Predictive Analysis of First ANDA Submission for NCEs Based on Machine Learning Methodology

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

Disclaimer

▶ The views and opinions expressed in the following PowerPoint slides are those of the individual presenter and should not be attributed to Drug Information Association, Inc. ("DIA"), its directors, officers, employees, volunteers, members, chapters, councils, Communities or affiliates, or any organization with which the presenter is employed or affiliated.

▶ Drug Information Association, Drug Information Association Inc., DIA and DIA logo are registered trademarks. All other trademarks are the property of their respective owners.

My 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

▶ An approved Abbreviated New Drug Application (ANDA) is required to ensure that generic drugs are available to facilitate drug availability and accessibility to US public

▶ The ANDA assessment process involves:
  – Multiple offices in FDA
  – Multiple disciplines
  – Tight timeframe in GDUFA II

▶ Predictive analysis of ANDA submission will greatly enhance the assessment process
  – Strategic planning
  – Workload management

Study Scope

- Time to *first submission* for ANDA referencing new chemical entity (NCE)

- Prediction of first ANDA submission is important for
  - Prioritizing product-specific guidance (PSG) development and research efforts
  - Optimizing resource allocation (e.g., pre-ANDA interactions)

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 Federal Food, Drug, and Cosmetic 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 (Paragraph IV certification).

▶ For ANDAs that reference an RLD with NCE and contain a Paragraph IV certification, the *earliest lawful ANDA submission date* (**ELASD**) is one year before the NCE exclusivity expires.

Collecting Data for Prediction

- **Drug product information**
  - Complex API
  - Complex Dosage Form
  - Complex Delivery Route
  - Complex Drug-Device Combination
  - Abuse Deterrent Formulation
  - Oral Modified Release
  - Anatomical Therapeutic Chemical (ATC)
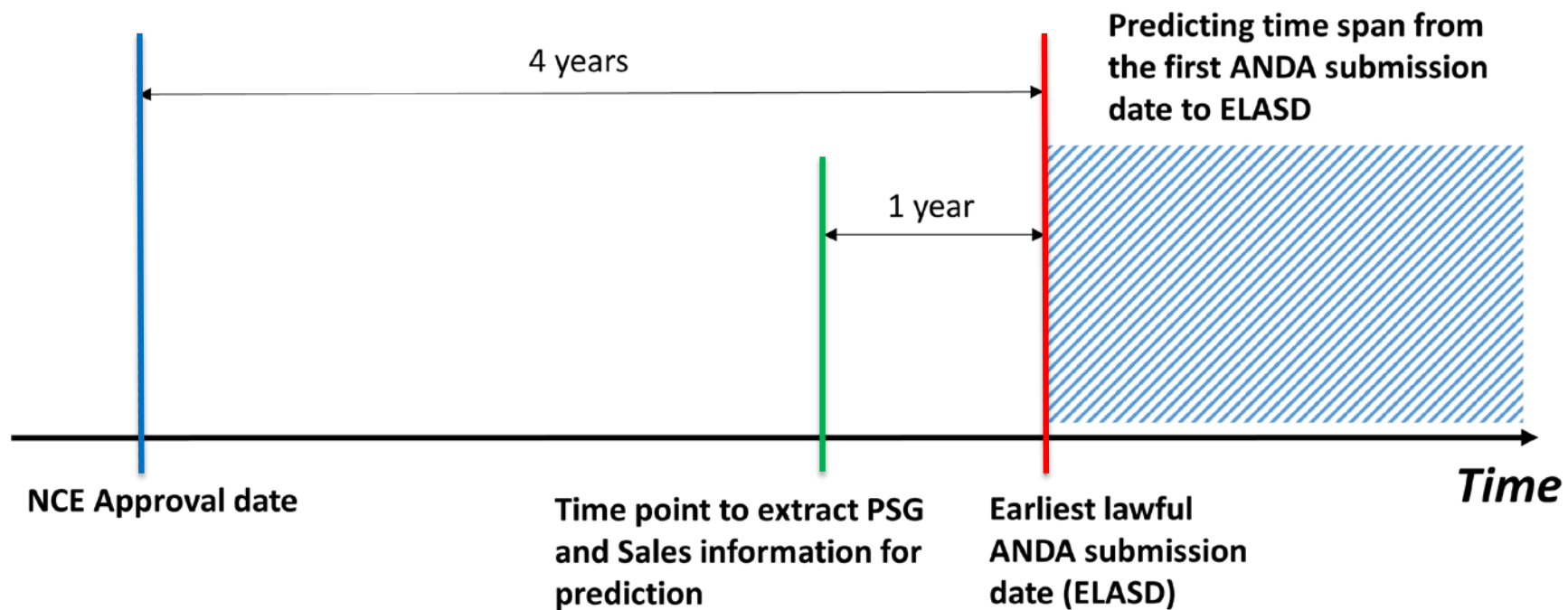 - Acute/Chronic Disease

- **Regulatory information**
  - NDA Approval Date
  - NCE Exclusivity Expiration Date
  - Patent Expiration Date
  - First ANDA Submission Date
  - First PSG Publication Date
  - Risk Evaluation and Mitigation Strategies (REMS)

- **Pharmacoeconomic information**
  - Drug sales from 2011 to 2017

Data Model

Time span from the first ANDA submission date to ELASD $\sim f$(Drug Information, Regulatory Information, Pharmacoeconomic Information)

