U.S. FOOD & DRUG ADMINISTRATION

# Value Creation of Big Data from the Regulatory Setting

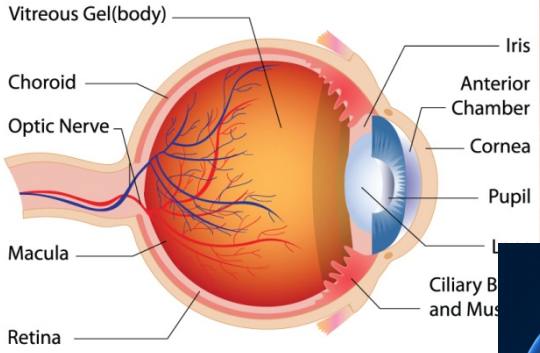
Liang Zhao, PhD

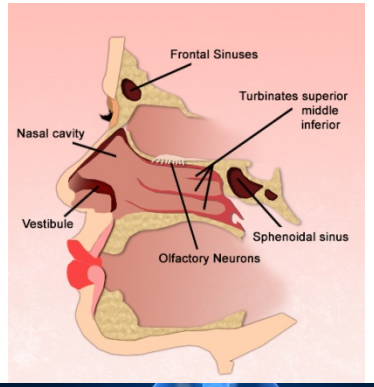
CBA 2017-2018 Workshop Series-3
**February 3rd , 2018, Rockville, MD**

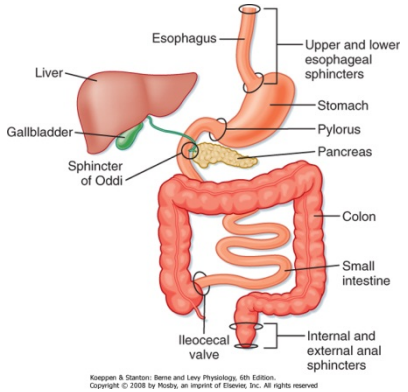
***Disclaimer: My remarks today do not necessarily reflect the official views of the FDA***

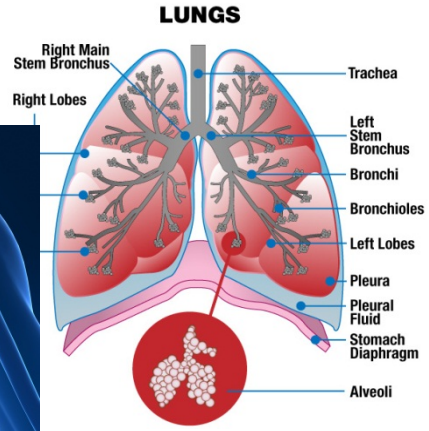
# Human Body are Reality of Big Data

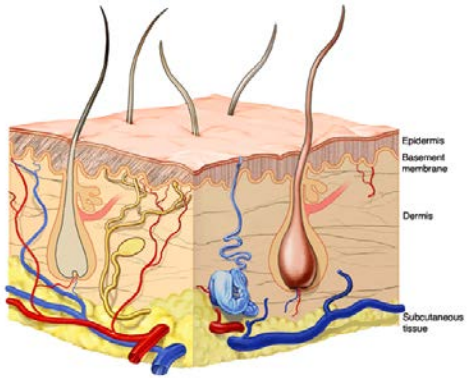
**Drug Substance Formulations In Vitro Testing** → **Physiological System** → **In Vivo Performance**



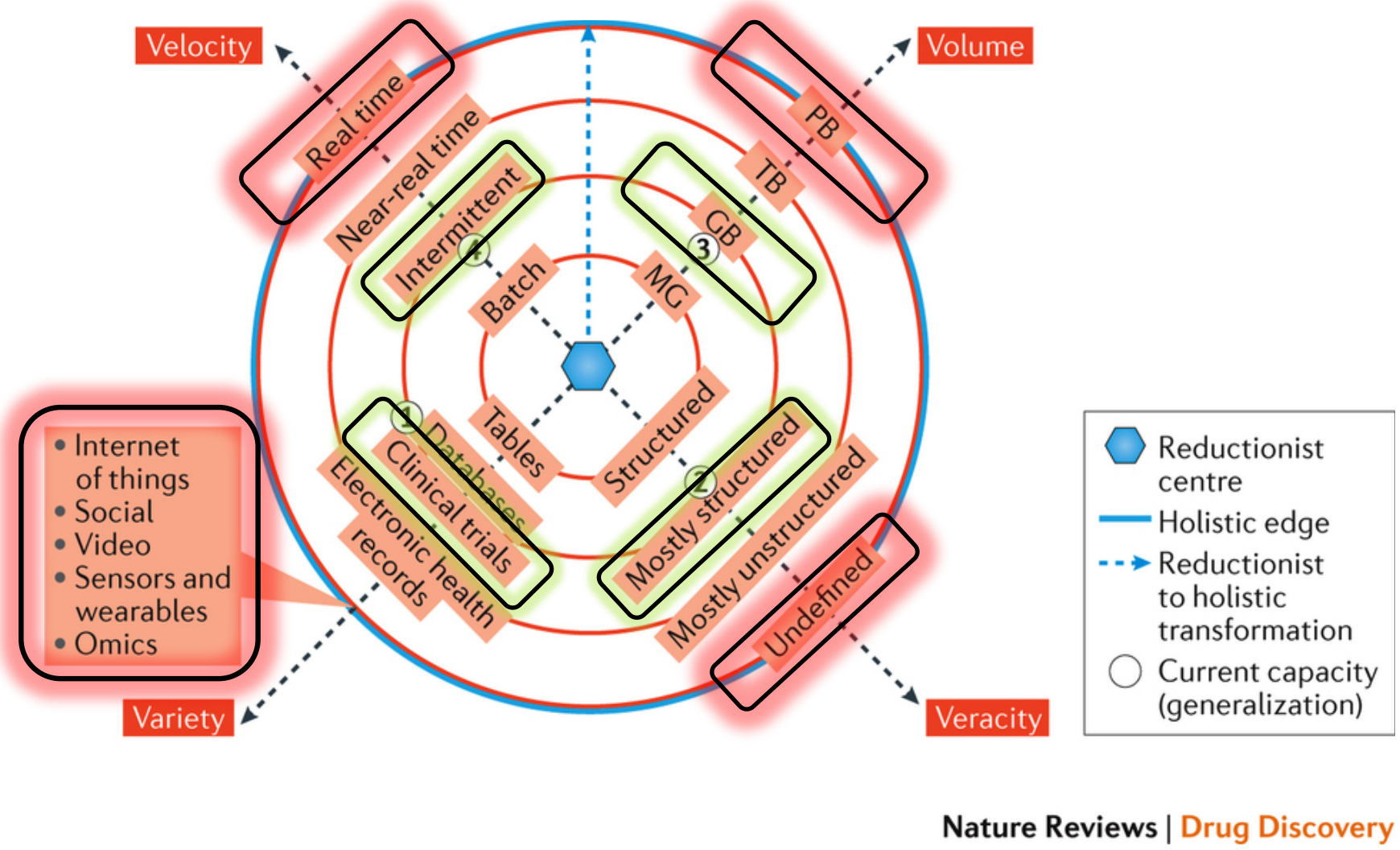




Based on the publication by Jiang W, Kim S, Zhang X, Lionberger RA, Davit BM, Conner DP, Yu LX. Int J Pharm. 2011 Oct 14;418(2):151-60.

# Examples of Big Data Related Activities at Agency

- Real world study
- Human genome-based data
- Post-market evaluations
- Precision medicine/Digital health

# From Preclinical/Clinical to Real World




Nature Reviews | Drug Discovery

Khozin S, et al. Nat Rev Drug Discov. 2017.

