Secondary use of historical clinical data in meta-analysis

Hao Meng, Hong Wang

July 26th , 2018

Disclaimer

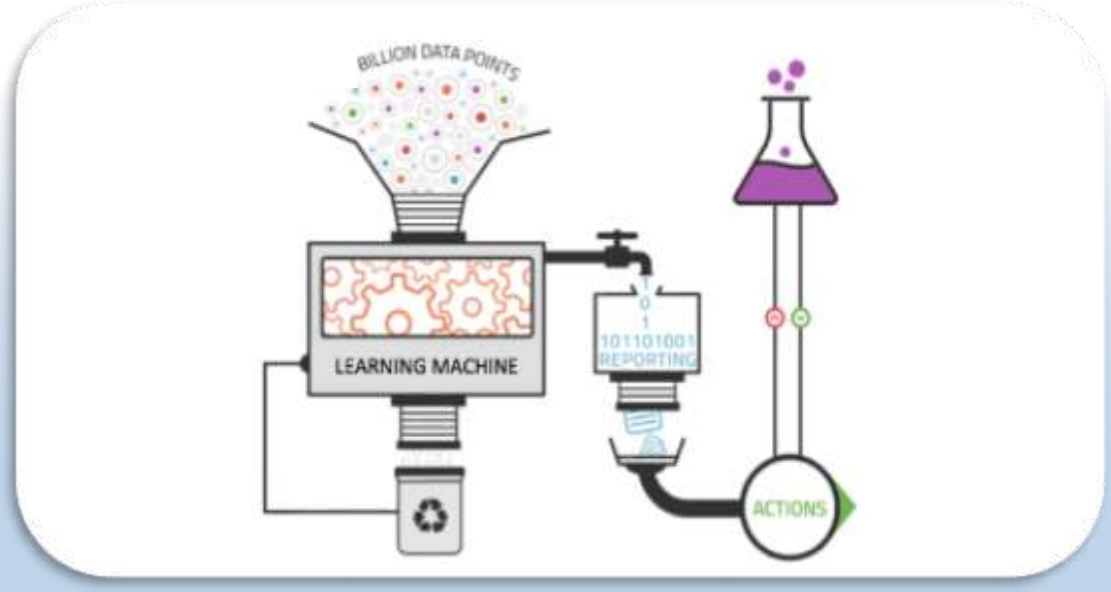
The views expressed in this presentation are our personal opinions and not those of Boehringer Ingelheim Pharmaceuticals, Inc.

Outline

- 1. Introduction
- 2. Data processing
- 3. Innovative programming methods in meta-analysis

Introduction

– A lot of pieces of clinical trial data, reuse them is hot topic
  • Artificial Intelligence, Machine learning, Big Data…..



How secondary use of data helps?

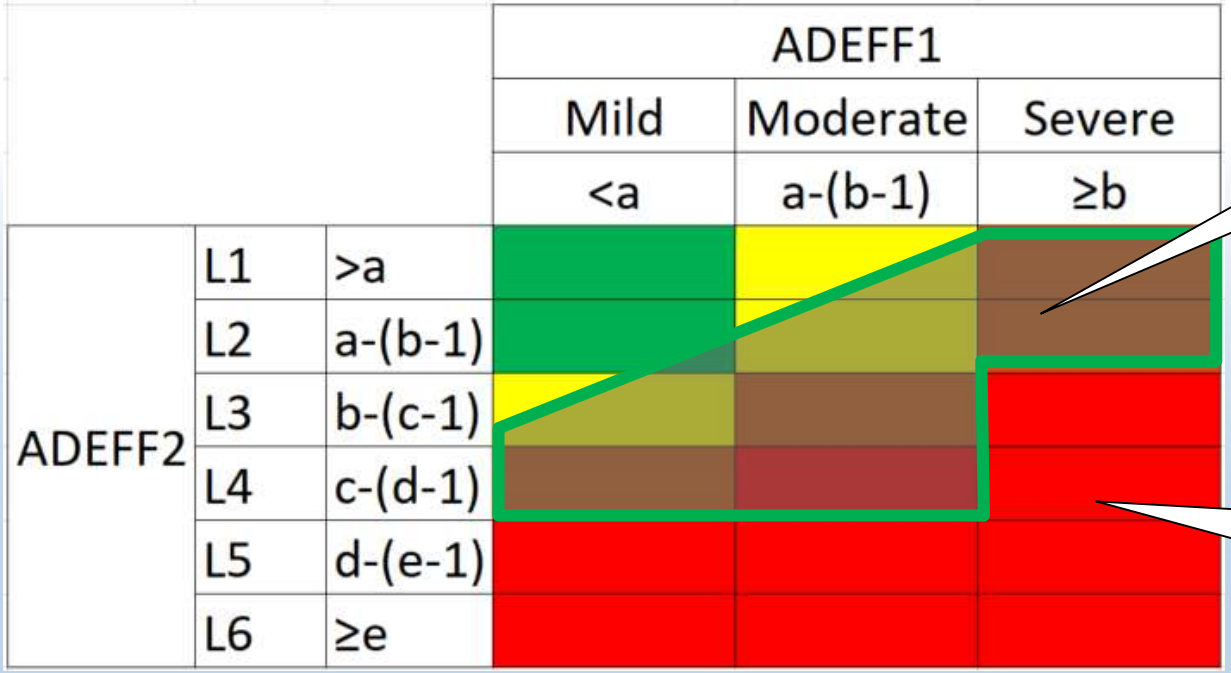
- Answer new questions
  i. Identify right patient populations that have fast disease progression
  ii. Identify demographic and disease characteristics that related to disease progression/clinical endpoint events
  iii. Explore the correlation between disease progression and occurrence of clinical endpoints
  iv. …

- To pool multiple projects/indications as one analysis database

- Covariates: special factors of study interest (ADEFF1, ADEFF2), co-morbidities and baseline characteristics (e.g., age, gender, BMI)