Method

▶ **Time-to-event analysis**

▶ **Analysis methods**

– 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
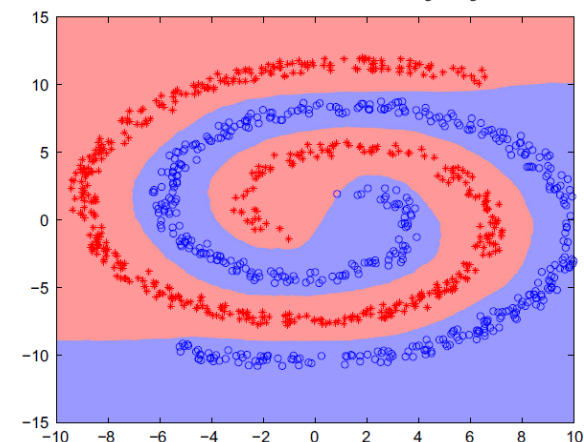
RSF as Supervised Machine Learning Method
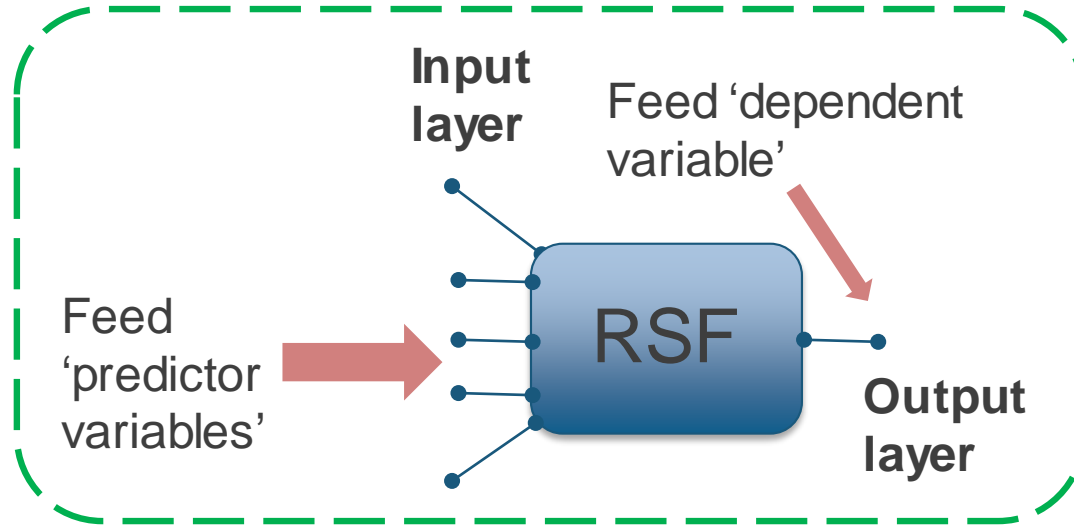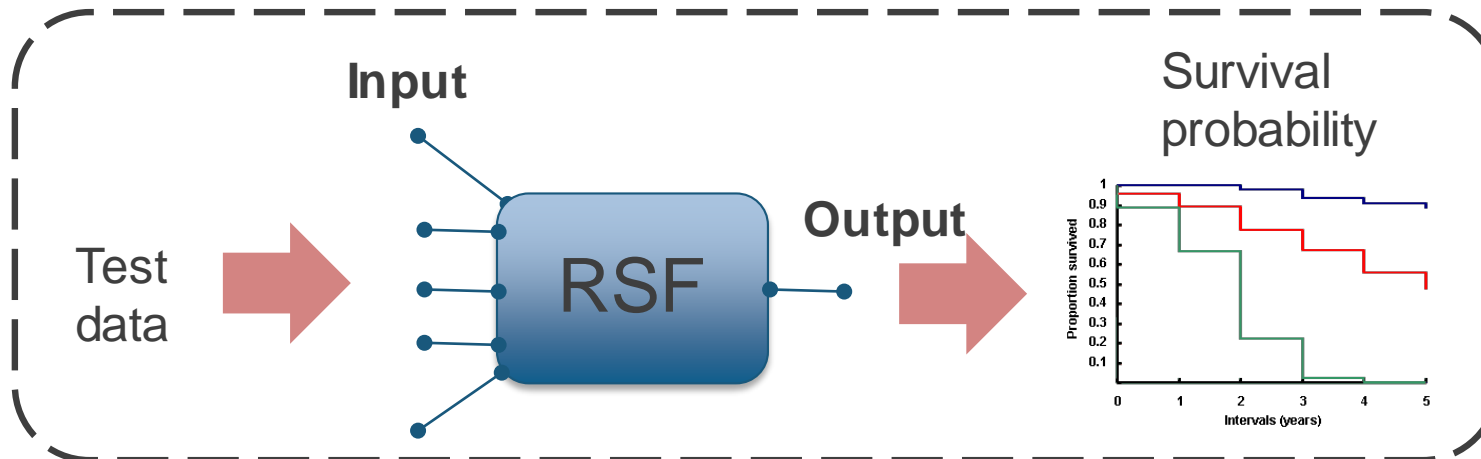
Predictive Performance Evaluated by the Concordance Index (C-index)

▶ 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)

▶ ***Approach III*** – 'predicting the future'
  – NCEs with ELASD before 2017 as training data
  – NCEs with ELASD in 2017 as testing data

