Predicting clinical trial duration via statistical and machine learning models

Joonhyuk Cho, Qingyang Xu, Chi Heem Wong, and Andrew W. Lo

Laboratory for Financial Engineering (LFE), Computer Science and Artificial Intelligence Laboratory (CSAIL)

Abstract

We apply survival analysis as well as machine learning models to predict the duration of clinical trials using the largest dataset so far constructed in this domain. Gradient boosting trees yield the most accurate predictions and we identify key factors that are most predictive of trial duration. This methodology may help clinical researchers optimize trial designs for expedited testing, and can also reduce the financial risk of drug development, which in turn will lower the cost of funding and increase the amount of capital allocated to this sector.

Introduction

Previous Works about
Analysis of Clinical Trial Characteristics

Calculation of Historical Probability of Success (PoS) of Clinical Trial *(Biostatistics, 2019)*

- Largest Dataset (Informa Citeline)
- 185,994 trials

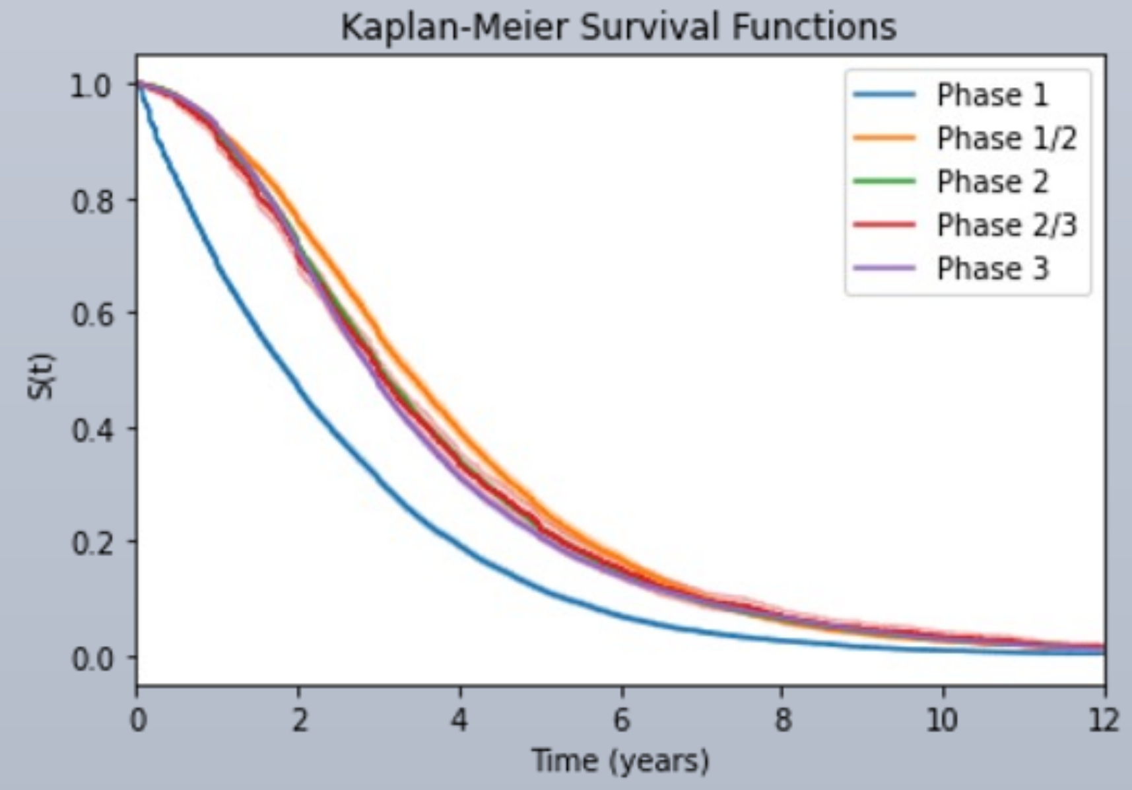
Machine Learning with Statistical Imputation for Predicting Drug Approvals *(HDSR, 2019)*