Disease Prognosis

- Low risk
- Moderate risk
- High risk
- Very high risk



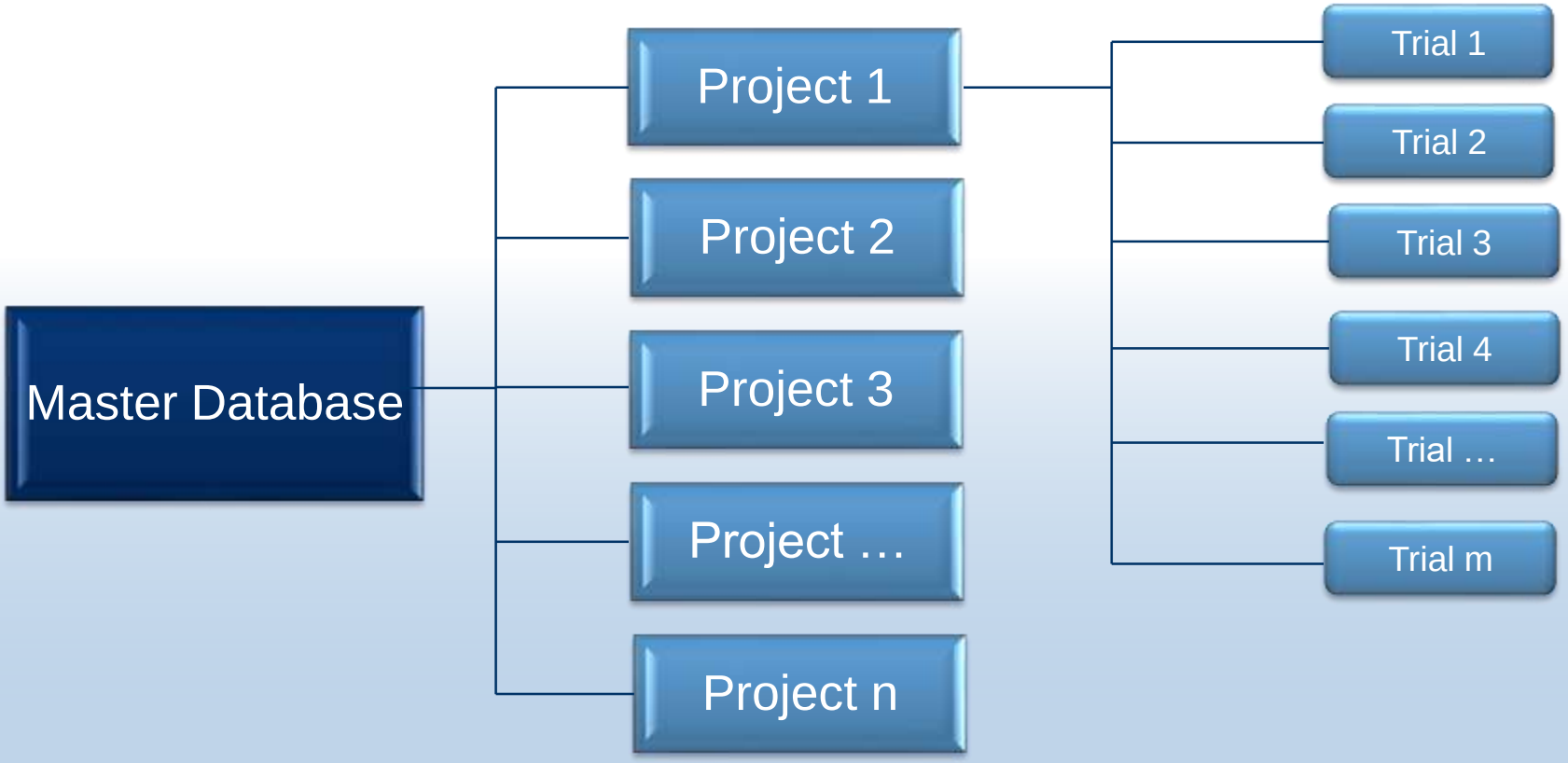
Unknown and under-studied but potentially benefitting population

Traditional study population

Data Processing

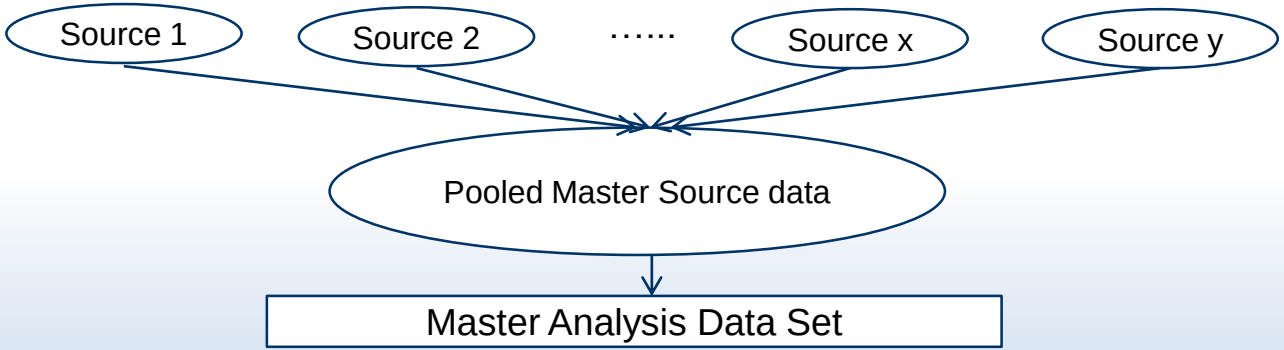
Hong Wang

Pooling database



Different ways to create master project analysis datasets

– To Harmonize at source data level (build master project database)