Results

- Approach I

- Approach II

- Approach III

Kaplan-Meier (KM) Plot from Original Data
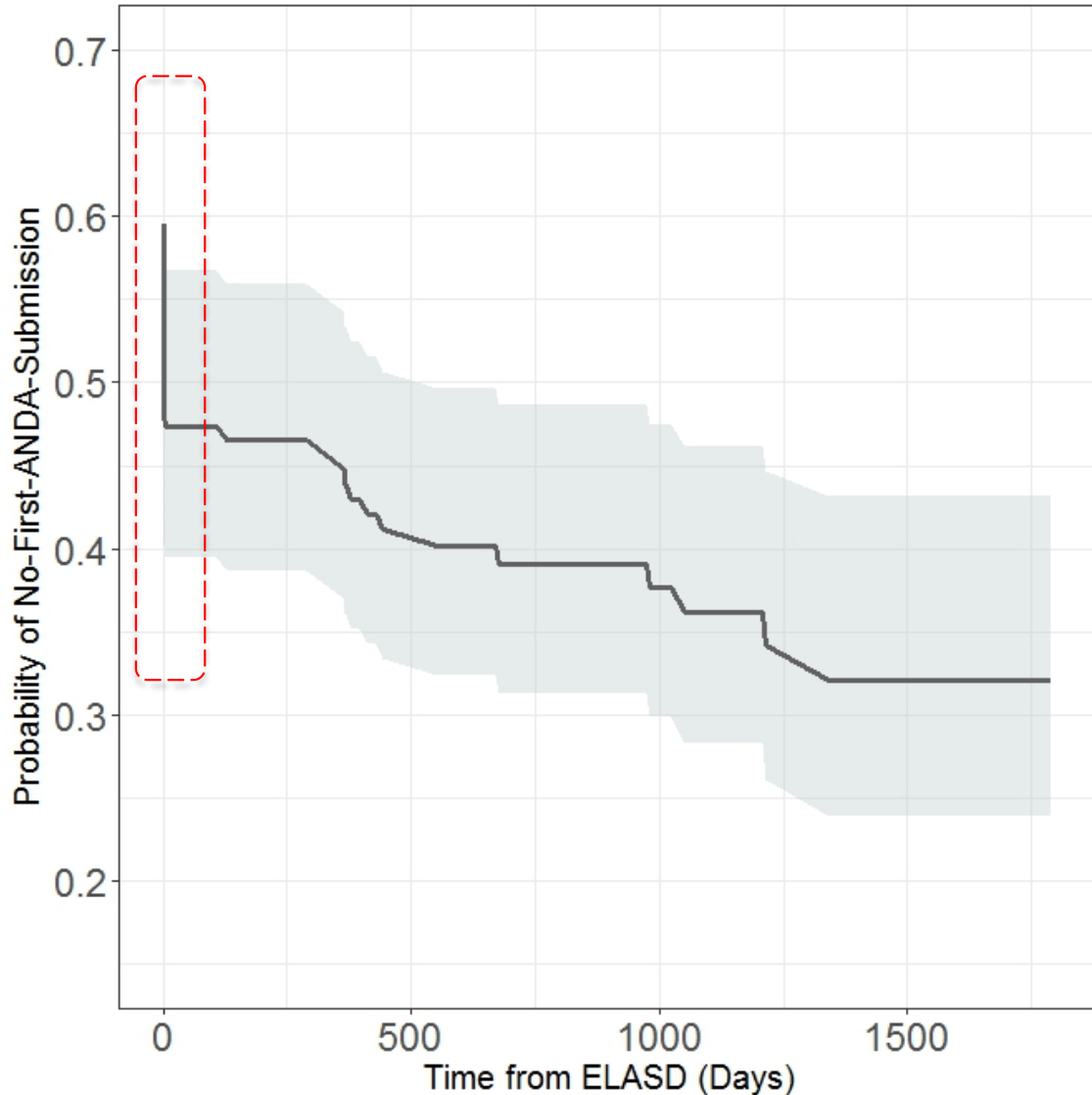
Survival plot:
Event: the first ANDA submission

