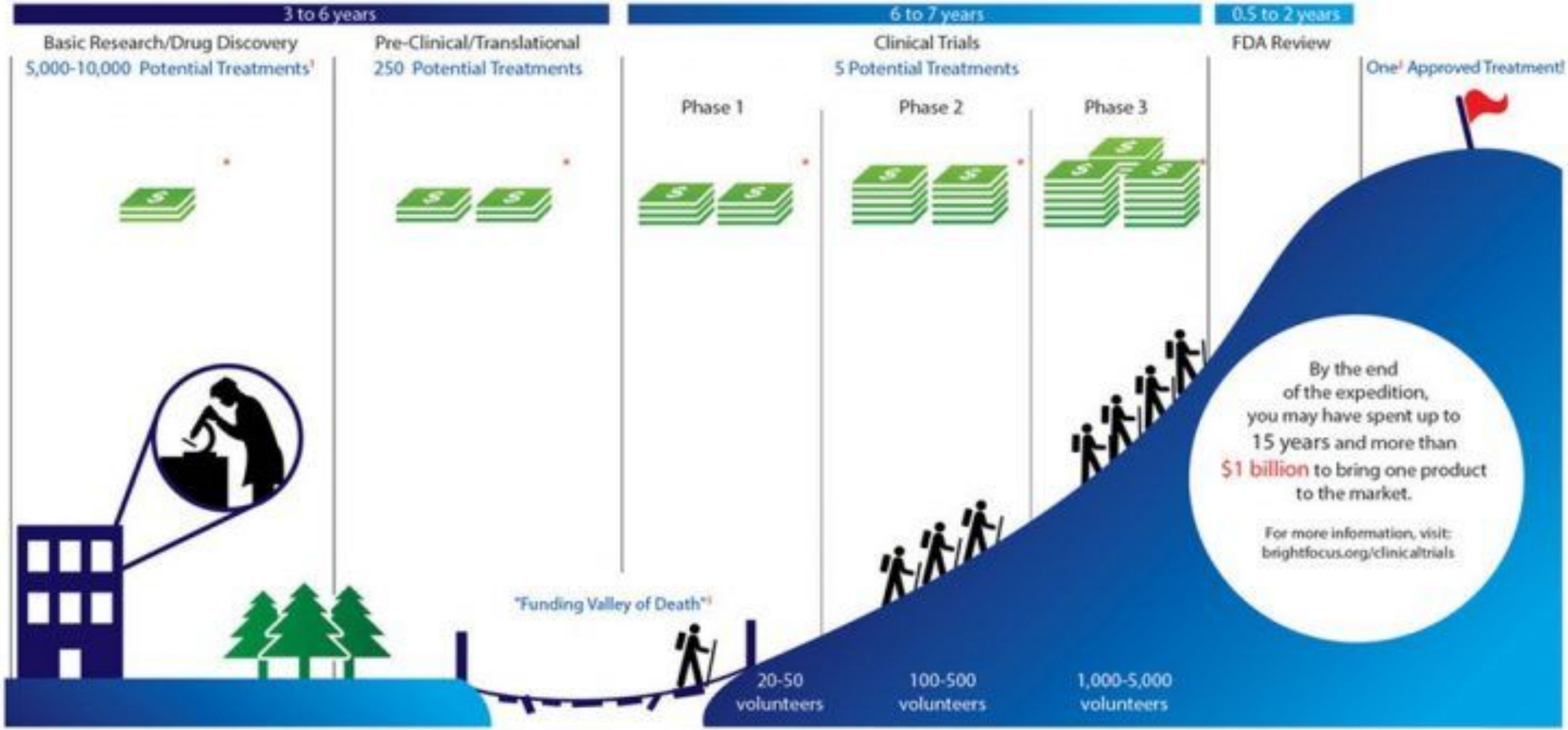
Predicting Clinical Trial Outcomes

Exploration of publicly available data

Nicholas McBride, PhD
May, 2024

Making medicine is expensive (in the US)

- Clinical trials for Investigational New Drugs (IND) are a significant driver of healthcare costs
- A single study costs $19M and up to $255M for pivotal Phase III trials
- Multiple clinical trials and phases required put the cost of marketing a new drug at $2.6B