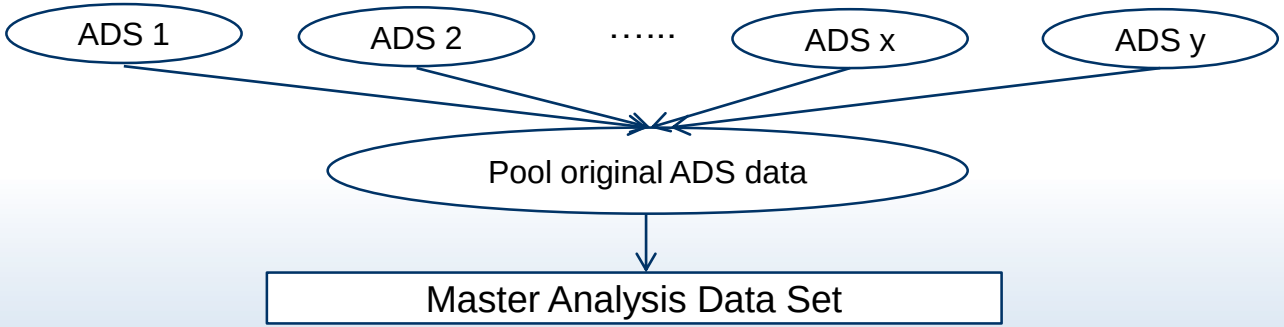


- Pro: first hand information, no black area, can get any kind of analysis dataset
- Con: take more time to harmonize if they are not in SDTM format (for example, in legacy format), or projects/trials included are spanned many years (e.g., over 20 years); some data issue is hard to know/take care

Suggestion: same format/type of source data

Different ways to create master project analysis datasets

– Harmonize at original Analysis Data Set (ADS) level

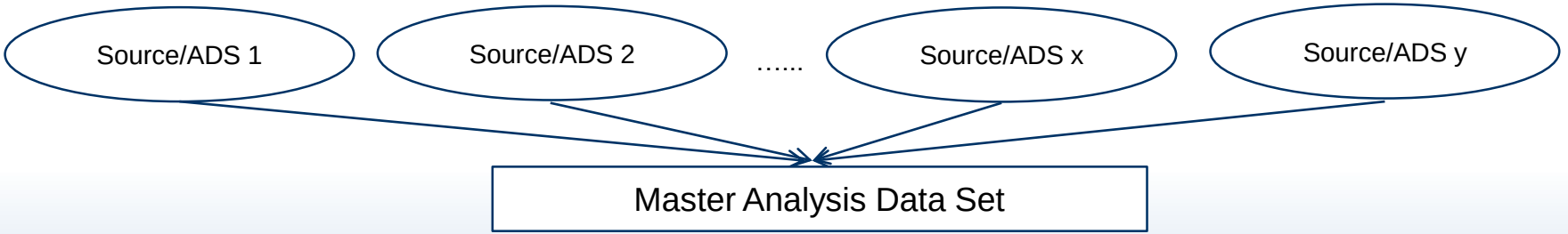


- Pro: easy to use, a lot of info has been taken care, make programing easier
- Con: no complete information for analysis purpose, especially not in ADaM (no traceability)

Suggestion: All info needs could be derived from original ADS, same derivation rule (e.g. All imputed using LOCF)

Different ways to create master project analysis datasets

– Part from ADS, part from source data, harmonize at master project ADS level



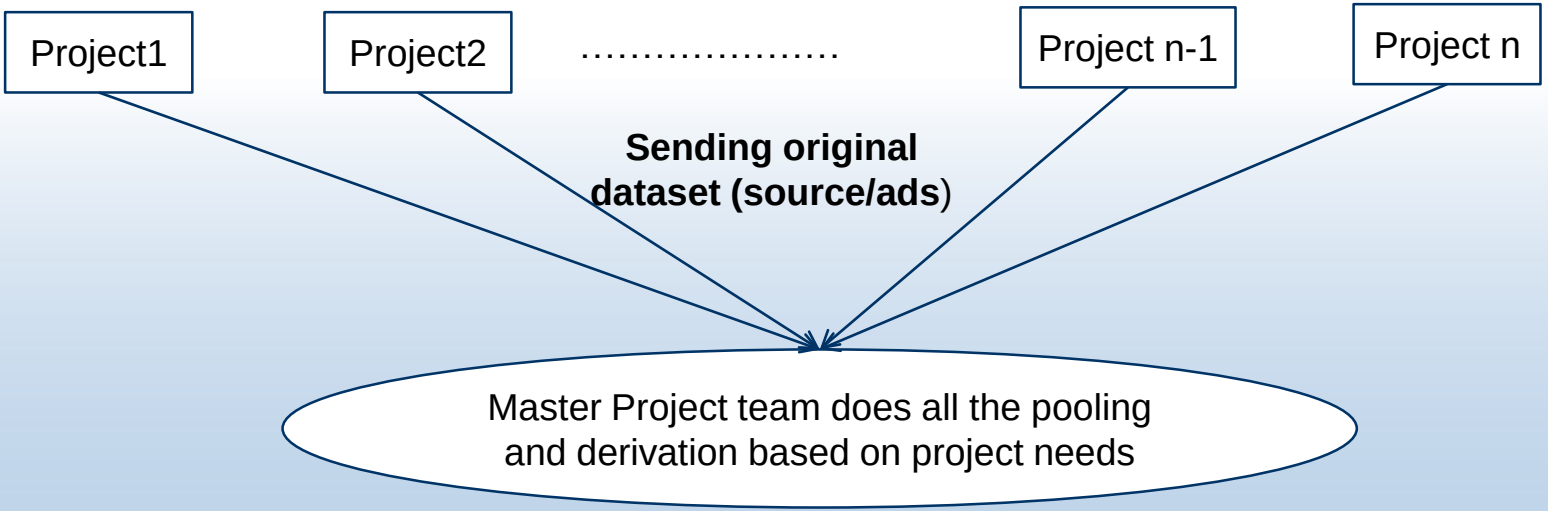
- Use original projects/trials ADS datasets as much as possible—some messy data has been taken care (e.g. ADSL)
- If can not get from original projects/trials ADS, then go to raw (as collected) data to derive (e.g. endpoint using different derivation rule)