# HIVE to Assist Big Data Review on Human Genome-based Data

**Storage:** ~2 Petabytes (comparable to 1 million HD movies), metal + SunGrid

**CPU:** 1500 cores, extensible to 3000–5000

**Network:** 10Gb ⇒ Internet2, 40Gb ⇒ Infiniband

**Mini-hive:** Research and scientific NGS portal with cutting edge production quality tools, White Oak/CBER server room

**Storage:** ~500 Terabytes, metal

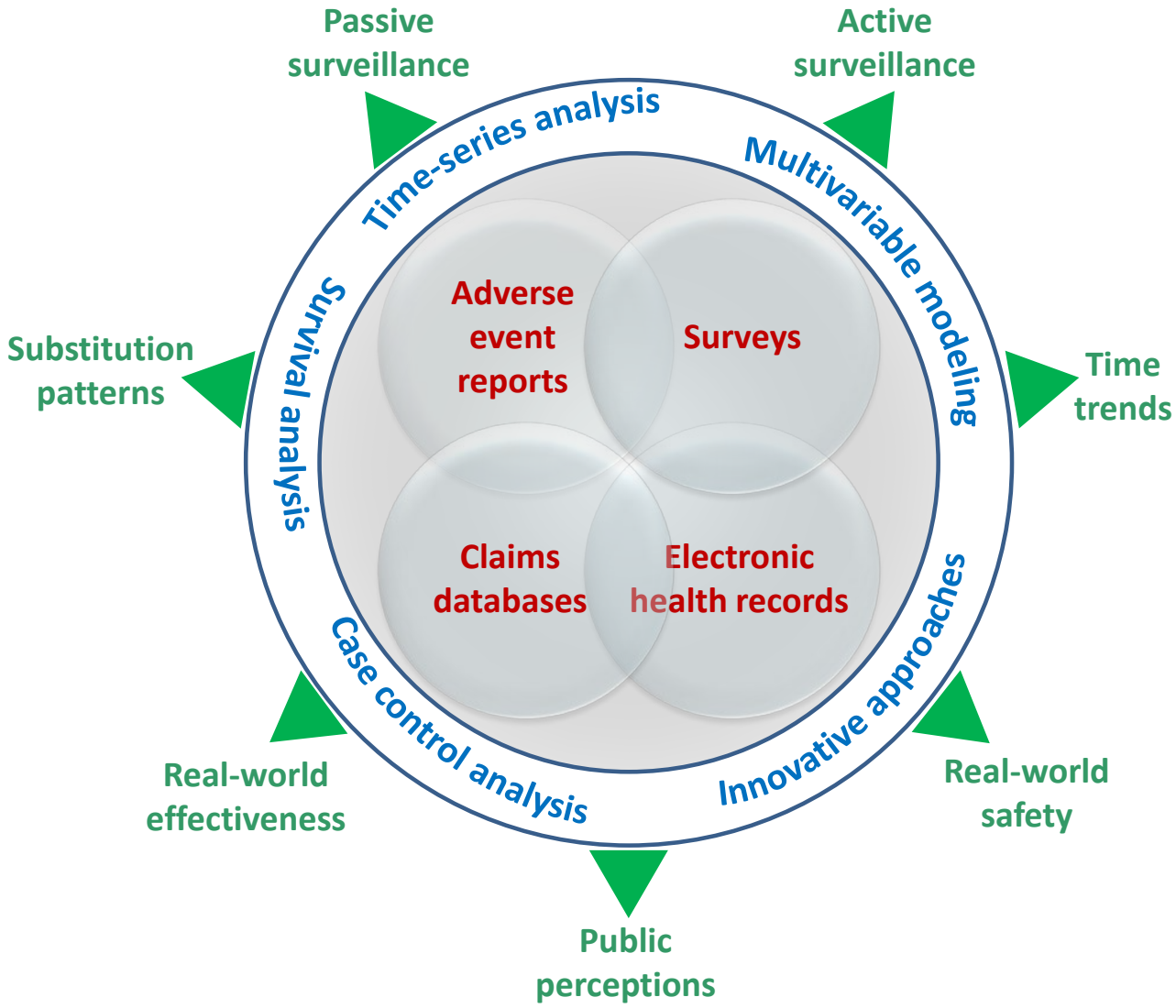
**CPU:** ~350 cores

**Network:** wan 1Gb, lan 40 GB

**Questions?** Contact FDA's Office of Media Affairs at 301-796-4540 or fdaoma@fda.hhs.gov

High-performance computation infrastructure performing NGS bioinformatics computations that are massively parallel (executed on multiple computers simultaneously)

# Quantitative Approaches in the Post-Marketing Evaluation of Generics



# Digital Health

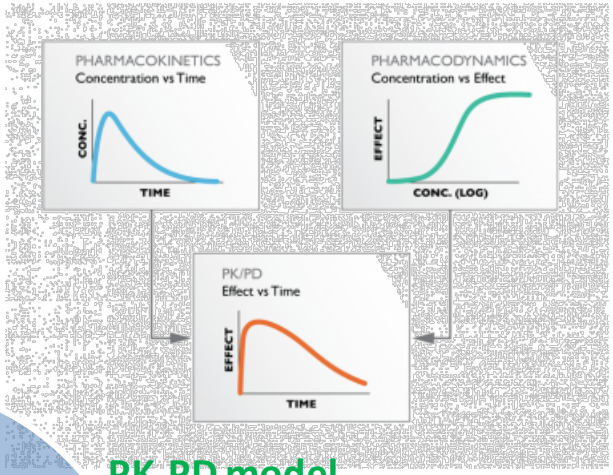
- Convergence of digital and genomic technologies with health, healthcare, living, and society to enhance the efficiency of healthcare delivery and make medicines more personalized and precise
- Use of information and communication technologies to help address the health problems and challenges faced by patients
- These technologies include both hardware and software solutions and services, including telemedicine, web-based analysis, email, mobile phones and applications, text messages, and clinic or remote monitoring sensors

https://en.wikipedia.org/wiki/Digital_health

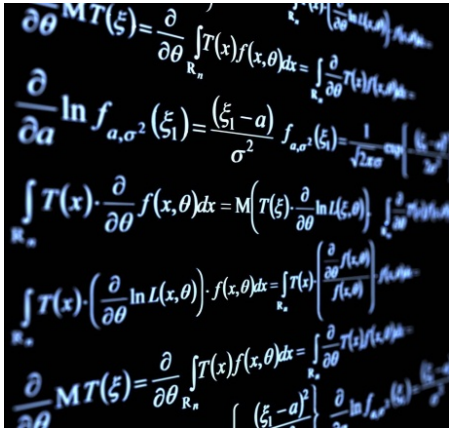
# Big Data vs Conventional Methods

Non-Oral Drug

Oral Drug




PK-PD model

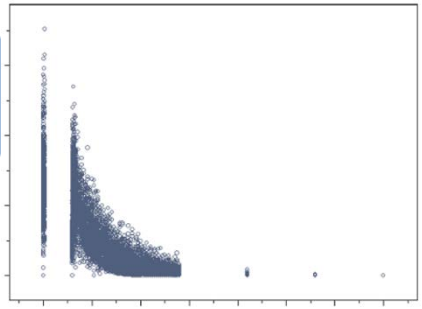


Release/ Absorption/ PBPK Models

Big Data

Pharmaco-metrics

Machine learning toolsets
Analytics for complex mixtures
Systems pharmacology
Risk-based models
Business process models



Population based model

# A Case Example

# Big Data to Understand Relationship Between the Biological Targets and Adverse Reactions for TKIs

- Tyrosine kinase inhibitors (TKIs): one of the most important classes of anti-cancer drugs
- Adverse reactions (ARs) by both on-target and off-target effects of TKIs
- Understanding the mechanisms of ARs are important for both drug development and post-market evaluation of other agents
- Past research are mainly based on summarization of clinical practices or in vitro/in vivo experiments
- Meta-analysis intends to take advantage of both vast individual data from registrational Phase 3 studies and the advancement of cutting edge quantitative methodologies

