Prediction of the first ANDA submission for NCEs utilizing machine learning methodology

Meng Hu, PhD
Division of Quantitative Methods and Modeling,
Office of Research and Standards, Office of Generic Drugs, CDER, FDA

FDA Public Workshop
October 3, 2017

Disclaimer

The opinions expressed in this presentation are those of the speaker and may not reflect the position of the U. S. Food and Drug Administration

Motivation

- Prediction of the first ANDA submission will play a significant role in enhancing the application review process.
  - Prioritize research efforts and product-specific guidance (PSG) development, especially for complex products
  - Optimize resource allocation (e.g., expected pre-ANDA interactions)

Scope

- In this study, we attempt to predict the first ANDA submission time for a reference-listed drug (RLD) of new chemical entity (NCE) , based on machine learning methodology.

- NCE
  - An NCE means a drug that contains no active moiety that has been approved by the FDA in any other applications submitted under section 505(b) of the Act.
  - A 5-year period of exclusivity is granted to NCE drugs, which means that no ANDA may be submitted during the 5-year exclusivity period.
  - An ANDA may be submitted 1 year prior to the NCE exclusivity expiration if they contain a certification of patent invalidity or noninfringement.
  - For ANDAs referencing an NCE, the *earliest lawful submission date* is one year before the exclusivity expires.

Hypothesis

- The following factors are hypothesized to be correlated with the first ANDA submission time for an NCE:

  - ❑ Market value (e.g., sales) [+]
  - ❑ Complexity of the RLD product (e.g., complex API and/or complex dosage form) [-]
  - ❑ Availability of PSG [+]

[+] means that the presence or increase of this factor will facilitate the earlier first ANDA submission.
[-] means that the presence or increase of this factor will hinder the first ANDA submission.

Data Collection

- A comprehensive list of all FDA-approved NCEs was generated.

- For each NCE, the corresponding variables were collected in the following 3 categories:
  - ❑ Drug product information
    e.g., Dosage form and Complexity of product
  - ❑ Regulatory information
    e.g., Earliest lawful submission date and PSG availability before first ANDA submission
  - ❑ Pharmacoeconomic information
    e.g., Sales before the first ANDA submission

Variables of Interest

Time to first ANDA submission
= *Actual first ANDA submission date* - *Earliest lawful ANDA submission date*

❑ Drug product information

- Simple dosage form (solution based)
- Complex dosage form (non-oral)
- Oral modified-release
- Complex API (e.g., peptide, polymers, heparin)
- Locally-acting gastrointestinal
- Containing nanomaterials
- Drug-device combinations
- Abuse-deterrent formulation
- Long acting injectable
- Anatomical Therapeutic Classification (13 categories)
- For acute or chronic disease
- Route of administration (5 categories)
- With a Risk Evaluation and Mitigation Strategies (REMS)