Suggestion: lots trials/projects in different data structure; same variable name may reflect different concept; some projects have paper based CRF not through EDC;

There is no One-Method-Fits-All solution

Different programming resources to create datasets

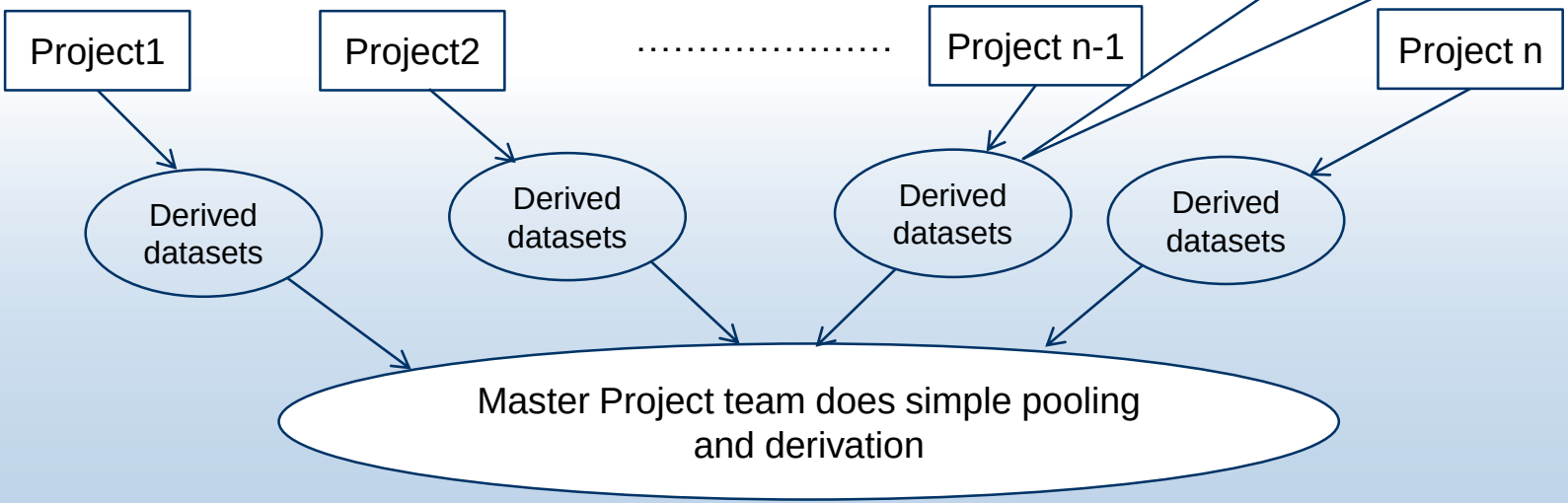
1. Master project team does all the pooling and derivation



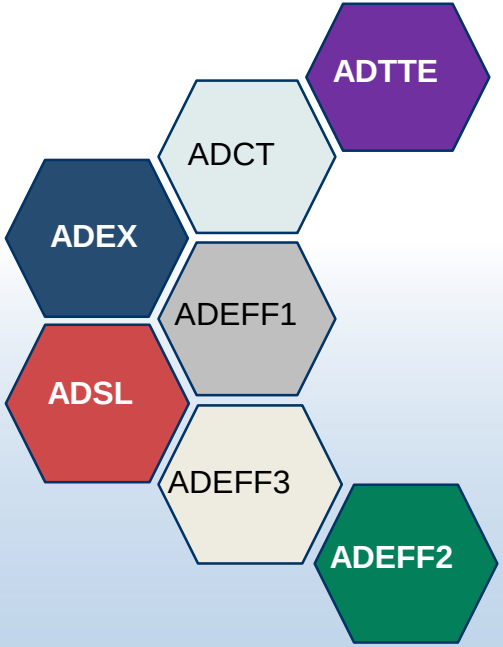
Different programming resources to create datasets:

2. Individual project programmer does derivation

Based on master project needs, each individual project team derives all needed variables using the master specification

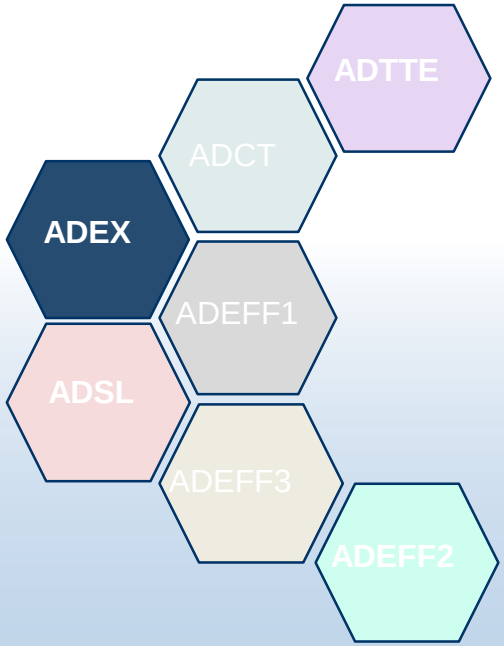


Challenge to Create Analysis Datasets



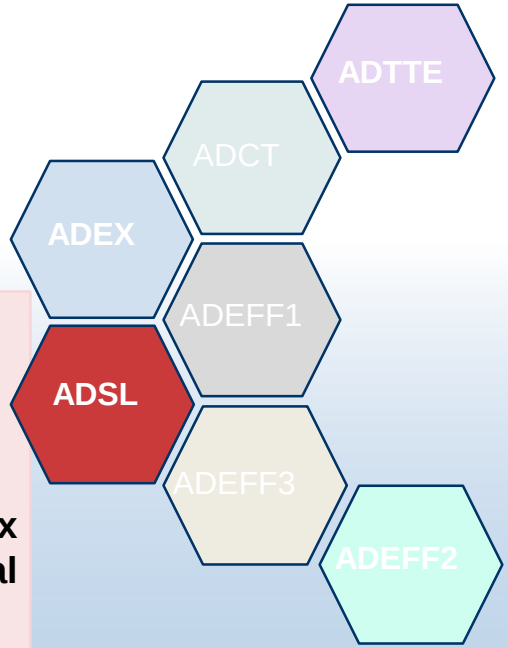
Challenge to Create Analysis Datasets

- Actual treatment period, ignore short wrong medication period;
- Need to specifically take care of cross over, extension trials
- Only include patients needed (eg. right population)