- 140 features
- 15 disease groups
- 0.81 AUC

Characteristic of the Prediction Model

- Right-Censored (Missing/Unreported trial duration)
- Still, we know the Trial Start Date

→ Survival Analysis rather than Regression Model



Method

Data Statistics

Phase	Trials	Drugs	Duration Mean	Duration SD	Duration 25% Qt.	Duration Median	Duration 75% Qt.
1	20260	7782	2.3	2.1	0.7	1.7	3.2
1/2	7455	3246	3.6	2.5	1.8	3.0	4.8
2	36066	6486	3.4	2.5	1.7	2.8	4.5
2/3	1905	1122	3.4	2.5	1.6	2.8	4.5
3	21152	3797	3.4	2.5	1.7	2.7	4.3
Total	86,838	12,454	3.2	2.5	1.4	2.6	4.2

- 86,838 Trials, 12,454 Drugs
- Phase 1 (2.3 years) significantly shorter than other phases (3.4 years)
- However, mean/std statistics is not enough!

Survival Analysis

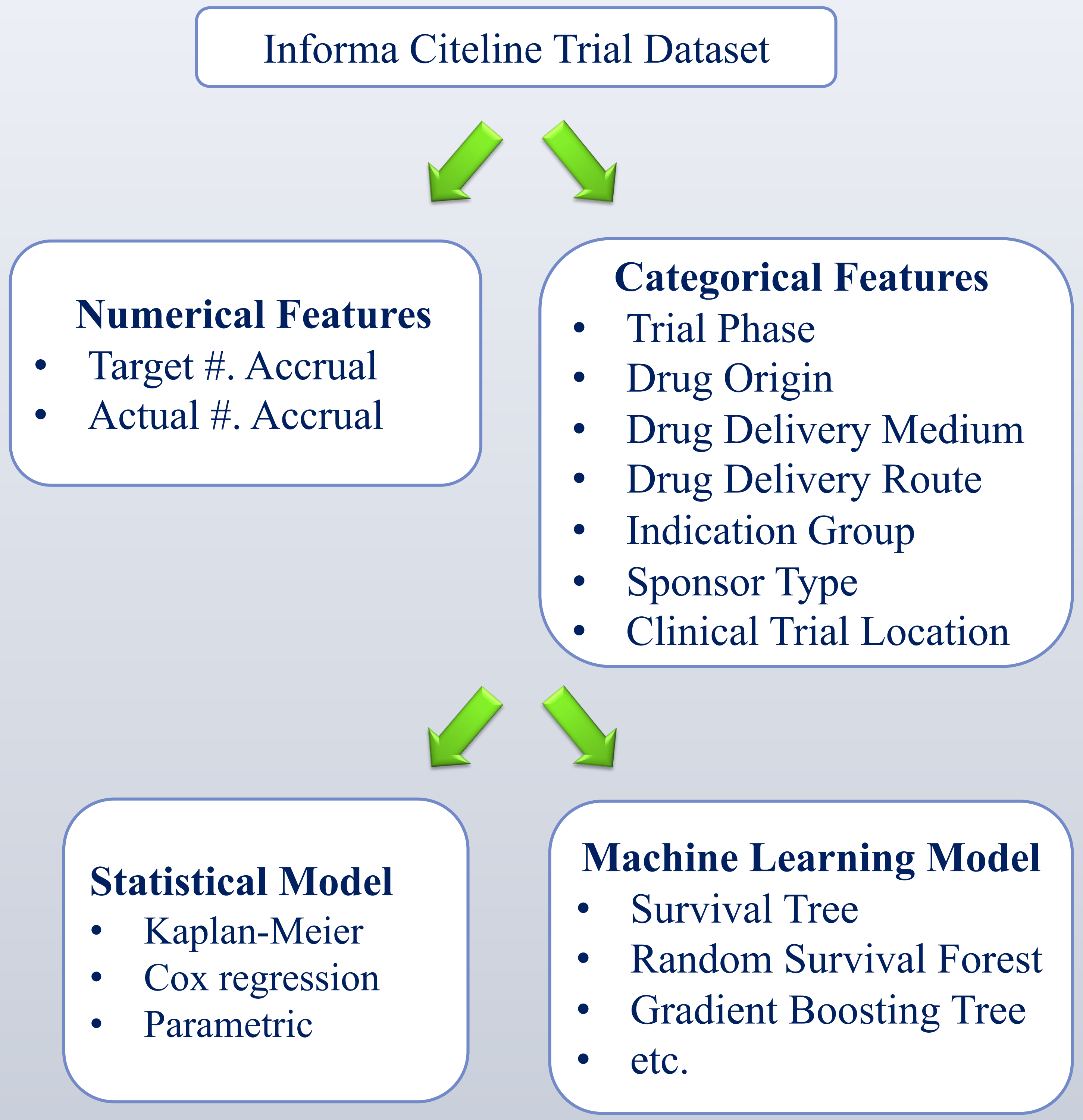
- Outcome variable of interest its time until event
- Useful with right-censored data
  - Hardware failure
  - Customer analytics
  - Human survival
- Evaluation metric: C-index (ranking statistics)