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

~40% of NCE RLDs had an ANDA submitted on ELASD

~50% of NCE RLDs had an ANDA submission within a few days from ELASD

Results from Approach I

▶ All the data were used as both training and testing datasets

▶ The C-index for the predictions (0.883) show that the whole NCE dataset can be well represented by the RSF model
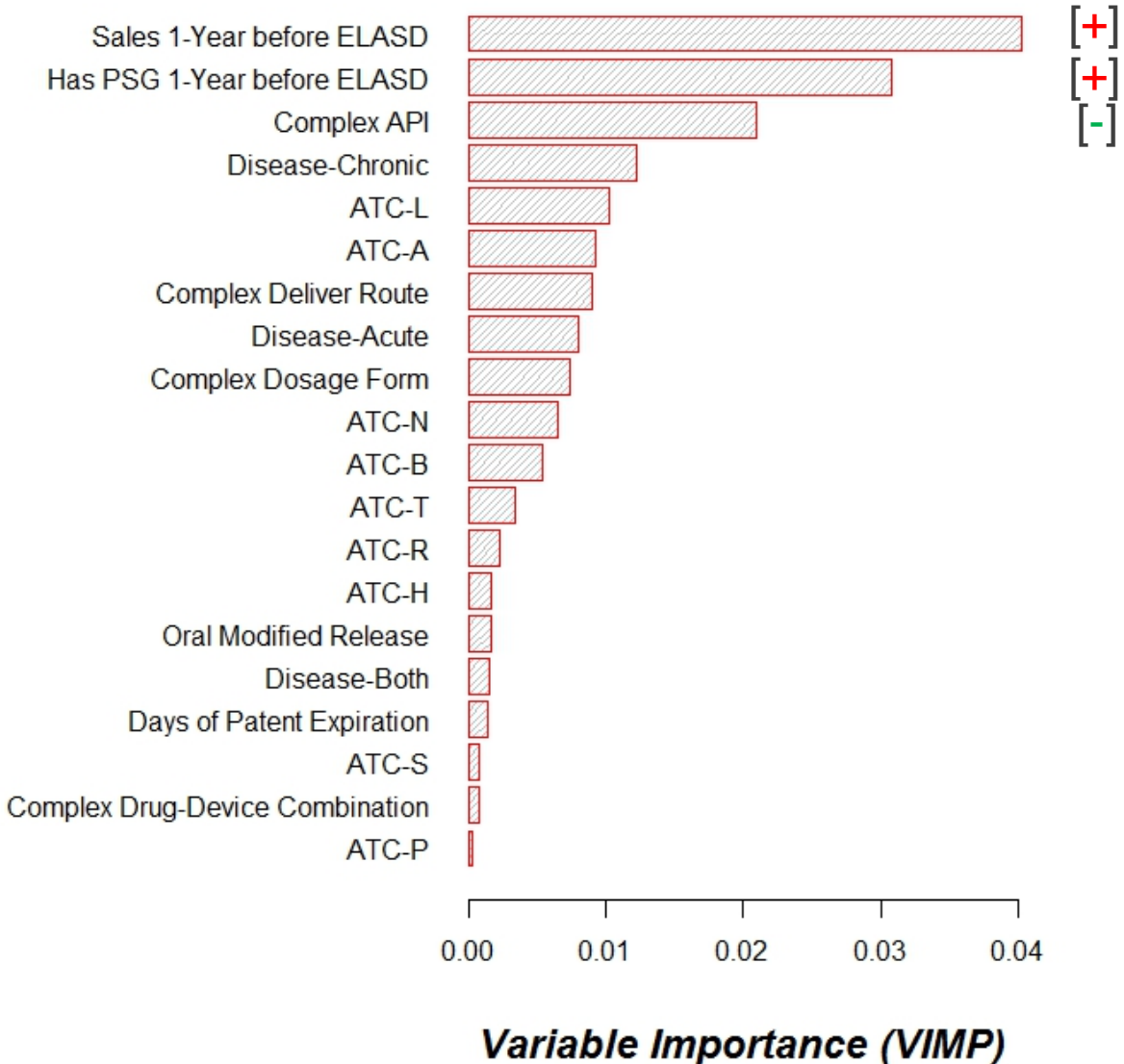
▶ Variable identification

Variable identification

The first 3 important variables:

(1) Sales 1-Year before ELASD

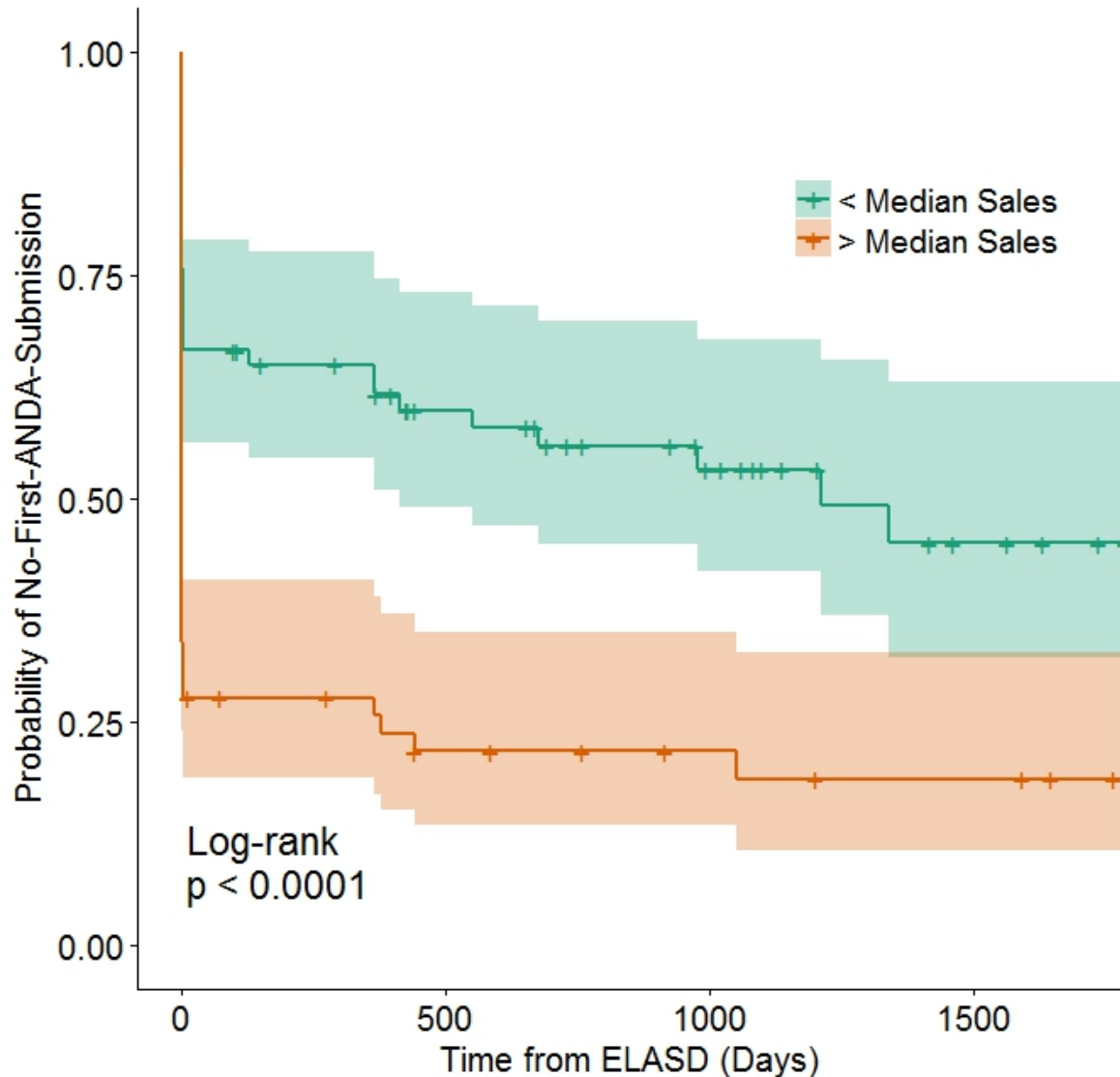**(2) PSG availability 1-Year before ELASD**