# Adverse Reactions of KIs

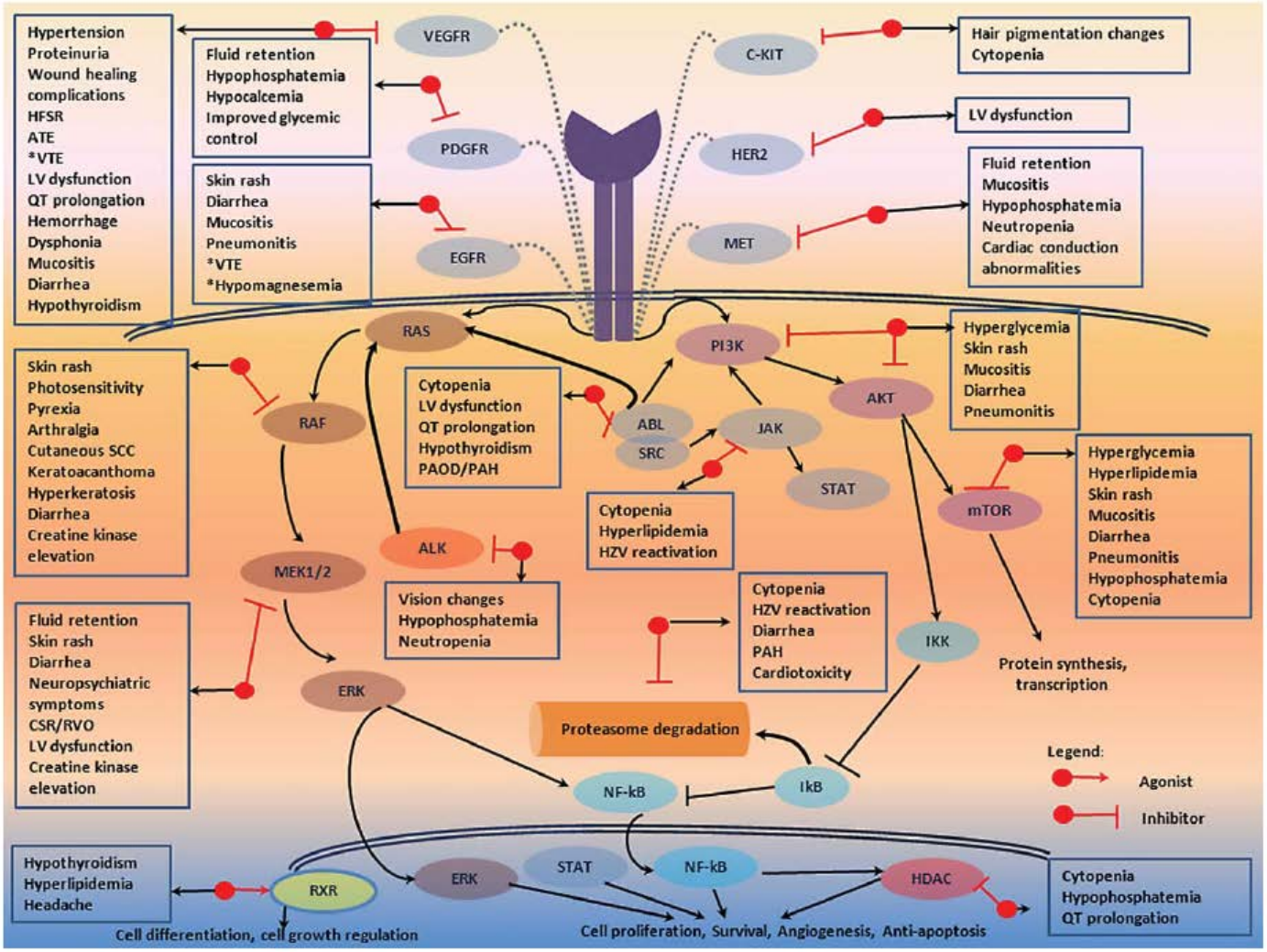



FIGURE 1. Toxicities Associated With Signal Transduction Inhibitors.*Associated predominantly with monoclonal antibodies. ATE indicates arterial thromboembolism; CSR, central serous retinopathy; HZV, herpes zoster virus; LV, left ventricular; PAH, pulmonary arterial hypertension; PAOD, progressive arterial occlusive disease; RVO, retinal vein occlusion; SCC, squamous cell cancer; VTE, venous thromboembolism.

Grace K. Dy and Alex A. Adjei, CA Cancer J Clin 2013; 63: 249–279

Results

# Data from 17 Kinase Inhibitors

| | Kinase Inhibitors (KIs) |
|---|---|
| 1 | Axitinib (Inlyta) |
| 2 | Pazopanib (Votrient) |
| 3 | Sorafenib (Nexavar) |
| 4 | Vandetanib (Caprelsa) |
| 5 | Crizotinib (Xalkori) |
| 6 | Erlotinib (Tarceva) |
| 7 | Gefitinib (Iressa) |
| 8 | Lapatinib (Tykerb) |
| 9 | Bosutinib (Bosulif) |
| 10 | Dasatinib (Sprycel) |
| 11 | Imatinib (Gleevec) |
| 12 | Nilotinib (Tasigna) |
| 13 | Sunitinib (Sutent) |
| 14 | Cabozantinib (Cometriq) |
| 15 | Ponatinib (Iclusig) |
| 16 | Regorafenib (Stivarga) |
| 17 | Afatinib (Gilotrif) |

**17 KIs**

1. Incidence of adverse reactions (ARs)
2. Inhibitory percent (%) data against 283 kinases

Reference for inhibitory percent data:
Uitdehaag JC et al. PLoS One. 2014 Mar; 9(3): e92146

**13 KIs**

1. Pharmacokinetic (PK) data
2. Dissociation constant ($K_d$) data against 257 kinases

Reference for $K_d$ data:
Davis MI et al. Nat Biotechnol. 2011 Oct; 29(11): 1046-51
Karaman MW et al. Nat Biotechnol. 2008 Jan; 26(1): 127-32

# Aim and Methods Outline

Aim: to assess the association between kinase inhibition and adverse reactions

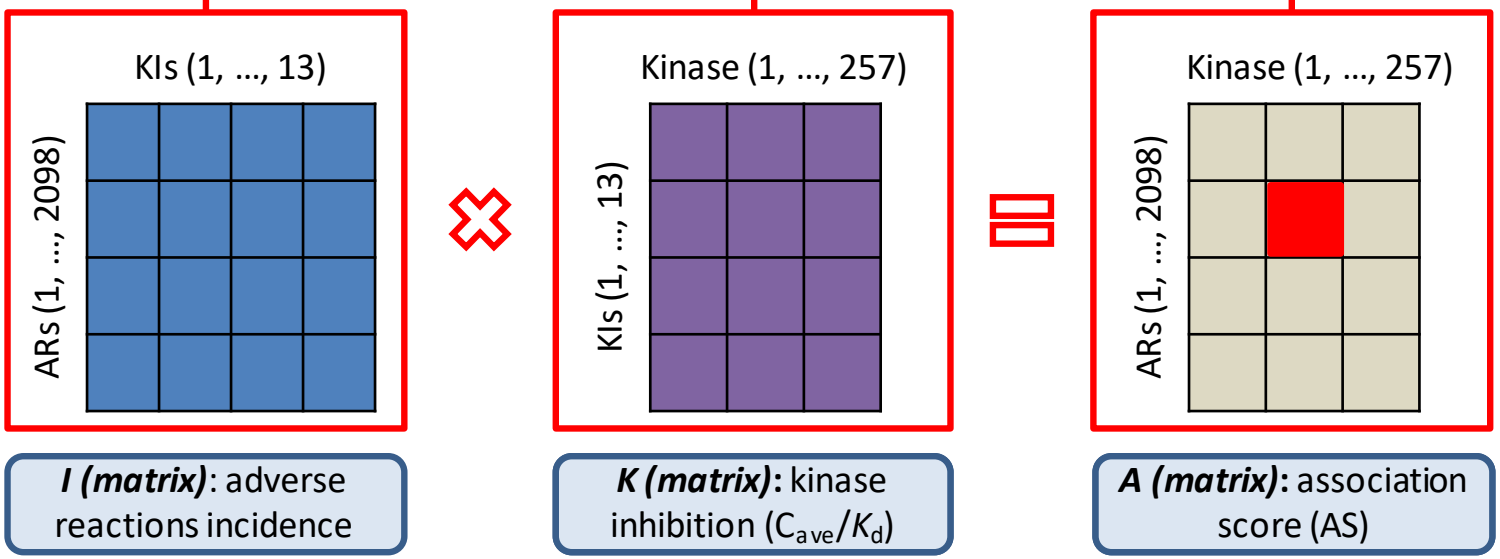
| A (drug exposure, PK) | B (*in vitro* kinase inhibitory activity) | C (incidence of adverse reactions) |
|---|---|---|
| 1. Collect pharmacokinetic (PK) data of current FDA-approved KIs<br><br>2. Conduct population PK modeling to calculate $C_{ave}$ (average plasma concentration) at steady state | 1. Literature search to collect constant dissociation ($K_d$) data<br><br>2. Literature search to collect inhibitory percent (%) data | 1. Collect safety data of current FDA-approved KIs<br><br>2. Standardize ontology of adverse reactions (ARs)<br><br>3. Calculate incidence of ARs |