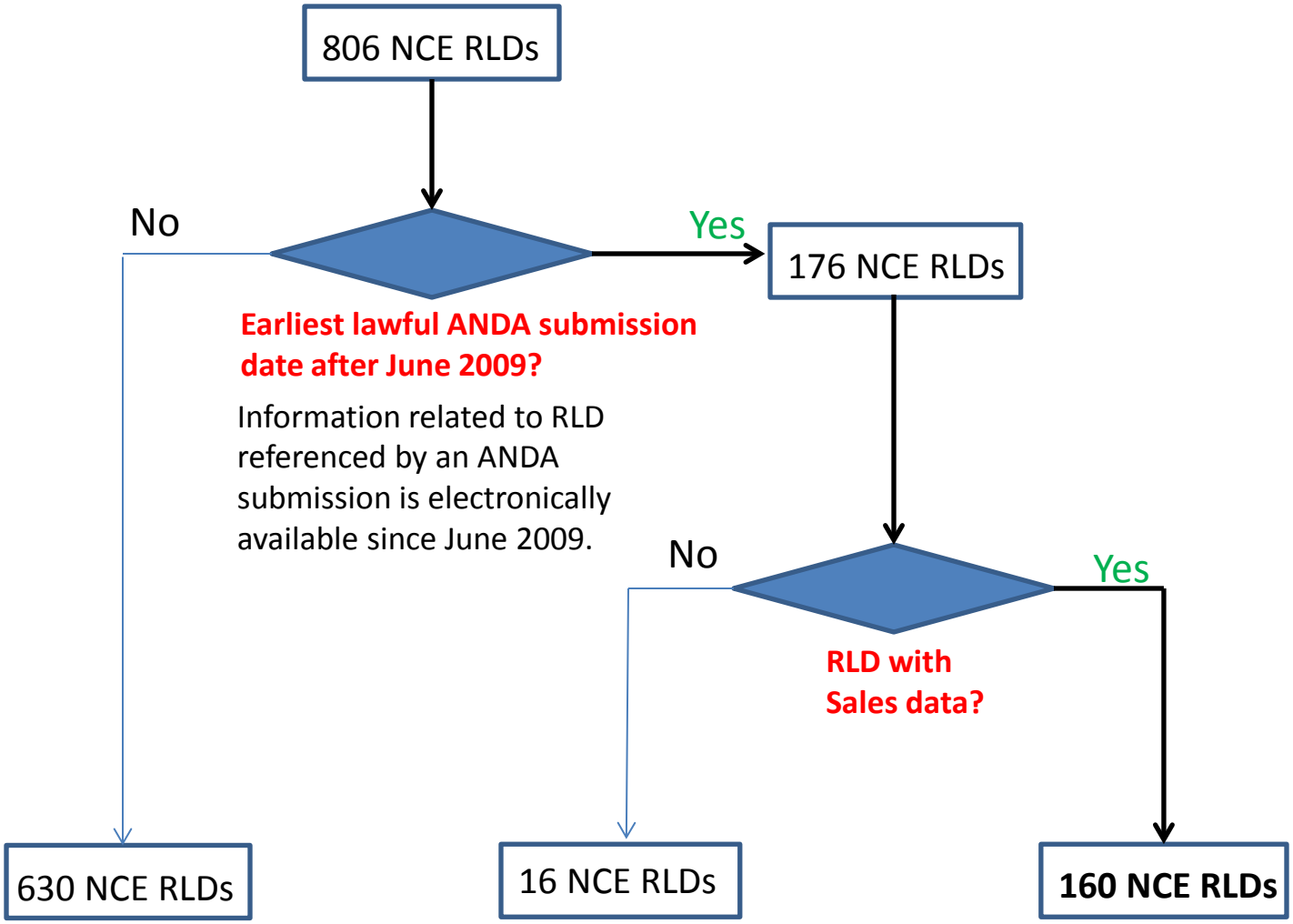
❑ Regulatory information

- Earliest lawful ANDA submission date *(1 year before exclusivity expiration)*
- Actual 1st ANDA submission date if applicable
- Availability of PSG before the first ANDA submission

❑ Pharmacoeconomic information

- Sales before 1st ANDA submission

Data for Analysis



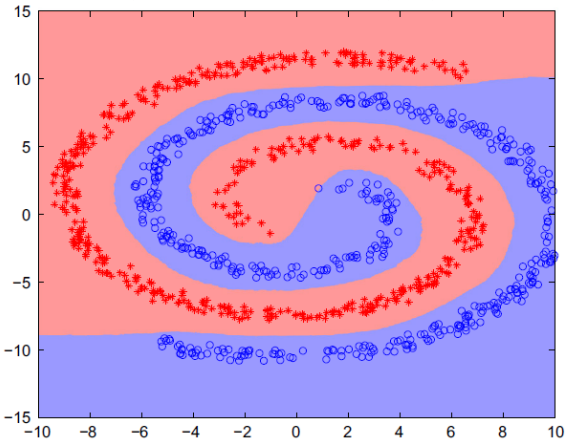
Method

- Formulate the prediction question:

  Time to first ANDA submission ~ f(product, regulatory, pharmacoeconomic)

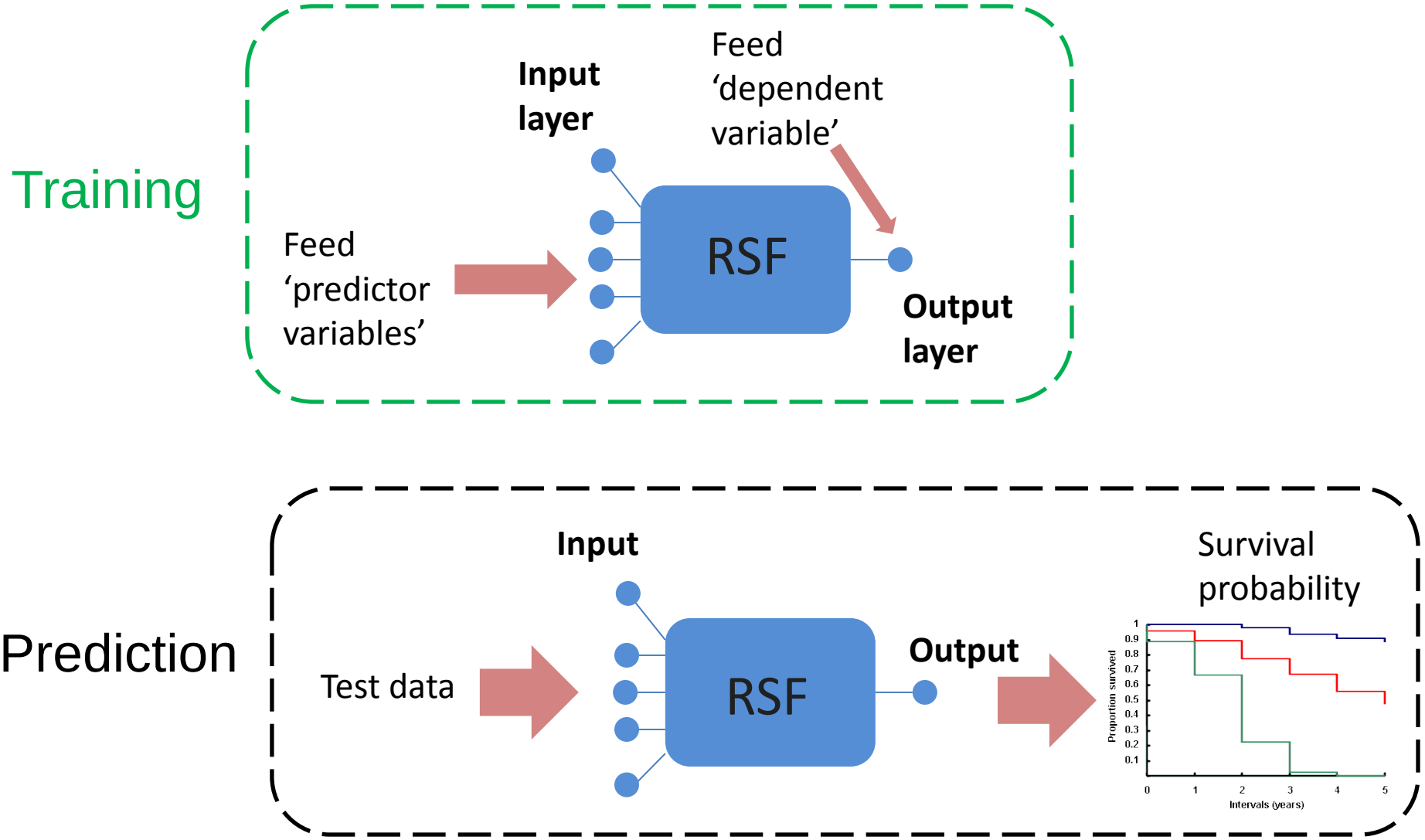
- Methods of analysis
  - Cox regression model
    - Model assumptions – proportional hazards, linear additive relationship between predictor variables
    - Difficult to converge with large number of predictor variables
  - **Machine-learning based method**
    - Random Survival Forest (RSF)
    - Data adaptive (no model assumptions)
    - Capable for large-feature problem