(3) Complex API

[+] 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

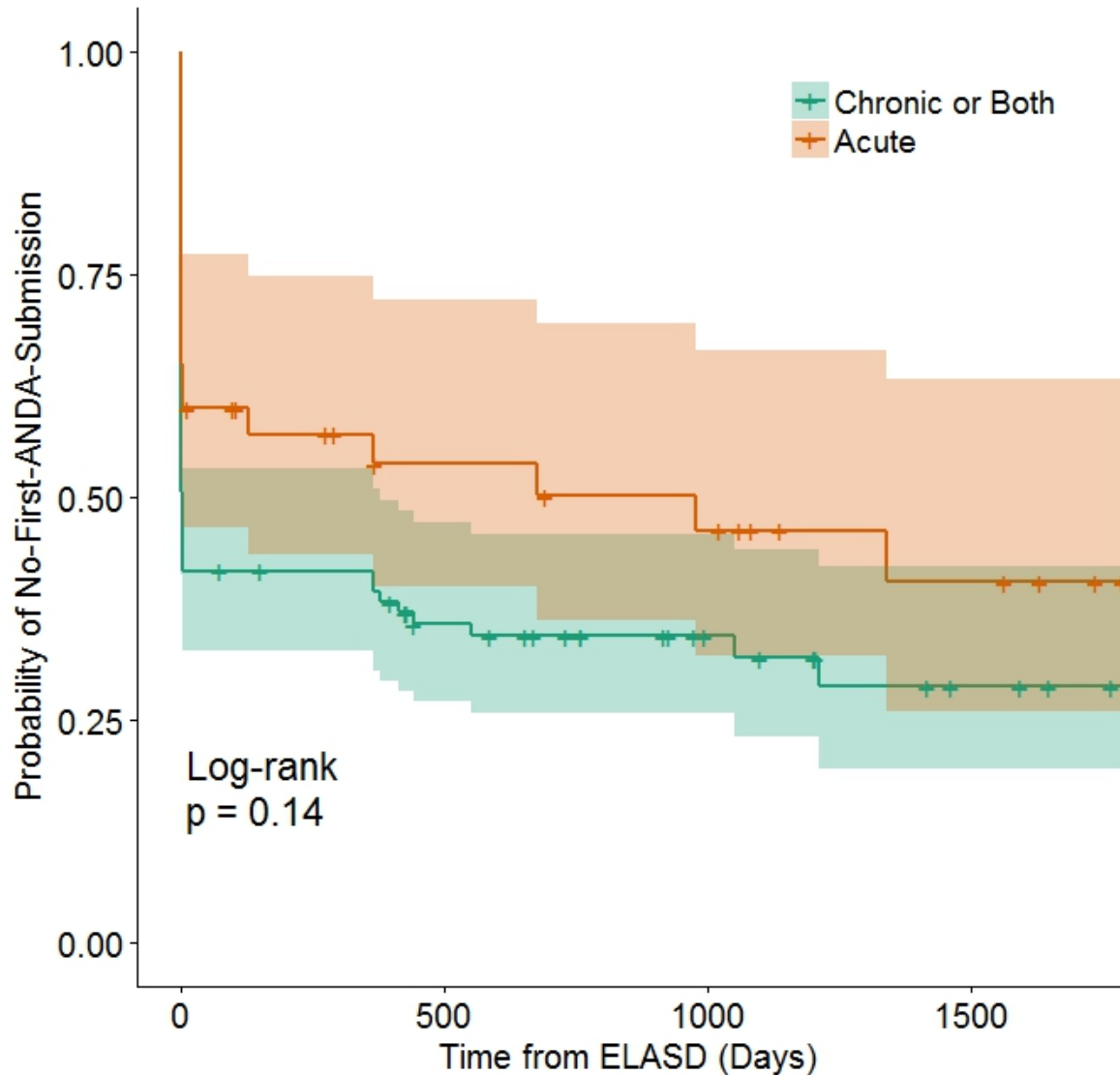Stratified KM plot by the median value (**$20M**) of all sales in year prior to the first ANDA submission.

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)

Stratified KM plot by indictor of the 'Acute Disease'

Log-rank test shows that two KM estimators have no significant difference (p=0.14).