| D (association score) | An association score matrix of 2098 ARs (preferred terms) and 257 kinases |
|---|---|

Methods

# Association Score Matrix

$$\textbf{Association score: } I \times K = \sum(i_{zx}) \times \left(\frac{C_{ave}}{k_{xy}}\right) = A$$




***I (matrix)***: adverse reactions incidence

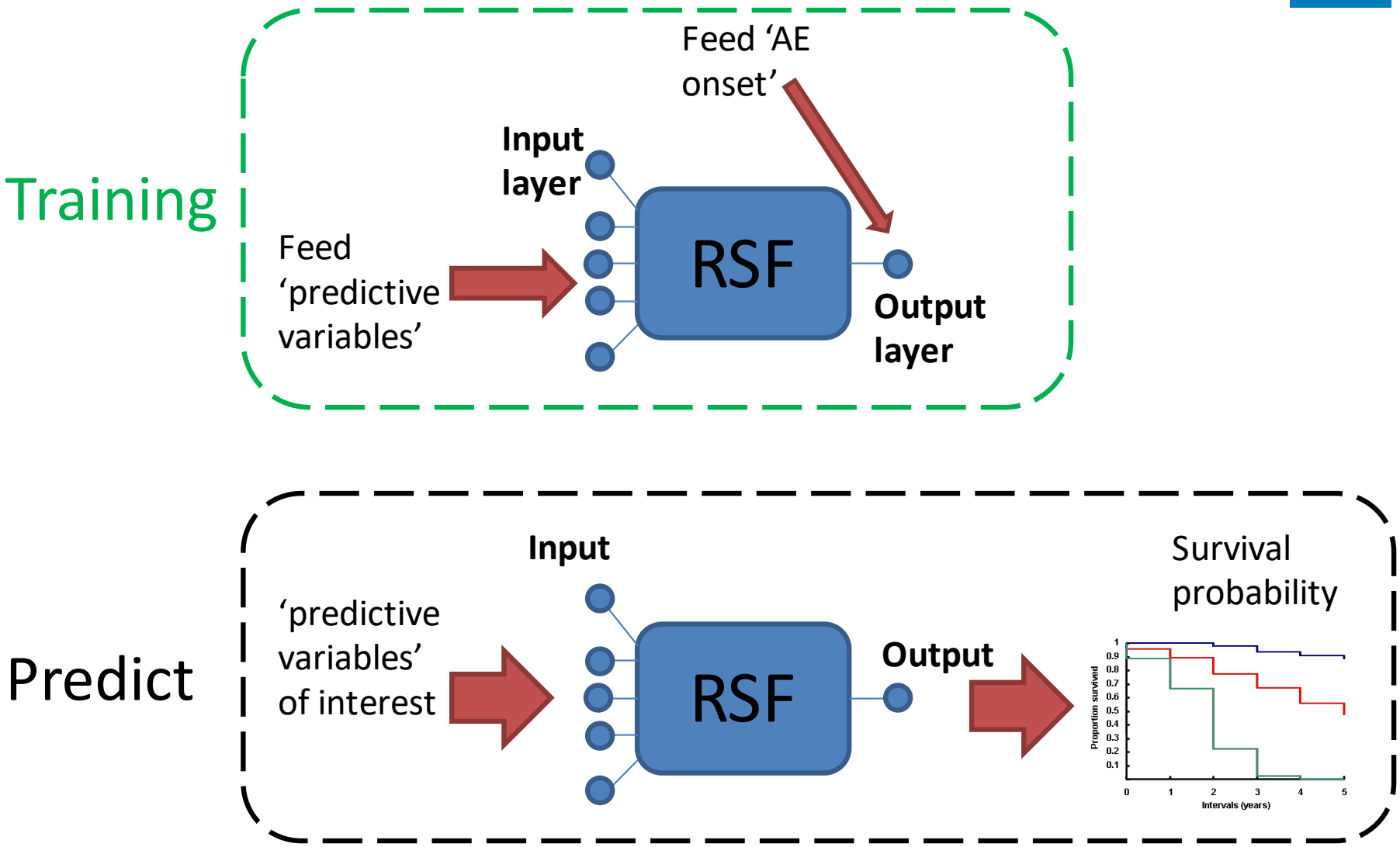
***K (matrix)***: kinase inhibition ($C_{ave}/K_d$)

***A (matrix)***: association score (AS)

| | |
|---|---|
| Limitation | A false positive may be included when a high association score was obtained with high AR incidence but moderate kinase inhibition. |

| | |
|---|---|
| Solution | After identifying AR associated KIs, only keep the preliminary identified kinases (by association score) which can be inhibited with > 95% activity by any identified KIs. |

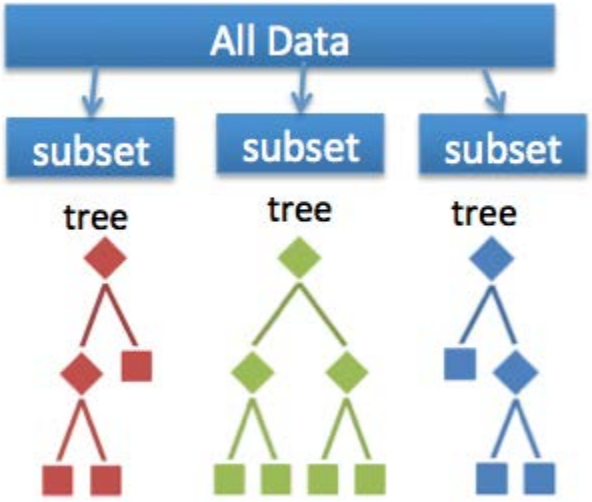
# Random Survival Forest



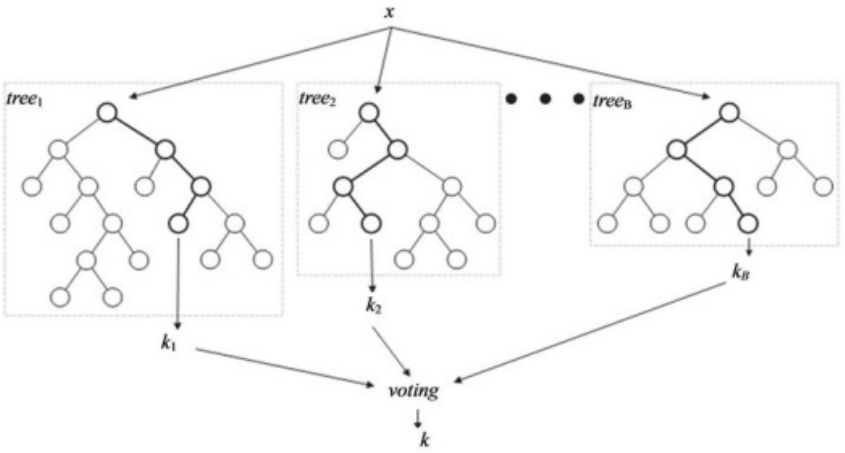
# Random Survival Forest

- Decision survival tree shares the same pitfall with the decision tree, as a 'greedy' algorithm.
- Random survival forest was developed to improve the decision survival tree.