Classification boundary by RSF



Ishwaran et al., The Annals of Applied Statistics, 2008, Vol. 2, No. 3, 841–860

RSF as Supervised Machine Learning Method



Predictive Performance Evaluated by the Concordance Index (C-index)

- C-index essentially measures the proportion of '*subject pairs with good predictions*', in which the subject who experiences the event earlier also has the lower predicted survival probability, over all eligible subject pairs.

An example of subject pair with good prediction

Subject	Real Event Time (day)	Predicted Survival Probability
A	10	0.4
B	40	0.9

- C-index = 1; perfect prediction
- C-index = 0.5; random guess

Performance Testing Approaches

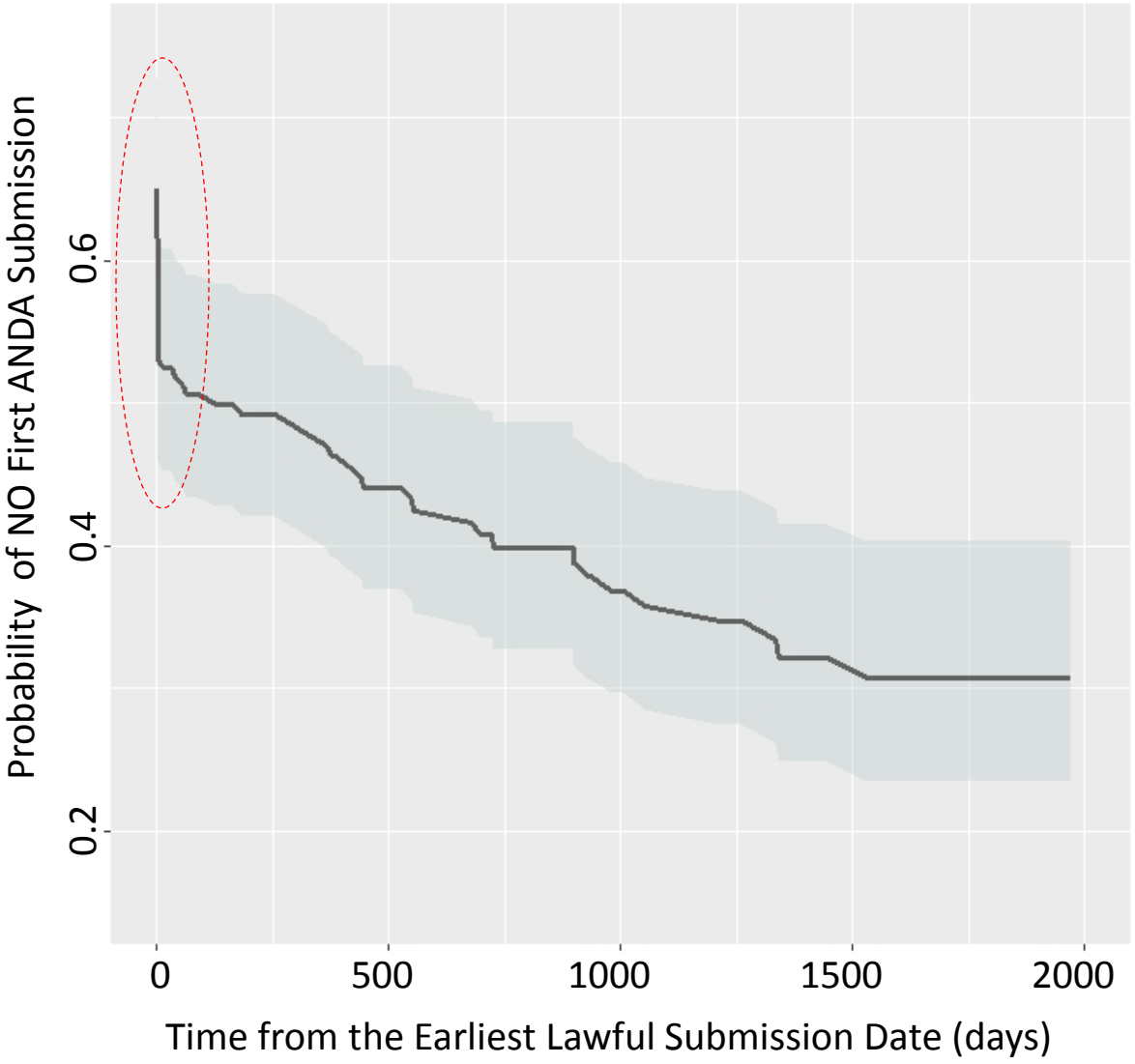
- ***Approach I*** - all data serves as both training and testing datasets
  - Goodness of fit
  - Variable identification

- ***Approach II*** - test by leave-one-out method
  - Leave one sample as testing data, use the rest of data as training data, and then rotate each sample as the testing data to conduct predictions for all samples.
  - Test generalization ability (prediction ability for unknown input)

Results

- Testing Approach I
- Testing Approach II

Kaplan-Meier Plot of Actual Time-Event Data of NCEs



Survival plot:

Event: the first ANDA submission

Survival probability: probability of **NO** first ANDA submission

~35% of NCE RLDs had an ANDA submitted on the earliest lawful ANDA submission date

>45% of NCE RLDs had an ANDA submission within a few days from the earliest lawful ANDA Submission date.