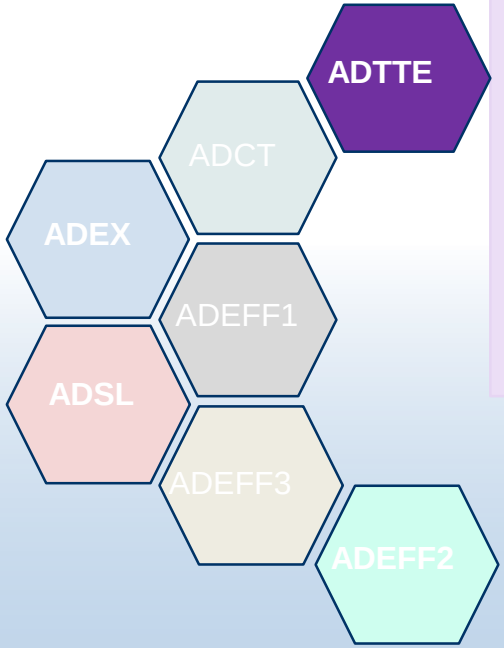


Challenge to Create Analysis Datasets

- Demographic
- Baseline characteristics
- Concomitant medications
- **Co-morbidities**
  Some projects collected tick box instead of MedDRA term, clinical experts mapped diagnosis to current disease category
  MedDRA/SMQ version

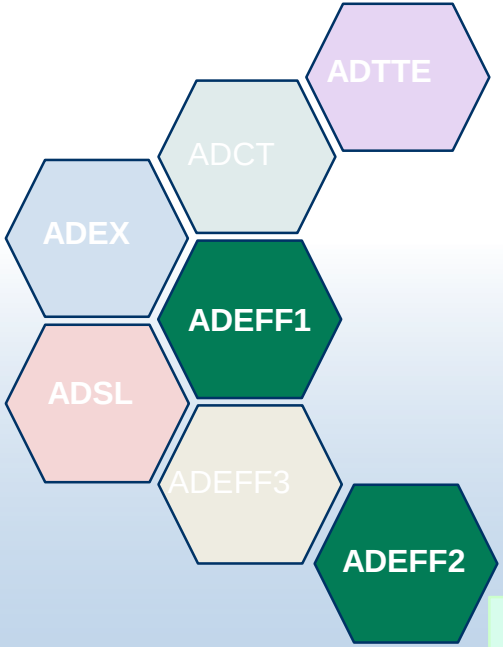


Challenge to Create Analysis Datasets



- Different trials had different way to get event/censor date, challenges:
  - Adjudicated events until primary analysis and then no adjudication any more
  - Extension trial
  - ……

Analysis Dataset



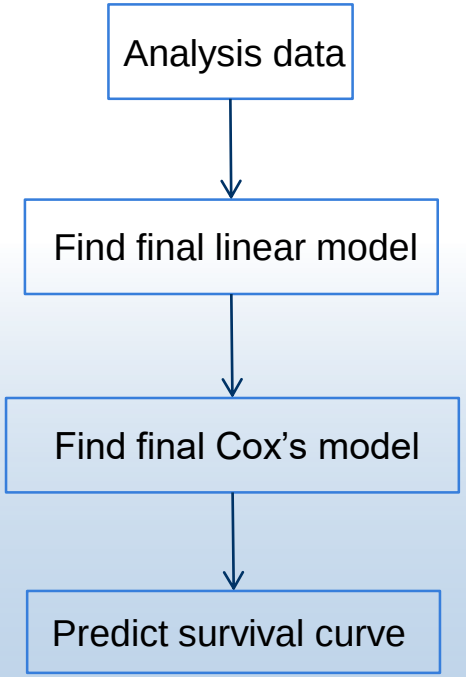
- Different trials collected value in different formula, need to convert them based on the unique formula (e.g. BSA, EGFR).

Innovative programming methods in meta-analysis

Hao Meng

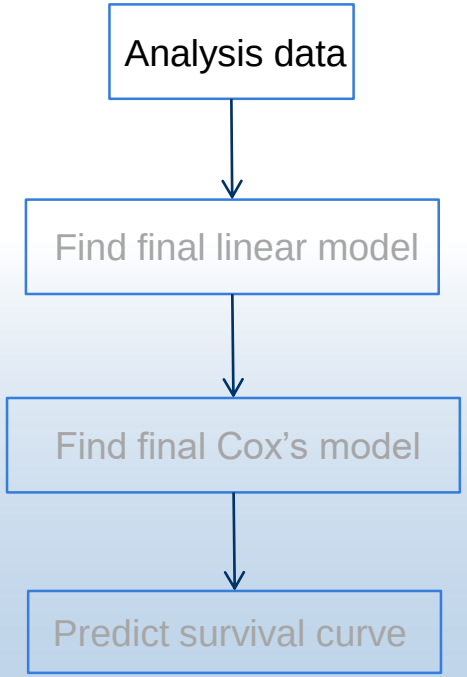
Statistical models and programming functions

- Data preparation and overview
  - ADS creation and cleaning (R haven, HPMIXED)
- Longitudinal Analysis
  - Linear mixed model (LMM/MMRM)
  - Generalized additive mixed model (GAMM)
  - Variable selection (glmmLasso)
  - Latent class mixed model (LCMM)
- Time-to-event Analysis
  - Kaplan-Meier curve
  - Cox model with time-dependent variable