## Training




## Prediction




https://mapr.com/blog/predicting-loan-credit-risk-using-apache-spark-machine-learning-random-forests/
http://www.hallwaymathlete.com/2016/05/introduction-to-machine-learning-with.html

# Results

4279 pairs of associations involving 534 ARs (preferred terms) and 140 kinases.

**Well-established pairs of kinase inhibition and ARs were confirmed:**

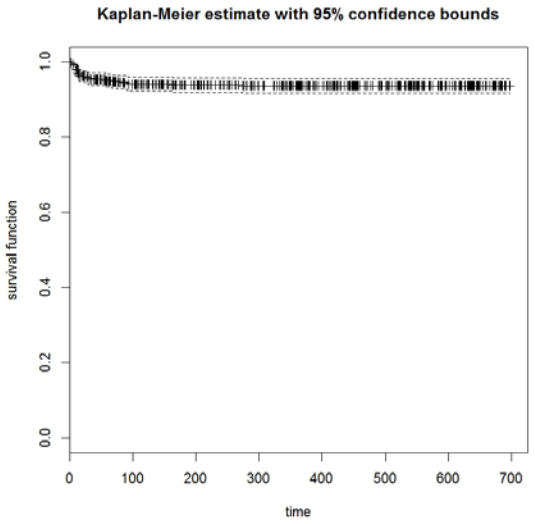
hypertension – VEGFR2;
acneiform rash – EGFR/HER4;
conjunctivitis – EGFR;
fluid retention – ABL;
hepatotoxicity – MET;
diarrhea – EGFR;
pulmonary hypertension – ABL;
QT prolongation – VEGFR;
proteinuria – VEGFR.

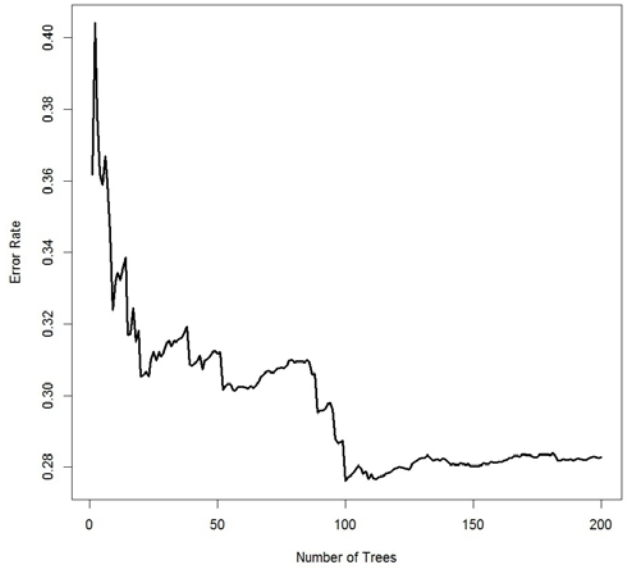
Visualize the results using a web app:
https://jzliu.shinyapps.io/KINASE

# Machine Learning Results

Consistent with DPA and BCPNN finding in general

Dermatitis acneiform as an example



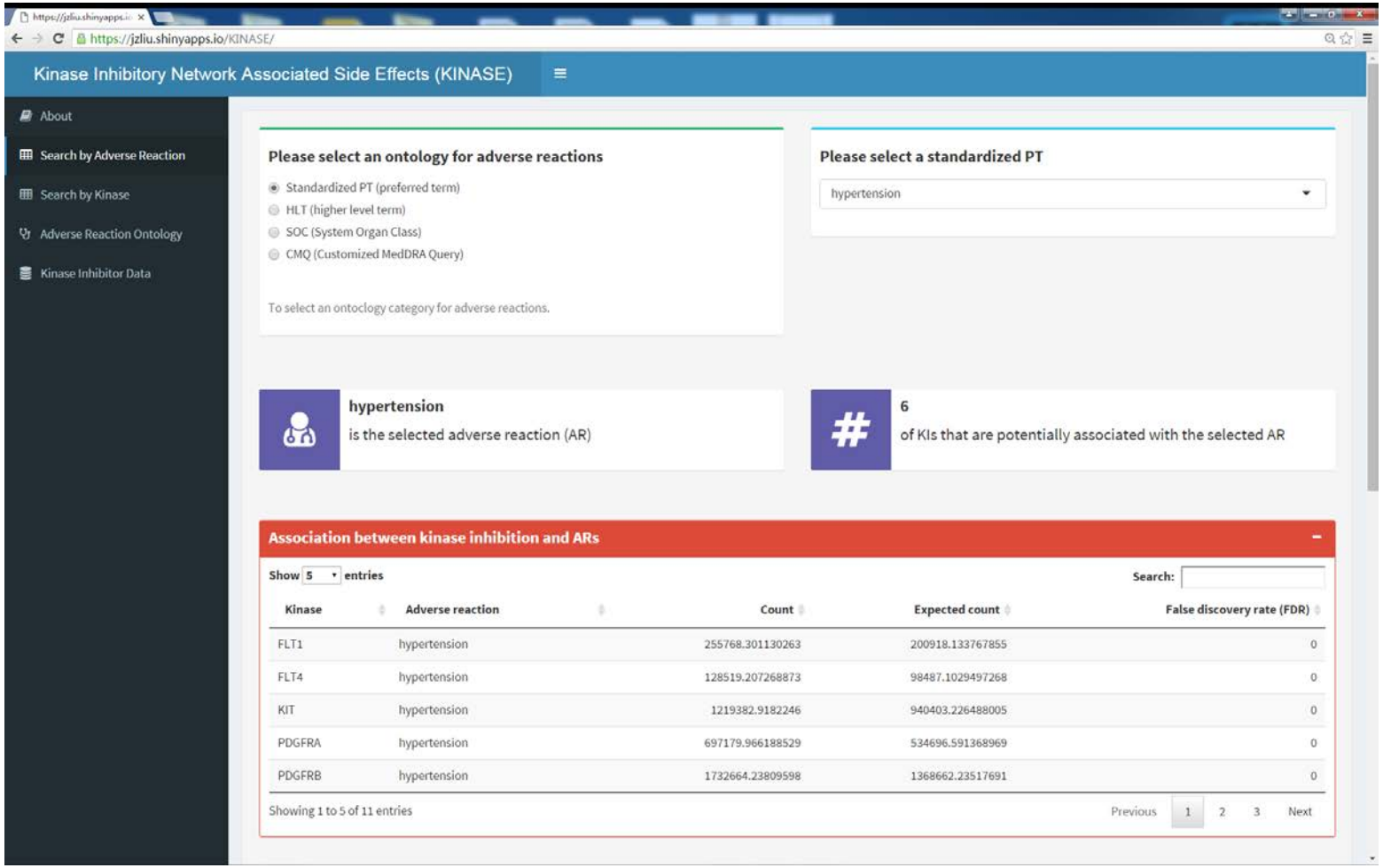

| | Importance | Relative Imp |
|---|---|---|
| EGFR.G719S. | 0.0086 | 1.0000 |
| EGFR.L747.T751del.Sins. | 0.0073 | 0.8550 |
| EGFR.E746.A750del. | 0.0072 | 0.8395 |
| EGFR | 0.0053 | 0.6214 |
| EGFR.L747.E749del..A750P. | 0.0051 | 0.6012 |
| EGFR.L747.S752del..P753S. | 0.0050 | 0.5866 |
| JAK2.JH1domain.catalytic. | 0.0050 | 0.5817 |
| EGFR.T790M. | 0.0048 | 0.5642 |
| EGFR.G719C. | 0.0039 | 0.4608 |
| EGFR.L858R.T790M. | 0.0036 | 0.4156 |
| MKNK1 | 0.0033 | 0.3856 |
| JAK3.JH1domain.catalytic. | 0.0032 | 0.3746 |
| ADCK4 | 0.0031 | 0.3653 |
| ERBB2 | 0.0030 | 0.3526 |
| DRAK1 | 0.0028 | 0.3302 |
| TYK2.JH1domain.catalytic. | 0.0025 | 0.2972 |
| SYK | 0.0025 | 0.2941 |
| JNK2 | 0.0025 | 0.2903 |
| EGFR.L858R. | 0.0024 | 0.2856 |