Results from Testing Approach I

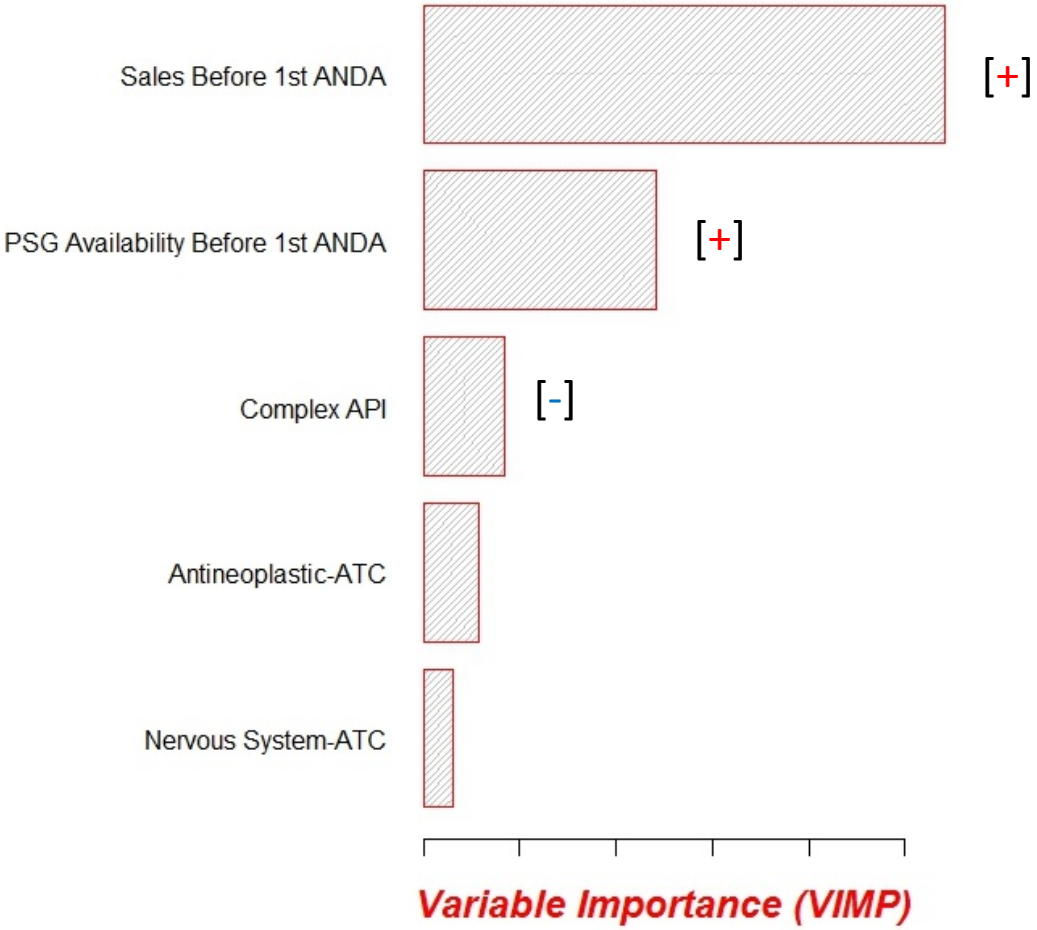
- All the data were used as both training and testing datasets.

- The C-index for the predictions (0.877) show that the whole NCE dataset can be well represented by the RSF model.

- Variable identification

Variable identification

First 5 variable of importance



The first 3 important variables identified by RSF are:

(1) Sales before the first ANDA submission

(2) PSG availability before the first ANDA submission

(3) Complex API

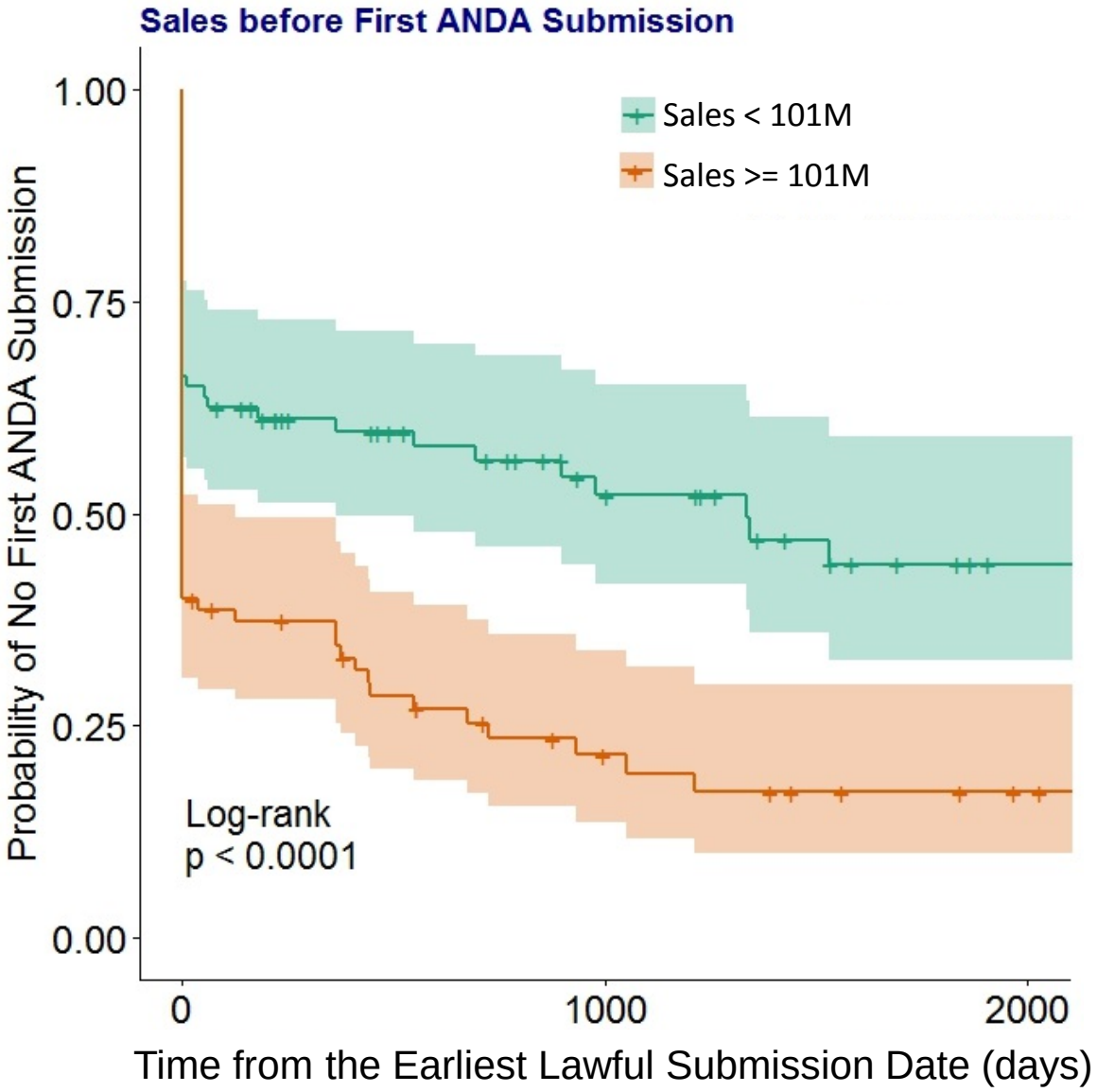
These findings support our hypothesis for predicting the first ANDA submission.

[+] means that the presence or increase of this factor will facilitate the earlier first ANDA submission.

[-] means that the presence or increase of this factor will hinder the first ANDA submission.

Verification for the Identified Variable

The 1st variable of importance



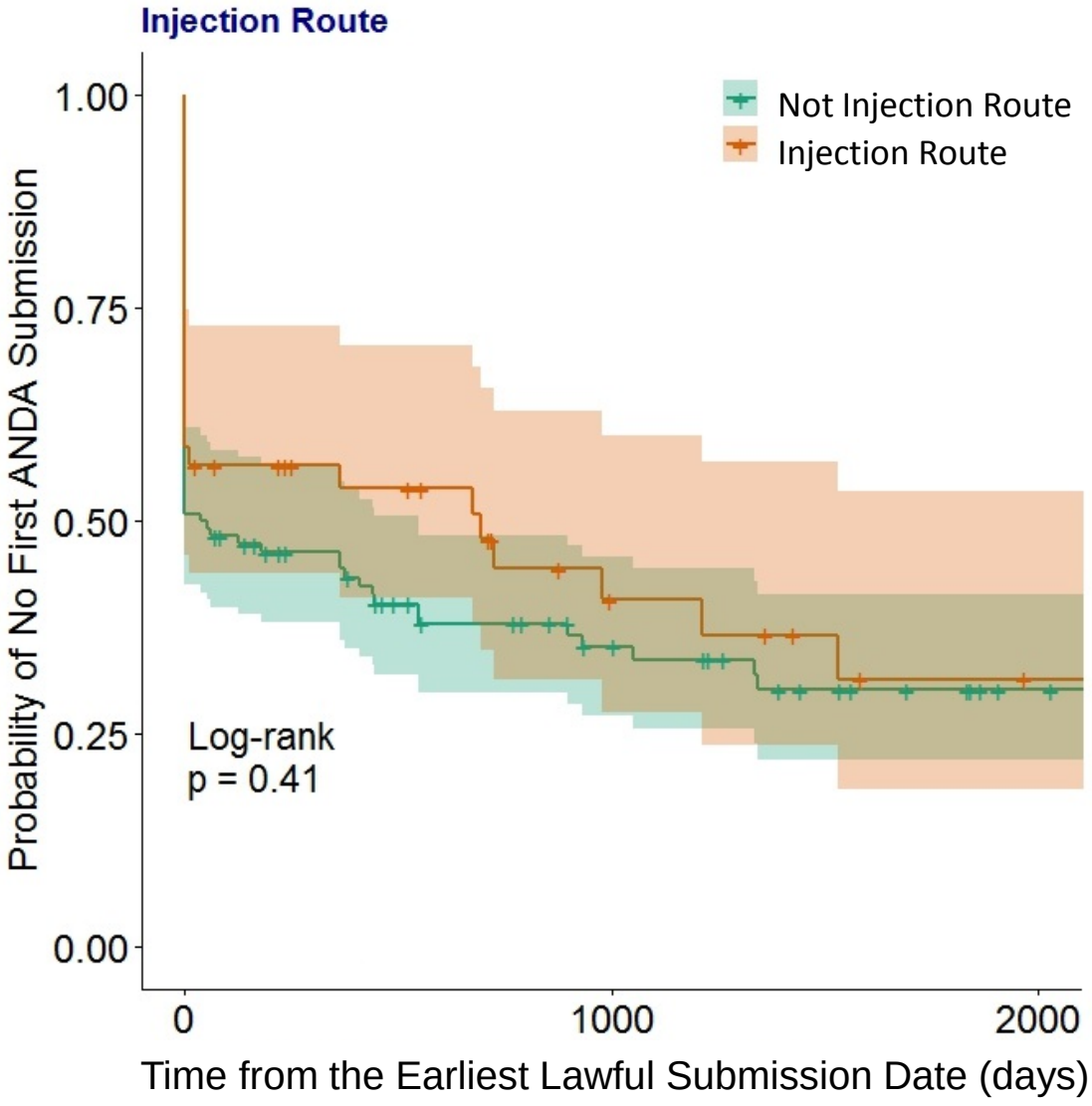
Stratified KM plot by the median value (**$101M**) of all sales in year prior to the first ANDA submission.