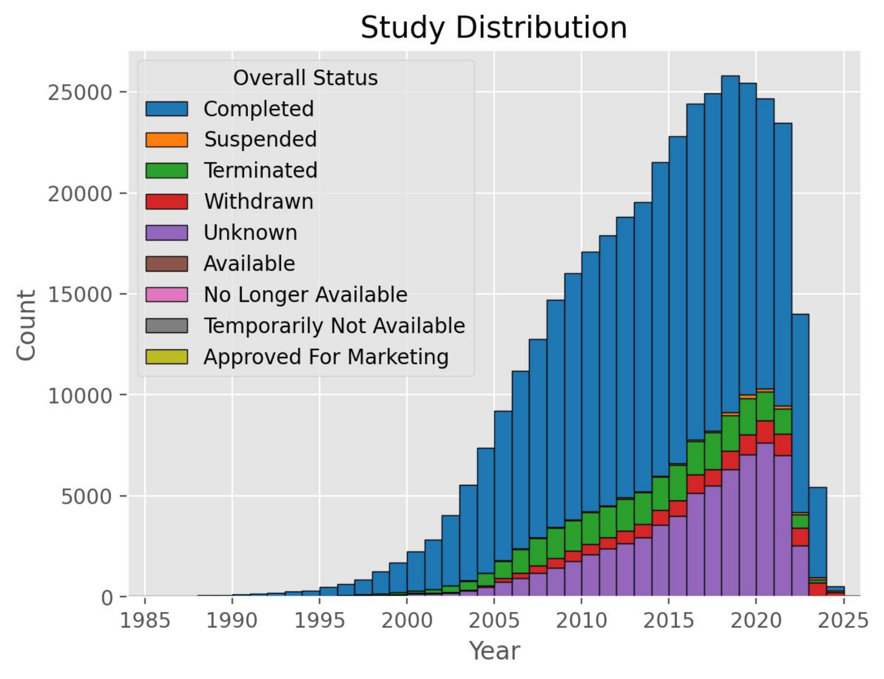
By the end of the expedition, you may have spent up to 15 years and more than $1 billion to bring one product to the market.

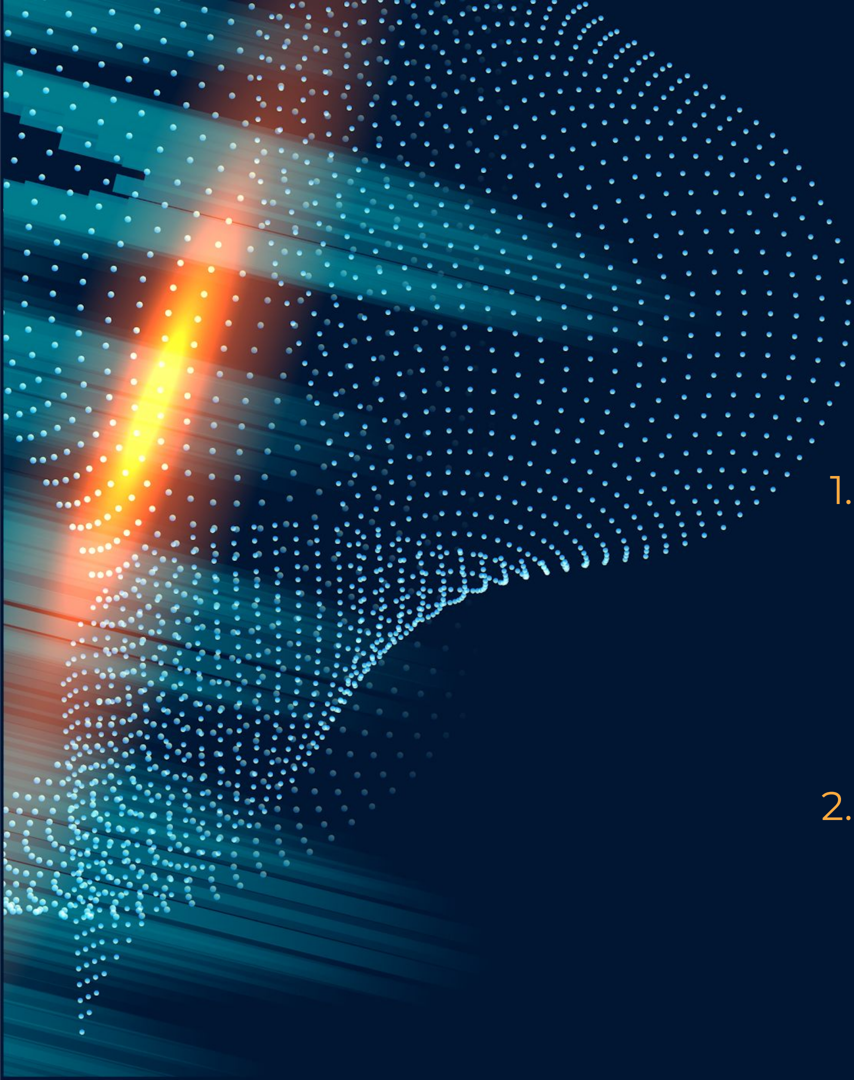
For more information, visit: brightfocus.org/clinicaltrials