Results from Approach II

▶ Leave-one-out method

  – Rotate each sample as testing data to conduct predictions for all samples
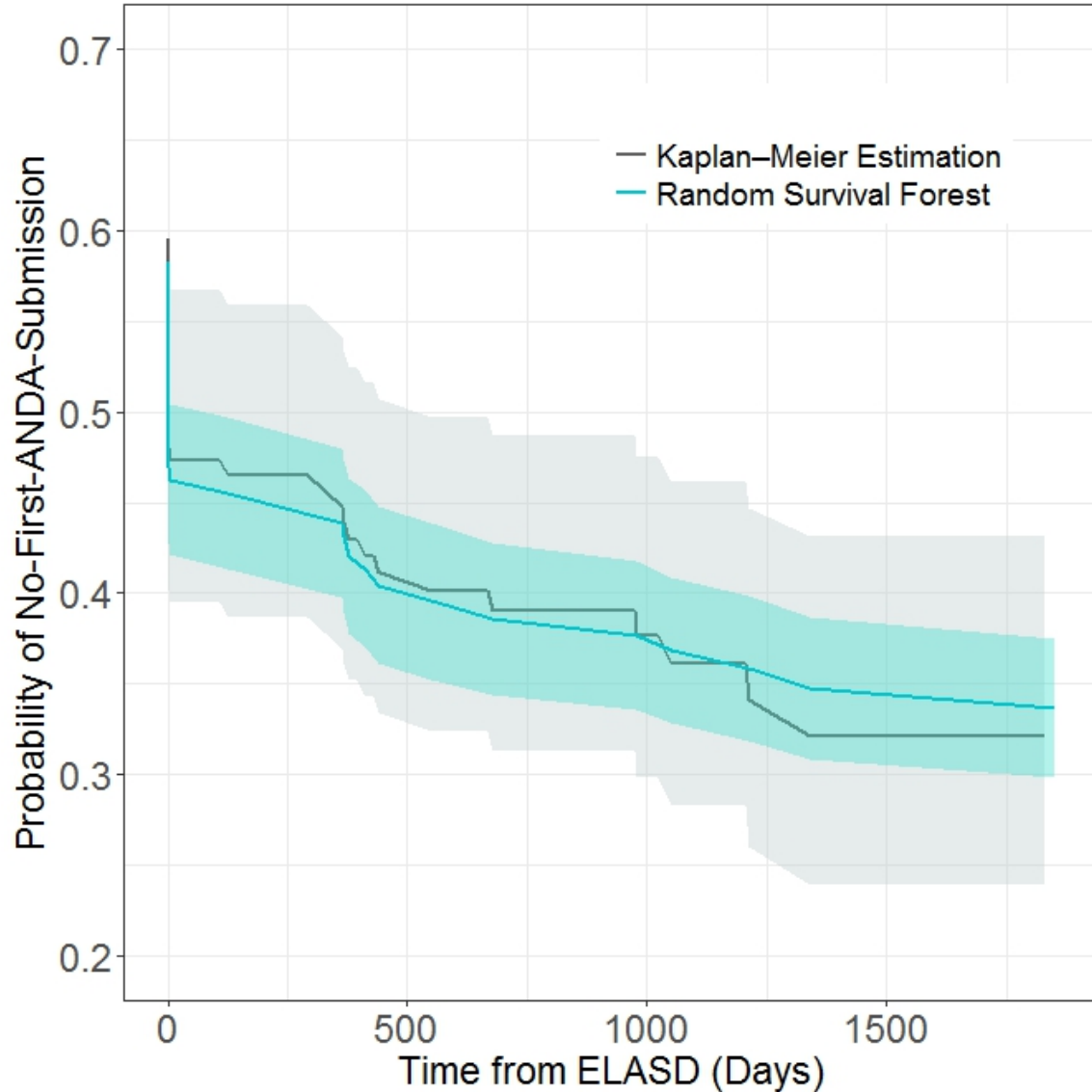
▶ Overall prediction