Goal: Predict the Hazard Function of Survival function!

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T > t)}{\Delta t} = -\frac{S'(t)}{S(t)}$$

$h(t) \geq 0$, S(t) monotonically decreases from 1 to 0

Prediction Model



Informa Citeline Trial Dataset

Numerical Features

- Target #. Accrual
- Actual #. Accrual

Categorical Features

- Trial Phase
- Drug Origin
- Drug Delivery Medium
- Drug Delivery Route
- Indication Group
- Sponsor Type
- Clinical Trial Location

Statistical Model

- Kaplan-Meier
- Cox regression
- Parametric

Machine Learning Model

- Survival Tree
- Random Survival Forest
- Gradient Boosting Tree
- etc.

Performance Evaluation

C-index:
Compare performance between models

Feature Importance:
Compare importance between features

Results

C-Index:

Model	Data Preprocess.	Split 1	Split 2	Split 3	Split 4	Split 5	c-index (mean)	c-index (SE)
Cox Regression	Original	0.686	0.687	0.681	0.680	0.683	0.683	0.001
Weibull AFT	Original	0.686	0.687	0.682	0.680	0.683	0.684	0.001
Survival Tree	Original	0.643	0.669	0.665	0.661	0.690	0.666	0.007
Survival Tree	WoE	0.679	0.673	0.674	0.679	0.674	0.676	0.001
Random Forest	Original	0.704	0.703	0.701	0.698	0.699	0.701	0.001
Random Forest	WoE	0.698	0.693	0.693	0.696	0.696	0.695	0.001
Gradient Boosting	**Original**	**0.714**	**0.715**	**0.713**	**0.709**	**0.712**	**0.713**	**0.001**
Gradient Boosting	**WoE**	**0.705**	**0.702**	**0.702**	**0.704**	**0.702**	**0.703**	**0.001**
DeepSurv	Original	0.710	0.705	0.703	0.701	0.704	0.704	0.001
DeepSurv	WoE	0.693	0.690	0.691	0.693	0.692	0.692	0.000
Neural-MTLR	Original	0.697	0.676	0.685	0.672	0.689	0.683	0.004
Neural-MTLR	WoE	0.668	0.674	0.628	0.674	0.666	0.662	0.008
Survival SVM	Original	0.682	0.683	0.677	0.676	0.679	0.679	0.001
Survival SVM	WoE	0.680	0.673	0.677	0.679	0.676	0.677	0.001

Gradient Boosting Survival Trees
best C-index with 0.714

Feature Importance:

Indication Group> Sponsor Type > Trial Phase
> Drug Characteristics (Medium, Delivery Route)

Conclusions

Developed a prediction model that can predict survival characteristic of each trial duration.

Next Step:
Incorporate analysis model of

- Probability of success
- Trial duration
- Financial modeling of the clinical trial

for precise estimation of financial Net Present Value of the clinical trial.

→ Megafund which invest in multiple clinical trials