Log-rank test shows that two KM estimators are significantly different ($p<0.0001$).

The NCEs with the greater sales tend to have the earlier first ANDA submission.

Verification for the Identified Variable (Negative Control)

The 8th variable of importance



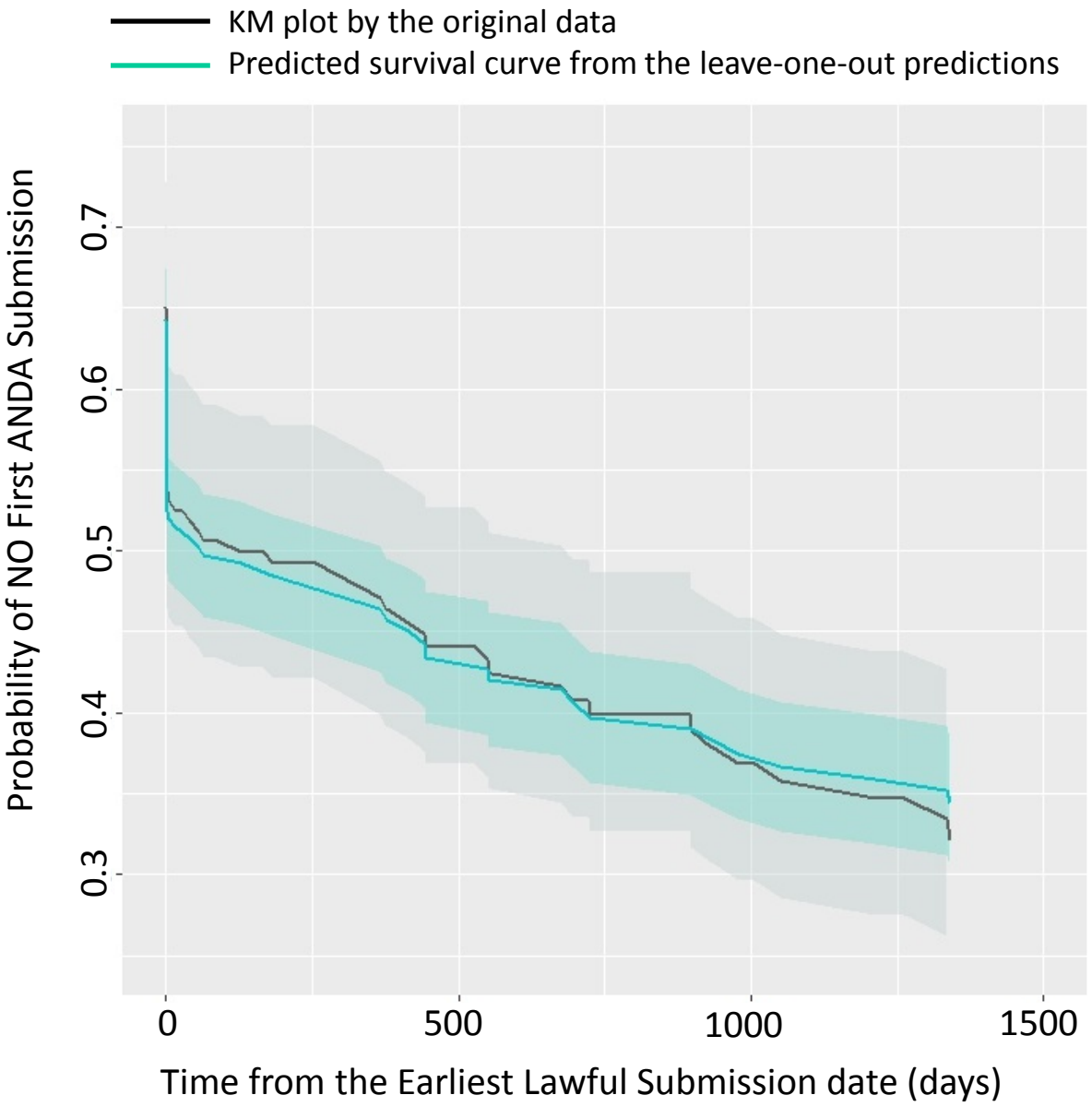
Stratified KM plot by the injection route.

Log-rank test shows that two KM estimators have no significant difference (p=0.41).