Work in progress and manuscript is accepted

DPA: Disproportionality Analysis; BCPNN: Bayesian Confidence Neural Network

Results

# KINASE: A Web App to Query the Results




YouTube by Dr. Liu: https://www.youtube.com/watch?v=O1kqbWFqhwc&t

# Summary for the Case

- Meta-analyses are based on Phase 3 data from 17 TKIs
- Analysis results for associations between kinases inhibitions and adverse reactions are consistent with research finding
- Caveat should be given before experimentally verifying other associations or claiming a causal relationship
- Novel methods including machine learning techniques can be used for analysis

# Take Home Message: (Big) Data Driven Decisions Makings in the Agency

Big data

Social media

Commercial/Sales

NDA

Internet

Literature

Post Market

## Data Mining/Selection

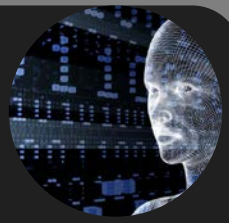
Proactively planned data collection

Secondary databases including Sentinel

Research/ Guidance Databases

Relational databases

## Analytics including Artificial Intelligence



Drug Disc.

Drug Develop.

Research

Generic Drug Competition-Pharmacoeconomics

Real world Assessment

# The World of Big Data

# Acknowledgement


**DOPI/OHOP/OND/CDER**

- **Geoffrey Kim**, M.D.
- **James Xu**, M.D.
- **Amy McKee**, M.D.

**ORS/OGD/CDER**

- **Jinzhong Liu**, Ph.D.
- **Meng Hu**, Ph.D.
- **Xiajing Gong,** Ph.D.
- **Robert Lionberger**, Ph.D.

**DHOT/OHOP/OND/CDER**

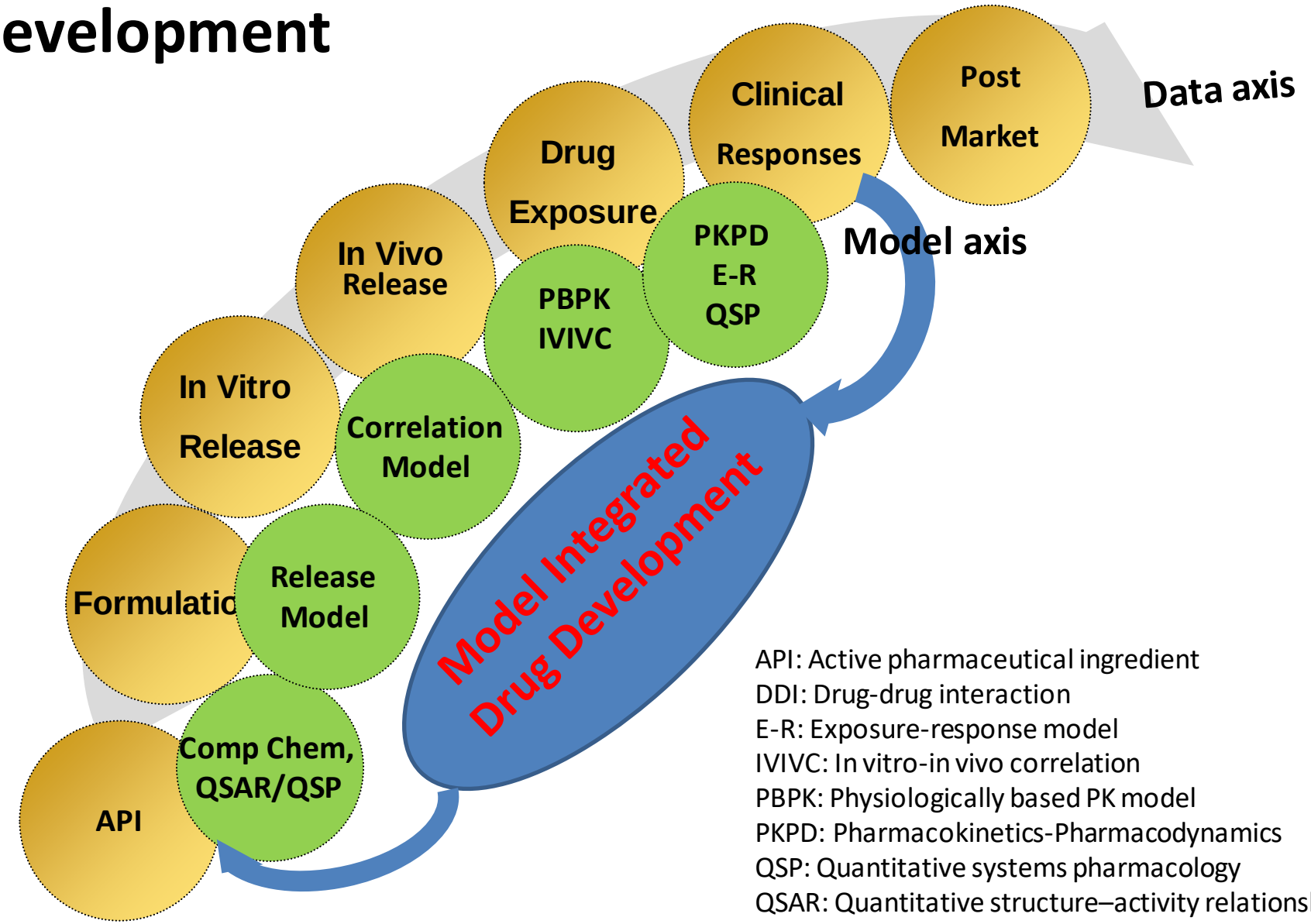
- **Todd Palmby**, Ph.D.

**DHP/OHOP/OND/CDER**

- **Angelo DeClaro**, M.D.

