hu M, Babiskin A, Wittayanukorn S, Schick A, Rosenberg M, Gong X, Kim MJ, Zhang L, Lionberger R, Zhao L. Predictive Analysis of First Abbreviated New Drug Application Submission for New Chemical Entities Based on Machine Learning Methodology. *Clin Pharmacol Ther*. 2019 Jul;106(1):174-181.

5. Gong, X, Hu, M, Basu, M, Zhao, L. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT Pharmacometrics Syst Pharmacol*. 2021; 10: 1433– 1443.

6. Hu M, Jiang X, Absar M, Choi S, Kozak D, Shen M, Weng YT, Zhao L, Lionberger R. Equivalence Testing of Complex Particle Size Distribution Profiles Based on Earth Mover's Distance. *AAPS J*. 2018 Apr 12;20(3):62.

7. Rogstad S, Pang E, Sommers C, Hu M, Jiang X, Keire DA, Boyne MT 2nd. Modern analytics for synthetically derived complex drug substances: NMR, AFFF-MALS, and MS tests for glatiramer acetate. *Anal Bioanal Chem*. 2015 Nov;407(29):8647-59.