Results from Testing Approach II

- Leave-one-out method
  - Rotate each sample as testing data to conduct predictions for all samples.
- Overall prediction
- Prediction at the individual level

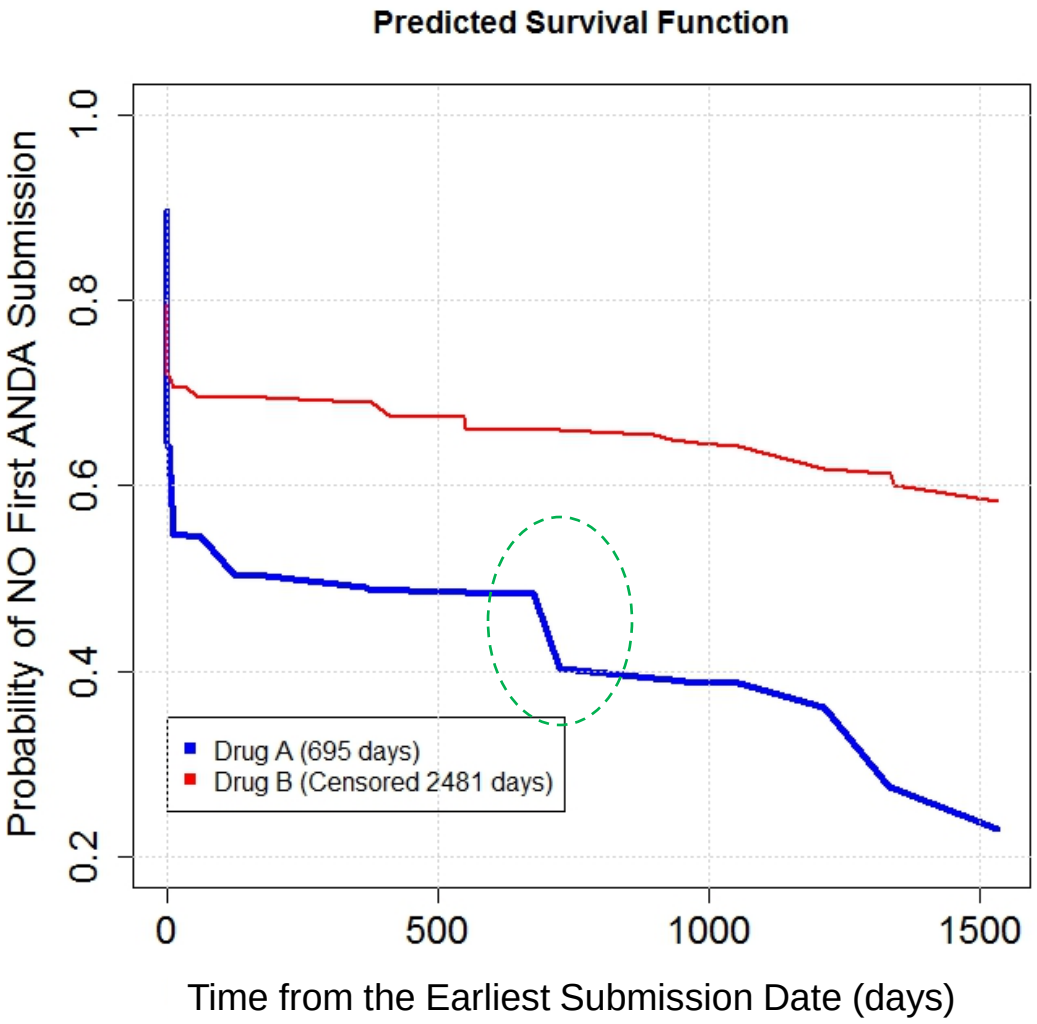
Overall Prediction



Predicted KM plot (green) overlaps with the KM plot from the original data (black).

C-index = 0.703, suggesting a good overall prediction performed by RSF.

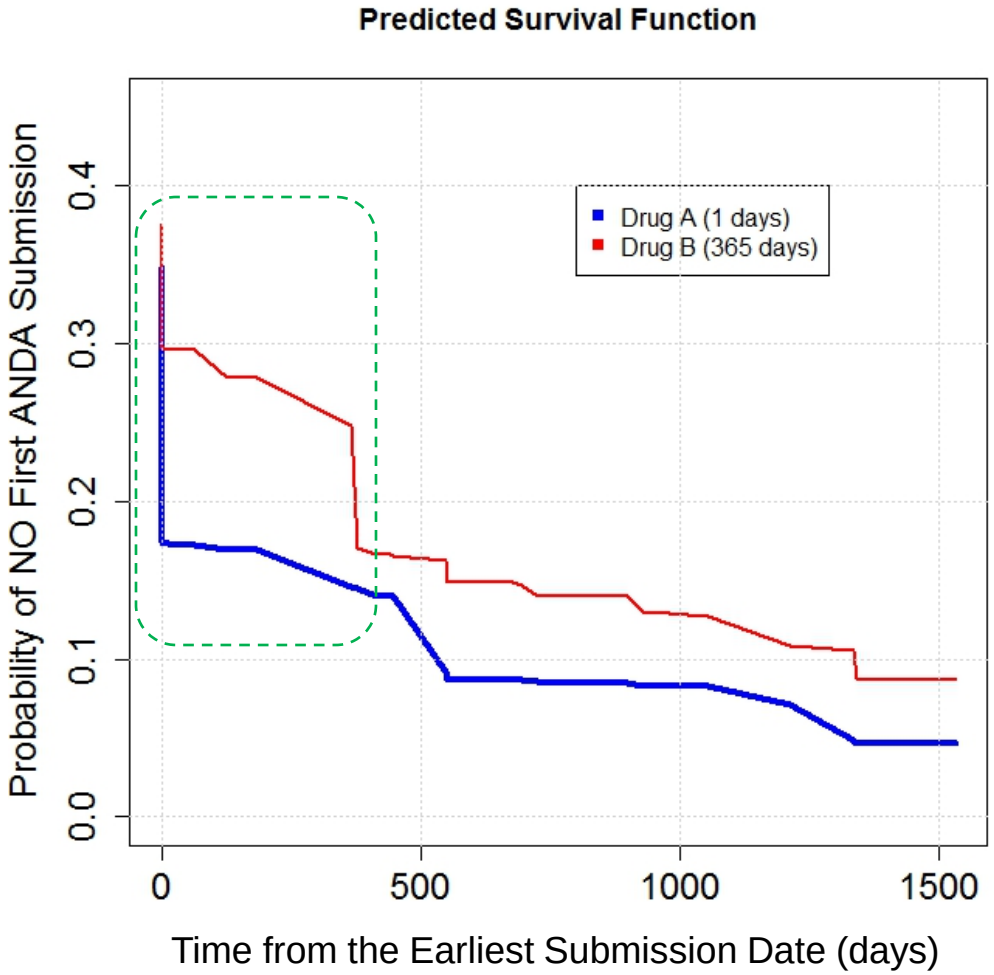
Prediction at the Individual Level (Example I)