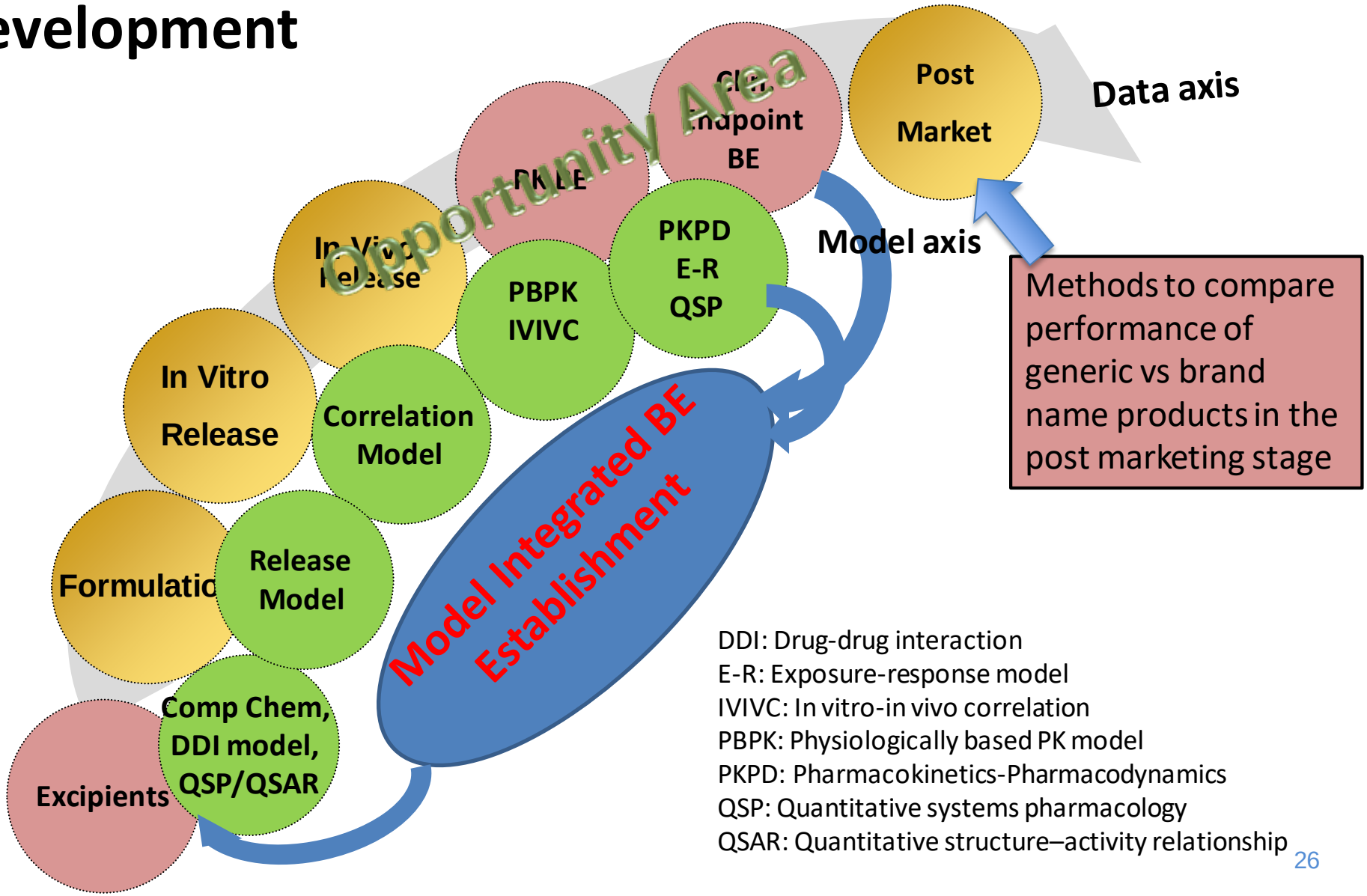
# Backups

# An Integrated Modeling System for New Drug Development




API: Active pharmaceutical ingredient
DDI: Drug-drug interaction
E-R: Exposure-response model
IVIVC: In vitro-in vivo correlation
PBPK: Physiologically based PK model
PKPD: Pharmacokinetics-Pharmacodynamics
QSP: Quantitative systems pharmacology
QSAR: Quantitative structure–activity relationship

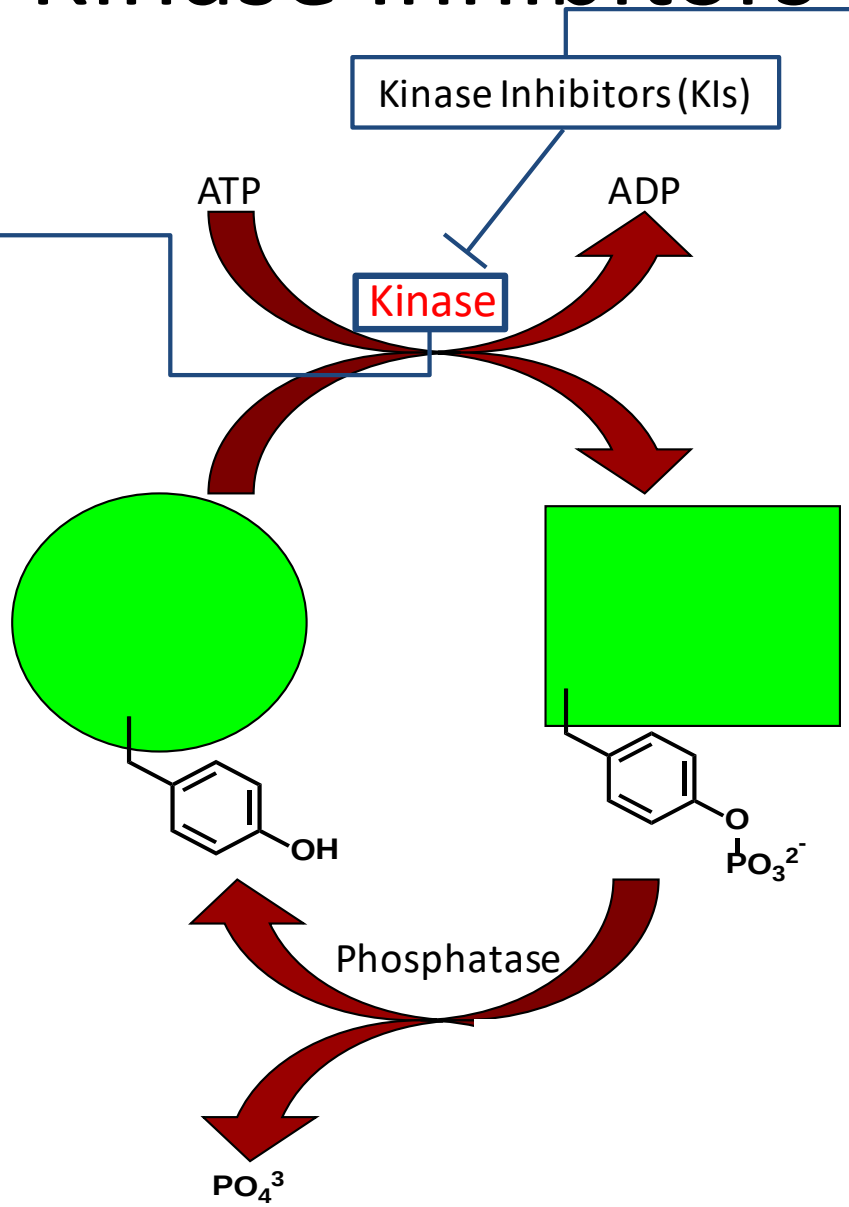
# An Integrated Modeling System for Generic Drug Development




DDI: Drug-drug interaction
E-R: Exposure-response model
IVIVC: In vitro-in vivo correlation
PBPK: Physiologically based PK model
PKPD: Pharmacokinetics-Pharmacodynamics
QSP: Quantitative systems pharmacology
QSAR: Quantitative structure–activity relationship

# Kinase Inhibitors

A kinase is a type of enzyme that transfers phosphate groups from high-energy donor molecules (such as ATP) to specific substrates, a process referred to as phosphorylation.

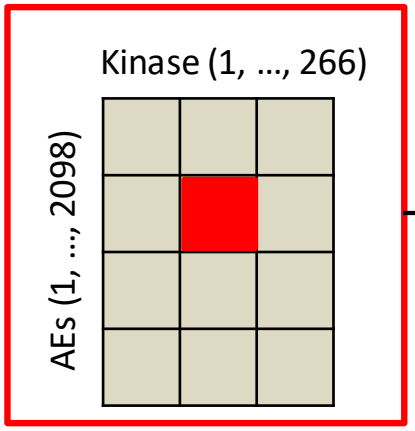



Kinase includes many oncogenes, so phosphorylation by kinases is a necessary step in some cancers.

Kinase inhibitors are used as drugs to treat these cancers by inhibiting kinases.