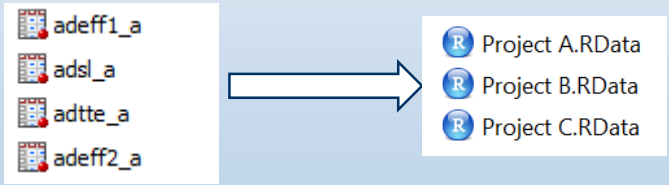
Statistical models and programming functions

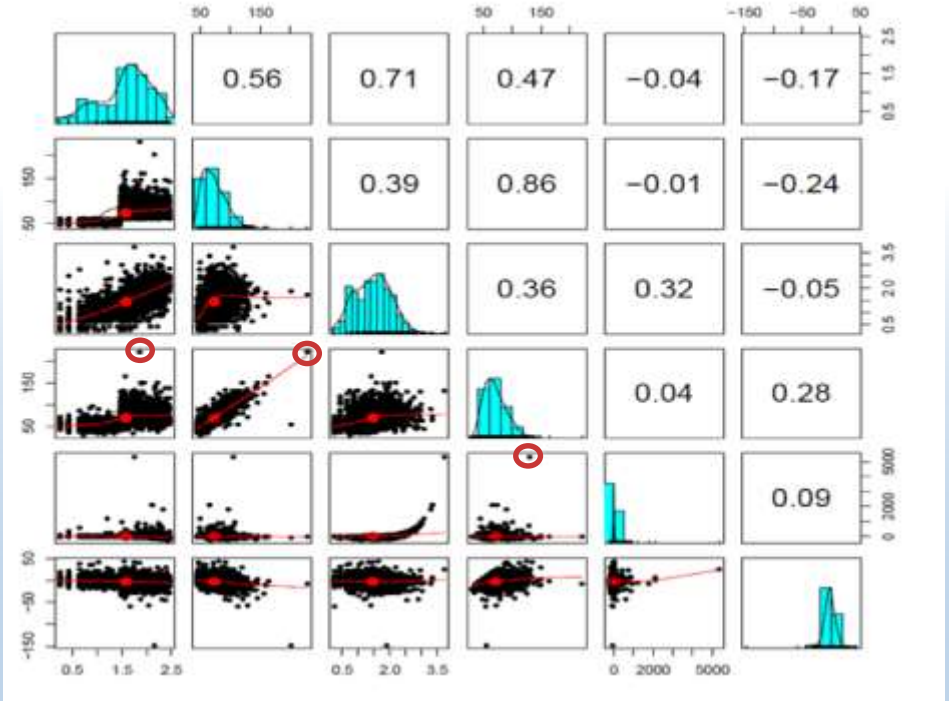
- Data preparation and overview
  - ADS creation and cleaning (R haven, HPMIXED)
- Longitudinal Analysis
  - Linear mixed model (LMM/MMRM)
  - Generalized additive mixed model (GAMM)
  - Variable selection (glmmLasso)
  - Latent class mixed model (LCMM)
- Time-to-event Analysis
  - Kaplan-Meier curve
  - Cox model with time-dependent variable



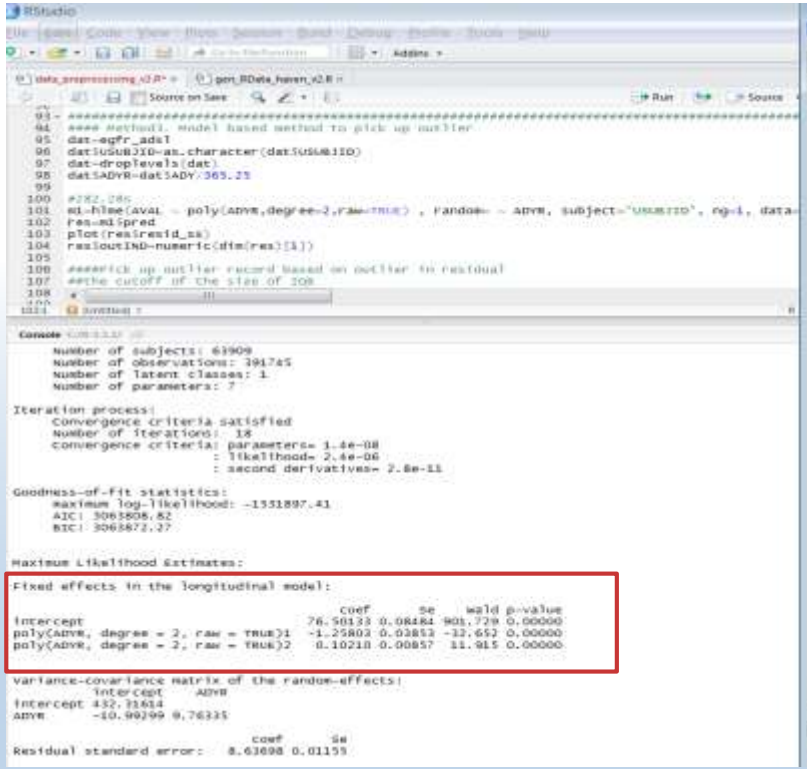
Data preparation and overview

- SAS to R data conversion (haven/foreign)
- Scatter matrix plot to see data trend (corrplot/PerformanceAnalytics)
- Use model based method to remove outlier (PROC HPMIXED vs. hlme)