True Time-to-Event Information

Drug	Submission time (Days)	Submission status
A	695	Submitted
B	2481	No ANDA Submitted

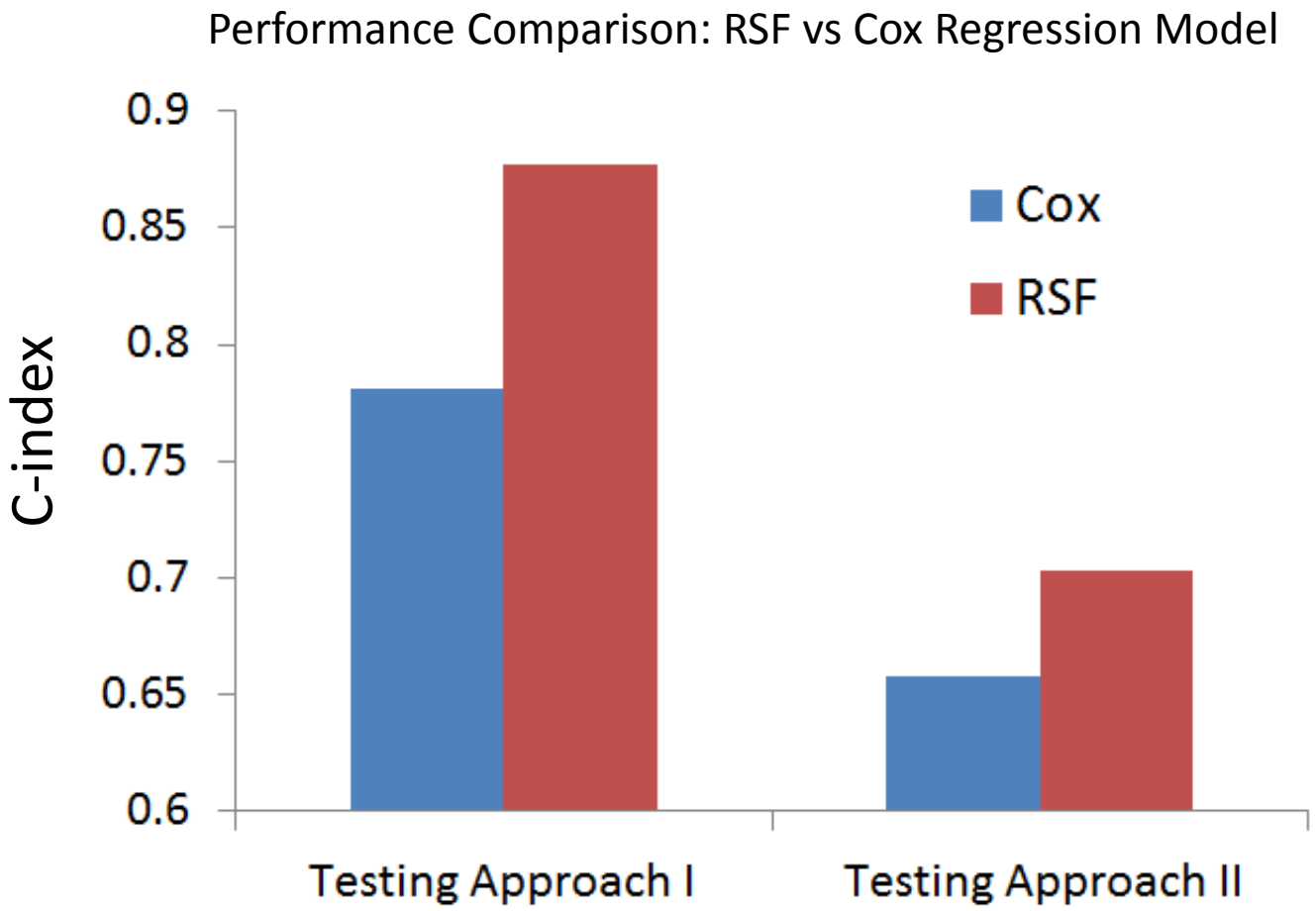
Prediction at the Individual Level (Example II)



True Time-to-Event Information

Drug	Submission time (Days)	Submission status
A	1	Submitted
B	365	Submitted

RSF model outperforms the Cox regression model



Take-Home Message

- Prediction of first ANDA submission timing facilitates generic drug workload management, e.g., prioritizing research efforts and PSG development, especially for complex products.

- The RSF is able to provide quantitative prediction for the first ANDA submission timing for NCEs.

- The RSF model outperforms the conventional Cox regression model in prediction, thus can be an important complement to conventional methods.

- This approach can be expanded to other prediction tasks, e.g., predicting the number of ANDAs submitted.

Team

CDER/OGD/ORS

- Liang Zhao
- Andrew Babiskin
- Saranrat Wittayanukorn
- Xiajing (Jean) Gong
- Zhong (John) Wang
- Meng Hu
- Robert Lionberger

CDER/OSP/OPSA/ES

- Andreas Schick
- Matthew Rosenberg

THANK YOU