To identify kinases associated with hypertension

# An Example




Preliminary identified kinases leading to hypertension: VEGFR1, VEGFR2, VEGFR3, KIT, PDGFR α, PDGFR β, TTK, … (27 kinases in total)

Identify hypertension associated KIs: pazopanib, axitinib, regorafenib, sorafenib, vandetanib, cabozantinib (6 KIs in total)

Only keep identified kinases which can be inhibited with > 95% activity by any identified 6 KIs.

| Kinase | # Inhibition |
|---|---|
| VEGFR2 | 6 |
| FLT1 | 6 |
| FLT4 | 6 |
| PDGFRA | 6 |
| PDGFRB | 5 |
| FGFR2 | 5 |
| KIT | 4 |
| FGFR3 | 3 |
| FGFR1 | 2 |
| RAF1 | 2 |
| AURKC | 1 |

# Machine Learning for Correlation Identification

$C_{AVG}$/Kd

| Subj# | Age | Gender | PT | AE_onset | $K_1$ | $K_2$ | … | $K_p$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 53 | M | A | 12 | $X1_1$ | $X1_2$ | … | $X1_p$ |
| 1 | 53 | M | A | 26 | $X1_1$ | $X1_2$ | … | $X1_p$ |
| 1 | 53 | M | B | 6 | $X1_1$ | $X1_2$ | … | $X1_p$ |
| … | … | … | … | … | … | … | … | … |
| 1 | 53 | M | Z | 130 | $X1_1$ | $X1_2$ | … | $X1_p$ |
| 2 | 48 | F | B | 3 | $X2_1$ | $X2_2$ | … | $X2_p$ |
| 2 | 48 | F | B | 78 | $X2_1$ | $X2_2$ | … | $X2_p$ |
| … | … | … | … | … | … | … | … | … |
| N | 59 | F | Y | 58 | $XN_1$ | $XN_2$ | … | $XN_p$ |

**The time factor is taken into account!**

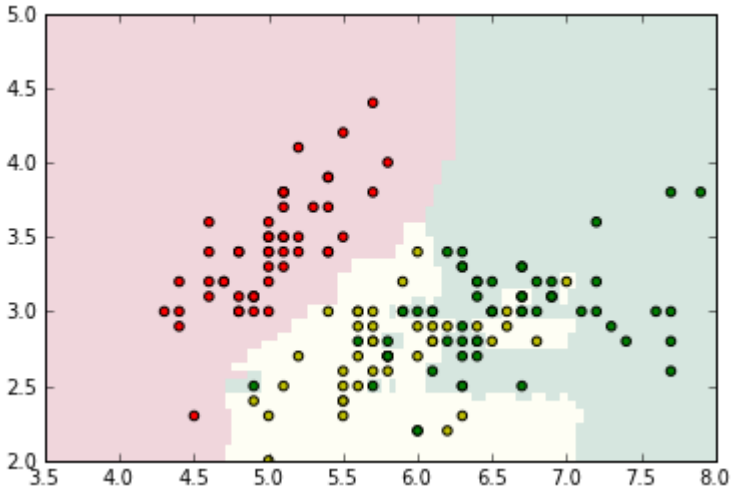
# Traditional methods

- Regression-based
  - Proportional hazards model
  - Accelerated failure time model
  - *Cox model* (semi-parametric)

- Issues
  - Distribution assumption
  - Model is difficult to converge due to large number of predictive variables
  - Linear relationships

# Machine learning

- Machine-learning-based
  - Artificial neural network
  - *Random forest*
  - Support vector machine

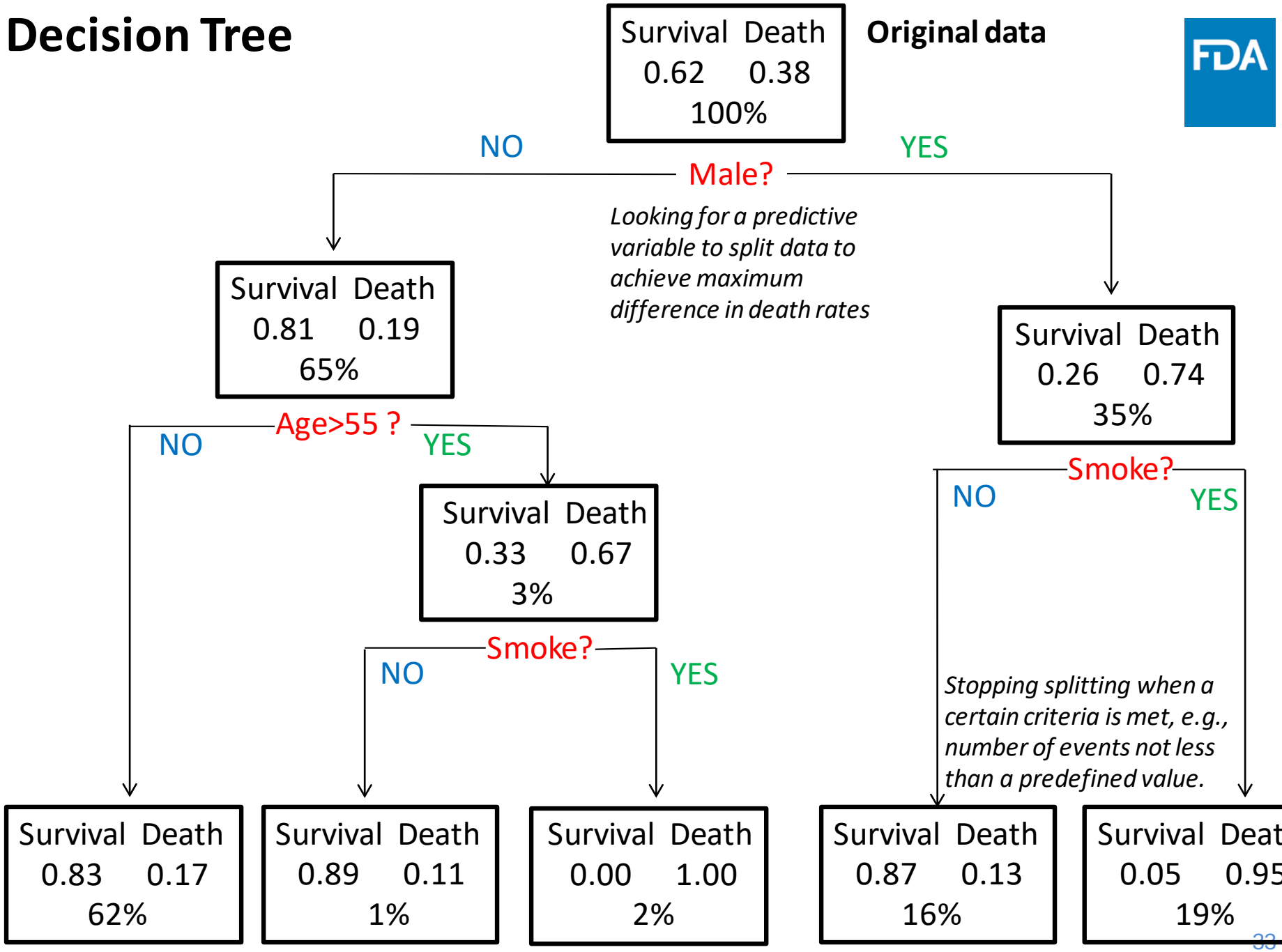



- Advantages
  - Less distribution assumption
  - Capable for large-feature problem
  - Nonlinear relationship
  - Able to describe the variable-variable interaction

# Random survival forest

- Artificial neural network
  - Over-learning
  - Inconvenient to identify importance of variable
- Support vector machine
  - Inconvenient to identify importance of variable
- ***Random survival forest***
  - Bagging (or boosting) technique to prevent from over learning
  - Established method to identify importance of variable
    - Variable importance
    - Minimal depth
    - Variable hunting

# Decision Tree



# How to grow a decision tree

- How to split
  - Searching a predictive variable to maximize event (e.g., death rates) difference between daughter nodes
- How to stop
  - A certain criteria is met, e.g., number of events no less than a certain value

# Why random forest?

- Decision tree is a '**greedy**' algorithm.
  - For example, given coins with values of 1, 15, 25 cents, how to get 30 cents using less coins.
  - Greedy: 30=25+1+1+1+1+1
  - Optimal: 30=15+15

- Decision tree is prone to over-learning or over-fitting.

- Random forest consists of many decision trees, each of which grows by a part of data and predictive variables.

# Relevant Research from Agencies

- GDUFA I supported the build out of the modeling and simulation tool chain for generic drugs