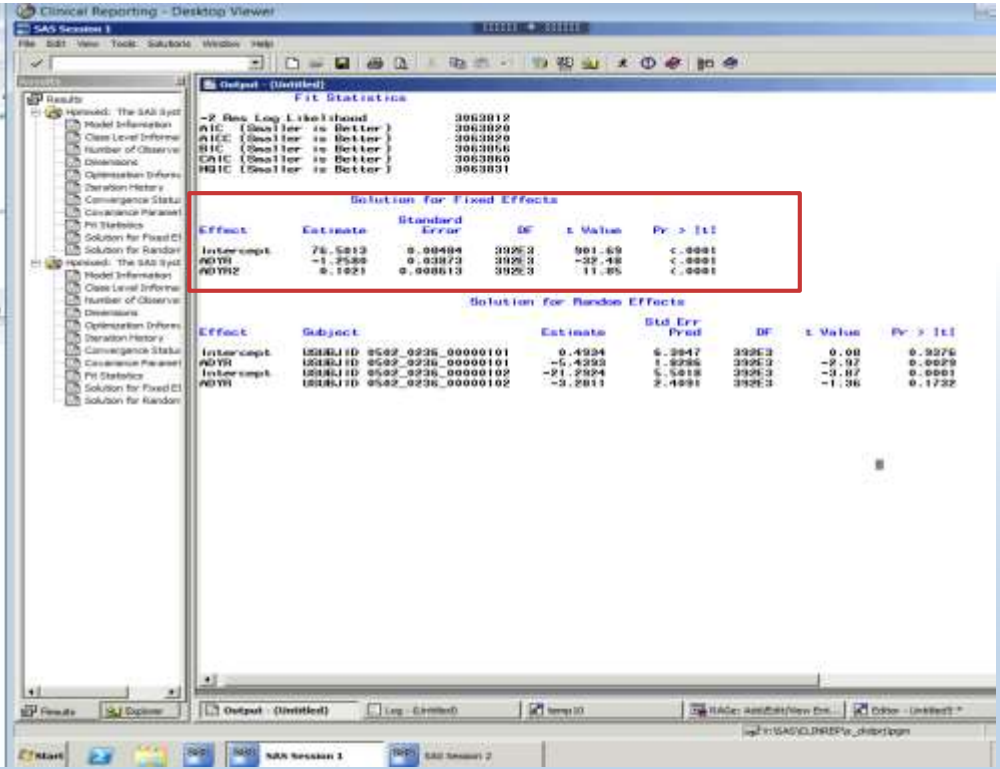




SAS vs. R



```
Number of subjects: 63909
Number of observations: 391745
Number of latent classes: 1
Number of parameters: 7

Iteration process:
Convergence criteria satisfied
Number of iterations: 18
Convergence criteria: parameters= 1.4e-08
                    : likelihood= 2.4e-06
                    : second derivatives= 2.8e-11

Goodness-of-fit statistics:
maximum log-likelihood: -1531897.41
AIC: 3063808.82
BIC: 3063872.27

Maximum Likelihood Estimates:

Fixed effects in the longitudinal model:

                                    coef      se    wald p-value
Intercept                        76.50133 0.08484 901.729 0.00000
poly(ADYR, degree = 2, raw = TRUE)1 -1.25803 0.03853 -32.652 0.00000
poly(ADYR, degree = 2, raw = TRUE)2  0.10210 0.00857  11.915 0.00000

Variance-covariance matrix of the random-effects:
          Intercept      ADYR
Intercept 432.31614
ADYR      -10.90299 0.76335

                     coef     se
Residual standard error: 8.63698 0.01155
```



Fit Statistics

-2 Res Log Likelihood	3063812
AIC (Smaller is Better)	3063820
AICC (Smaller is Better)	3063820
BIC (Smaller is Better)	3063856
CAIC (Smaller is Better)	3063860
HQIC (Smaller is Better)	3063831

Solution for Fixed Effects

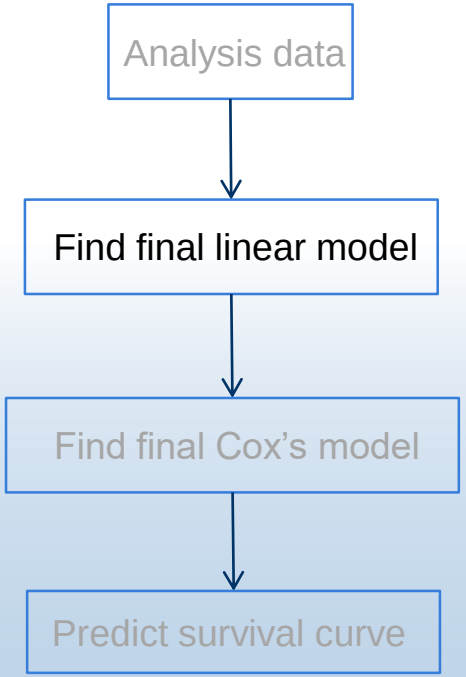
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	76.5013	0.08484	39263	901.69	<.0001
ADYR	-1.2580	0.03873	39263	-32.48	<.0001
ADYR2	0.1021	0.008613	39263	11.85	<.0001

Solution for Random Effects

Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	USUBJID 0502_0236_00000101	0.4934	6.3947	39263	0.08	0.9276
ADYR	USUBJID 0502_0236_00000101	-5.4393	1.9286	39263	-2.82	0.0029
Intercept	USUBJID 0502_0236_00000102	-21.2924	5.5818	39263	-3.82	0.0001
ADYR	USUBJID 0502_0236_00000102	-3.2811	2.4491	39263	-1.36	0.1732

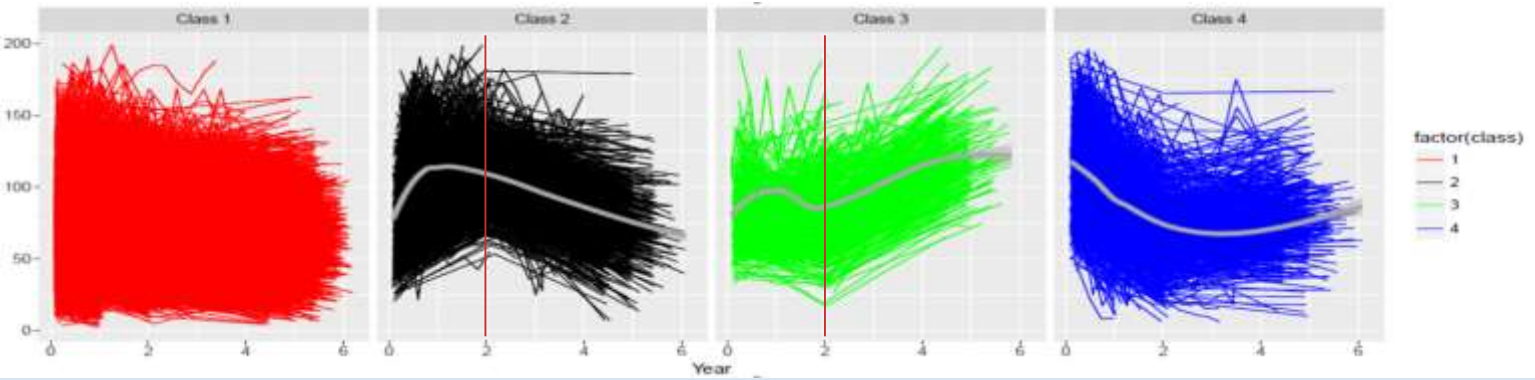
Statistical models and programming functions