▶ Prediction at the individual level

Overall Prediction

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

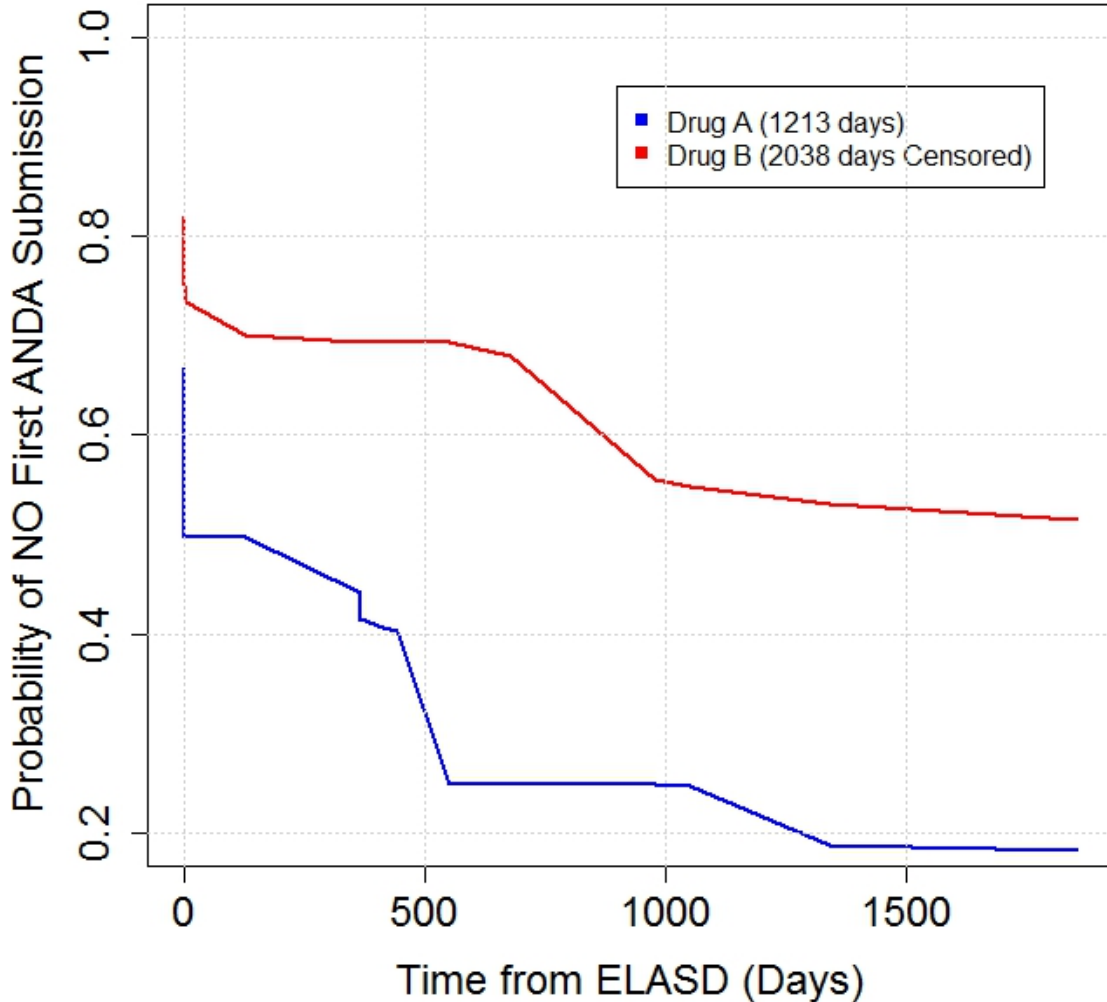C-index = 0.767, suggesting a good overall prediction performed by RSF

Prediction at the Individual Level (Example I)

True Time-to-Event Information

Drug	Time since ELASD (Days)	Submission status
A	1213	Submitted
B	2038	No ANDA Submitted

Prediction at the Individual Level (Example II)
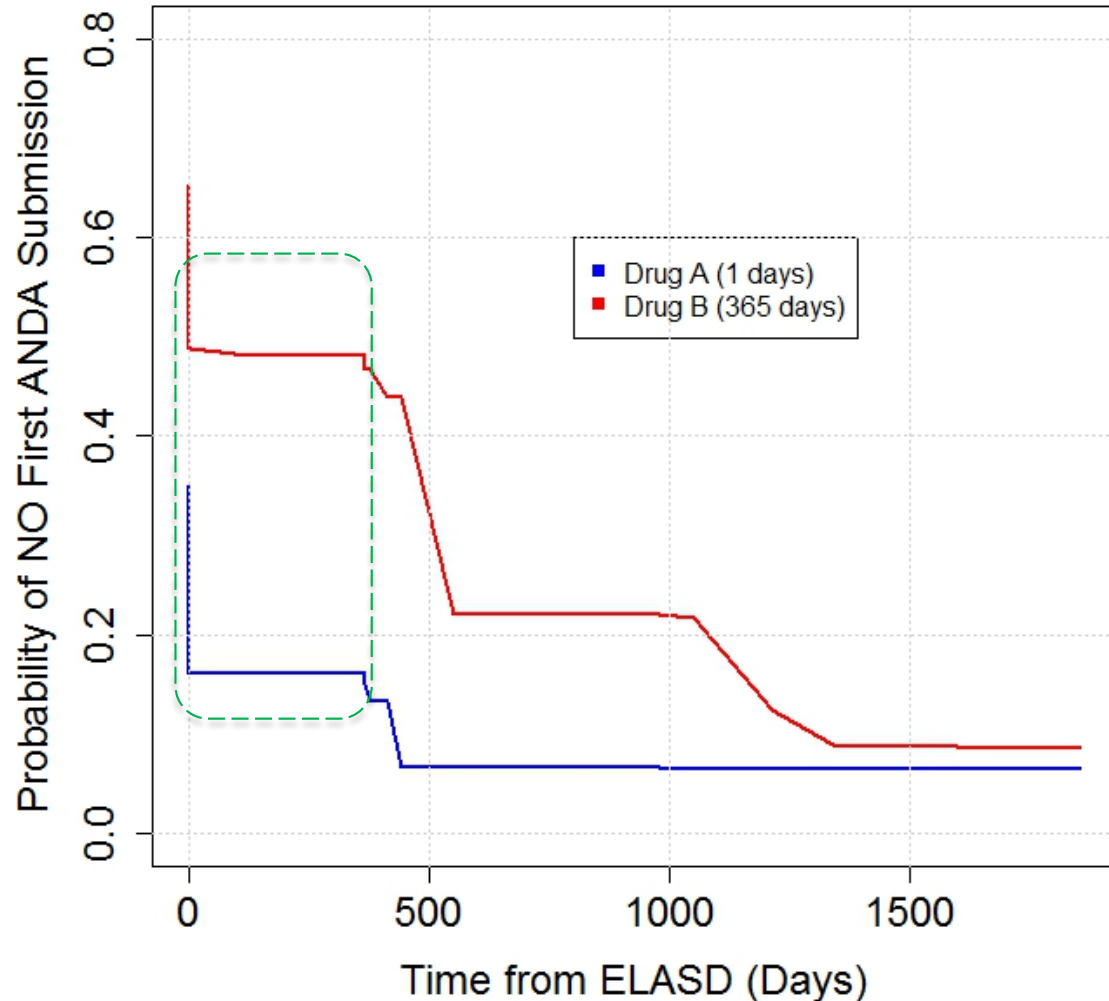
True Time-to-Event Information