Image credit Texas A&M University

ClinicalTrials.gov

- Database of all clinical trials and results became publicly available in 2008
- API provides access to 568 fields of data
- 495,650 indexed studies as of 10:26 pm last night





Project Goals

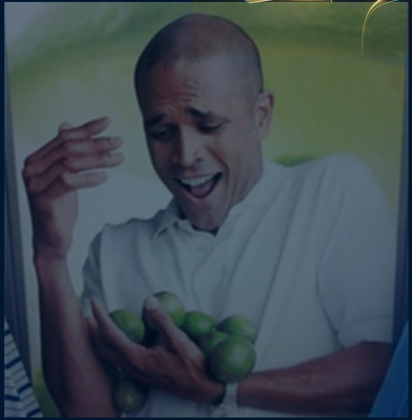
1. Identify data features within ClinicalTrials.gov that may predict clinical trial outcomes
    a. Features of study design and protocol prior to initiation
    b. Features that may be monitored as clinical trials progress

2. Develop a predictive model for clinical trial completion vs. suspension, termination, withdrawal, or abandonment
    a. Achieve predictive results better than the baseline mean

Data Cleaning

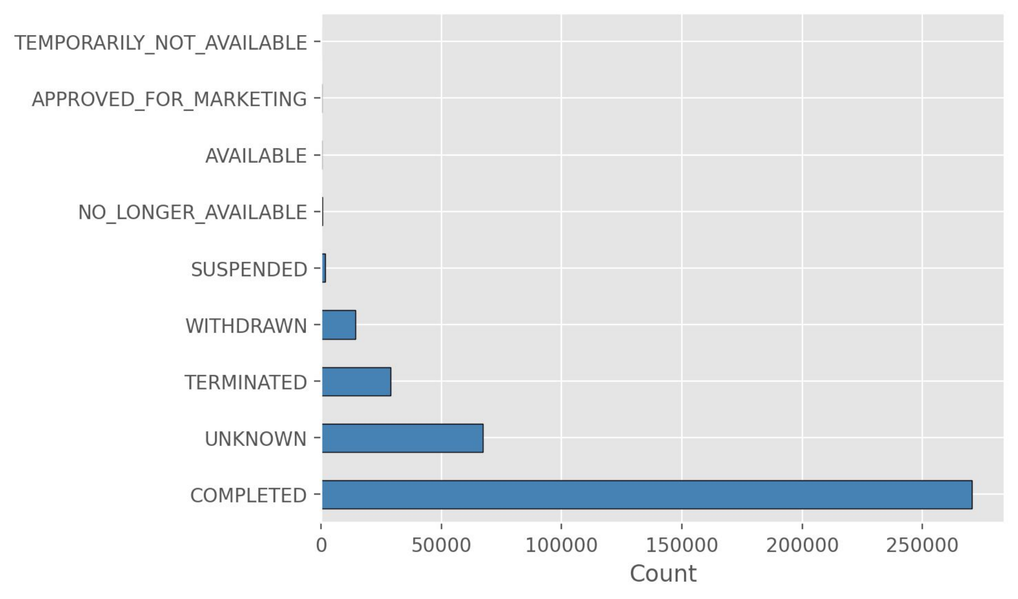
Challenges

- Selection of fields:
  - 115 out of 568 fields selected for initial download and exploratory analysis
  - Field definitions vague and unreliable
- New version of API debuted in March, 2024
- Fields contain multiple, nested entries
- Many fields contain free-text data with 100,000s of unique values
- 69 fields selected for encoding and analysis

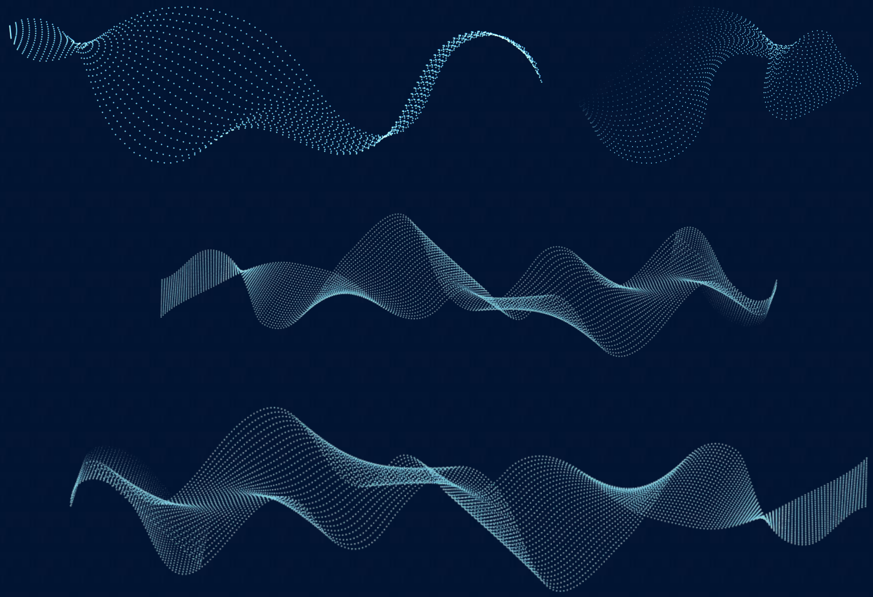


Study Overall Status

- Target class is unbalanced
- ‘Approved for marketing’ is an infrequent class
- Expanded Access records have different characteristics



Simplified Model

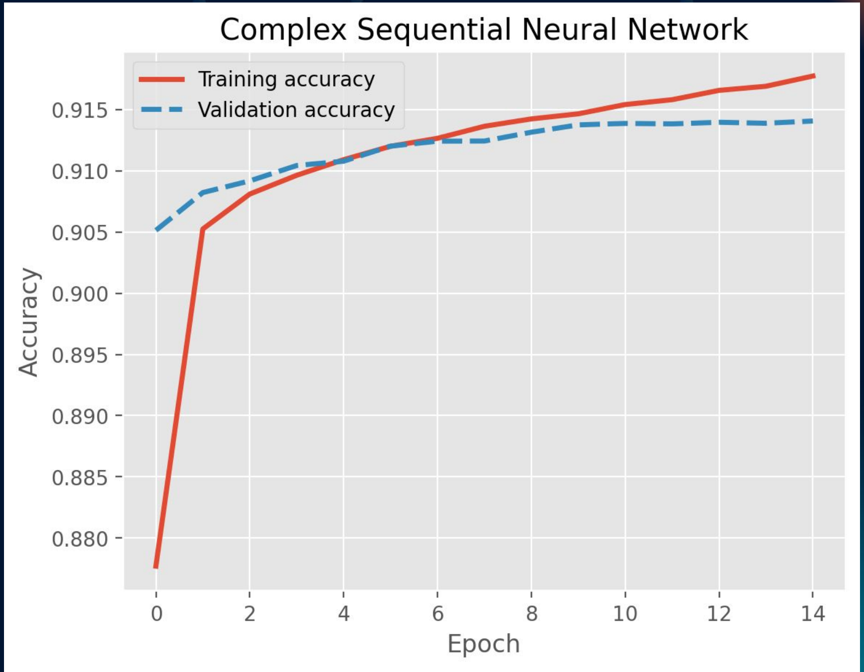


Binary classification

- Expanded access records excluded
- Develop a model to predict status ‘Completed’
- Combine ‘Suspended’, ‘Terminated’, ‘Withdrawn’, and ‘Unknown’ (abandoned) statuses as ‘Not Complete’