Drug	Time since ELASD (Days)	Submission status
A	1	Submitted
B	365	Submitted

Results from Approach III

▶ Predict the future

– Training data: NCEs with ELASD before 2017

– Testing data: NCEs with ELASD in 2017

– C-index = 0.728

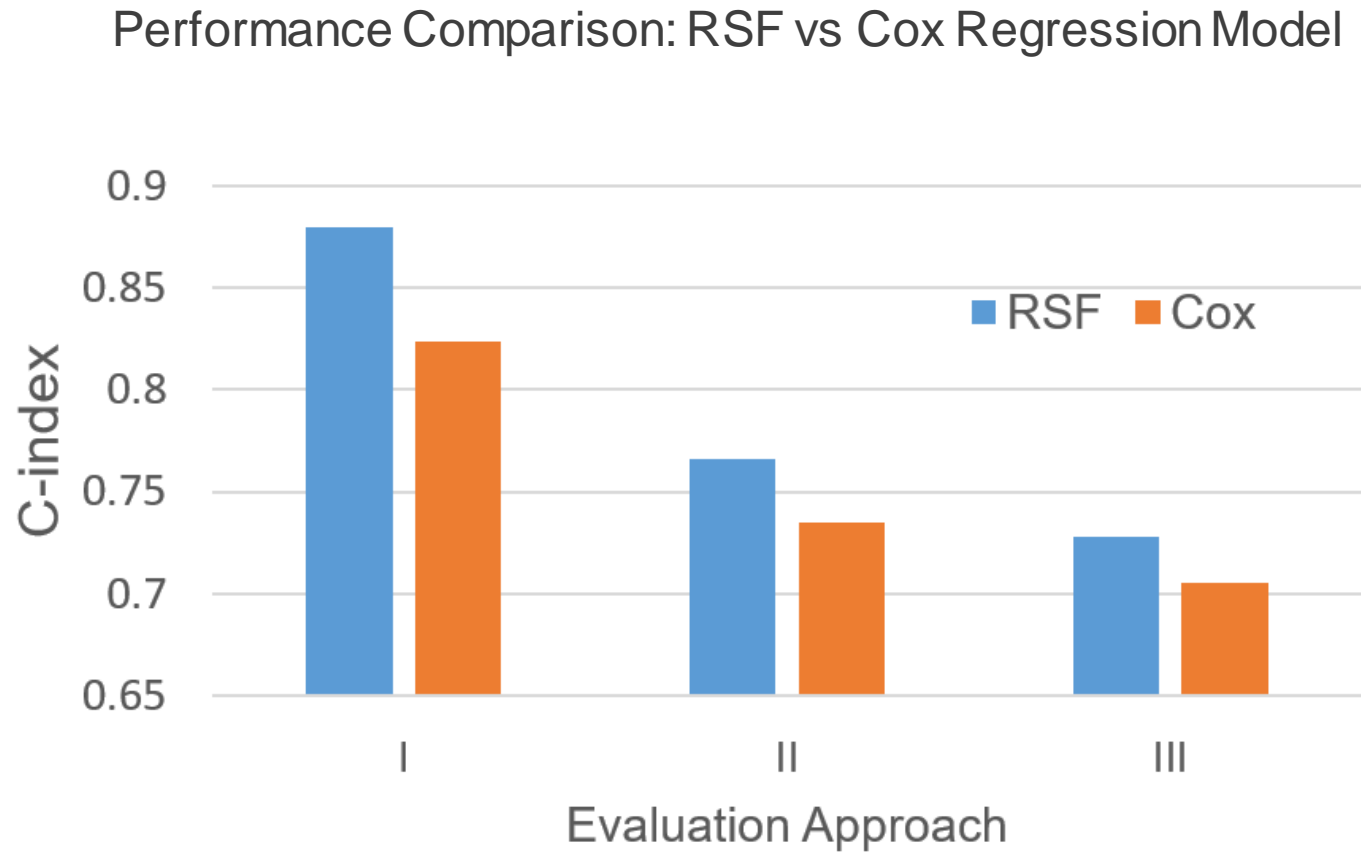# RSF model outperforms the Cox regression model

Performance Comparison: RSF vs Cox Regression Model

Take-Home Message

▶ Prediction of first ANDA submission facilitates generic drug workload management, e.g., prioritizing research efforts and PSG development.

▶ The RSF is able to provide quantitative prediction for the time to first submission for ANDA referencing NCE.

▶ 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

Join the conversation #DIA2018

https://www.linkedin.com/in/meng-hu-6620511a/

Backup

857 NCE RLDs

No

ELASD after June 2009?

Information related to RLD referenced by an ANDA submission is electronically available since June 2009.

Yes → 192 NCE RLDs

No

Available sales data?

Yes

665 NCE RLDs

61 NCE RLDs

131 NCE RLDs