Model Results

Model	Accuracy	Balanced Accuracy	Train Time
Baseline	*70.7%*		
Logistic Regression	90.9%	84.8%	50s
Logistic Regression with SMOTE	89.5%	85.7%	381s
Basic Neural Network	91.2%	86.0%	21s
Complex Sequential Neural Network	91.4%	86.0%	187s

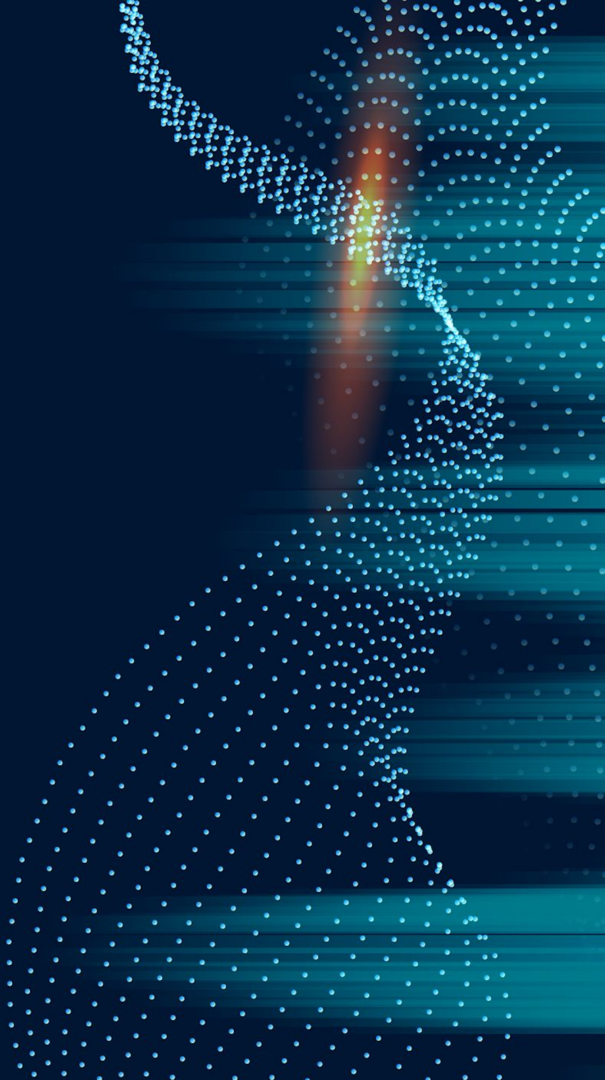


Logistic Regression Coefficients

Top 10			Bottom 10	
Feature	Coef		Feature	Coef
CompletionDateType_ACTUAL	2.385820		Phase_No_data	-0.161504
LocationStatus_No_data	1.711084		LocationCountry_No_data	-0.217455
CentralContactRole_No_data	1.225026		OrgClass_NIH	-0.219640
PrimaryCompletionDateType_ACTUAL	1.168753		PrimaryCompletionDate	-0.226498
CompletionDateType_No_data	1.091474		ReferenceType_No_data	-0.321632
StudyFirstSubmitDate	0.498214		LocationStatus_RECRUITING	-0.423382
PrimaryCompletionDateType_No_data	0.388943		CentralContactRole_CONTACT	-0.702567
ReferenceType_DERIVED	0.382309		StartDateType_ESTIMATED	-1.489931
CompletionDate	0.328414		PrimaryCompletionDateType_ESTIMATED	-1.509925
StartDateType_No_data	0.310194		CompletionDateType_ESTIMATED	-3.270344

Conclusions

- Data in ClinicalTrials.gov holds valuable insight to conducting successful clinical trials
- Snapshotting data to multiple time points in clinical trial design and progress would enable the best benefits of predictive modelling
- Detailed protocol design and study results data is ripe for detailed NLP analysis



Thank you!

All data sourced from ClinicalTrials.gov

- Comprehensive API documentation on ClinicalTrials.gov

Research

- Data Science by Nicholas McBride, PhD
- Supported by Adobe Digital Academy