- Data preparation and overview
  - ADS creation and cleaning (R haven, HPMIXED)
- Longitudinal Analysis
  - Linear mixed model (LMM/MMRM)
  - Generalized additive mixed model (GAMM)
  - Variable selection (glmmLasso)
  - Latent class mixed model (LCMM)
- Time-to-event Analysis
  - Kaplan-Meier curve
  - Cox model with time-dependent variable



Explore ADEFF1 with LCMM

Heterogeneity due to different study scenario, study design and duration of the trial etc.
To find subgroup patients with the same time trend of ADEFF1 or ADEFF2.

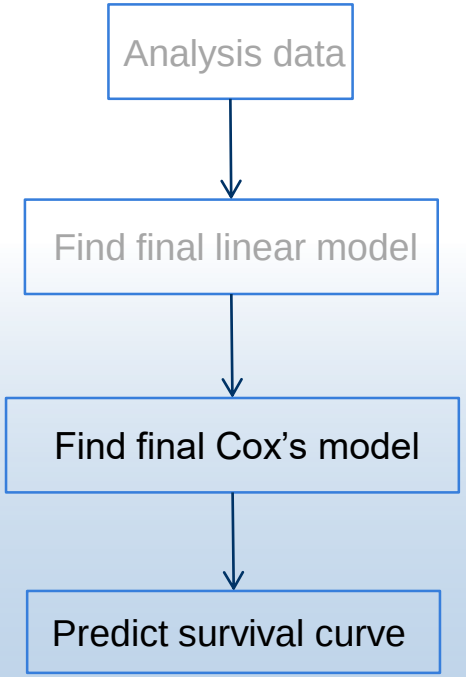


N	21527	588	714	283
Age (year)	69	57	62	65
Baseline ADEFF1	73	98	87	83
Baseline ADEFF2	5	5	4	6
….				

Identify demographic and disease characteristics that related to disease progression/clinical endpoint events

Statistical models and programming functions

- Data preparation and overview
  - ADS creation and cleaning (R haven, HPMIXED)
- Longitudinal Analysis
  - Linear mixed model (LMM/MMRM)
  - Generalized additive mixed model (GAMM)
  - Variable selection (glmmLasso)
  - Latent class mixed model (LCMM)
- Time-to-event Analysis
  - Kaplan-Meier curve
  - Cox model with time-dependent variable



Explore Time Dependent Variable in Cox's model

Challenge

- To include time dependent covariates in Cox's model, it requires analysis data structure to reflect the value changes along the time (e.g., creatinine at different visits)

subject	time1	time2	status	age	creatinine	...
1	0	15	0	25	1.3	
1	15	46	0	25	1.5	
1	46	73	0	25	1.4	
1	73	100	1	25	1.6	

'survival' package in R:

```
survsummary<-coxph(Surv(time1, time2, status) ~ age + creatinine + steroids +
                     + cluster(subject), mydata)
```

Explore Time Dependent Variable in Cox's model

Before grouping to intervals

	START	END	STUDYID	USUBJID	ARM	PARAMN	AVAL	CNSR	AGE	SEX	BMI	ADY	ADEFF1	LOG_ADEFF1
1	1	2	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
2	2	3	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
3	3	4	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
4	4	5	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
5	5	6	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
6	6	7	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
…														
	START	END	STUDYID	USUBJID	ARM	PARAMN	AVAL	CNSR	AGE	SEX	BMI	ADY	ADEFF1	LOG_ADEFF1
362	362	363	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
363	363	364	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
364	364	365	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
365	365	366	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
366	366	367	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
367	367	368	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
368	368	369	XXXX	XXXX_XX_XXXX	Arm X	1	369	0	68	1	36.00	161	7933.333	0.2247257486

After grouping

	START	END	STUDYID	USUBJID	ARM	PARAMN	AVAL	CNSR	AGE	SEX	BMI	ADY	ADEFF1	LOG_ADEFF1
1	1	161	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	1	4728.571	0
2	161	368	XXXX	XXXX_XX_XXXX	Arm X	1	369	1	68	1	36.00	161	7933.333	0.2247257486
3	368	369	XXXX	XXXX_XX_XXXX	Arm X	1	369	0	68	1	36.00	161	7933.333	0.2247257486

Explore Time Dependent Variable in Cox's model

Before grouping to intervals

Time to event summary for End stage renal disease (ESRD)

Statistics

Number of subjects(censor/event)	63911
Time to censor/event in days	
Minimum	1
25% Quantile	398
Median	1291
75% Quantile	1707
Maximum	4008

63911*1291 = **82,509,101 rows**

After grouping

The SAS System

The CONTENTS Procedure

Data Set Name	WORK.DAT_USE_50	Observations	329672
Member Type	DATA	Variables	41
Engine	V9	Indexes	0
Created	05/18/2018 04:08:20	Observation Length	448
Last Modified	05/18/2018 04:08:20	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Data size: 99.6% reduced
Computation capacity: from can't to can based on the same computing memory level

Explore Time Dependent Variable in Cox's model

Result

Before grouping to intervals

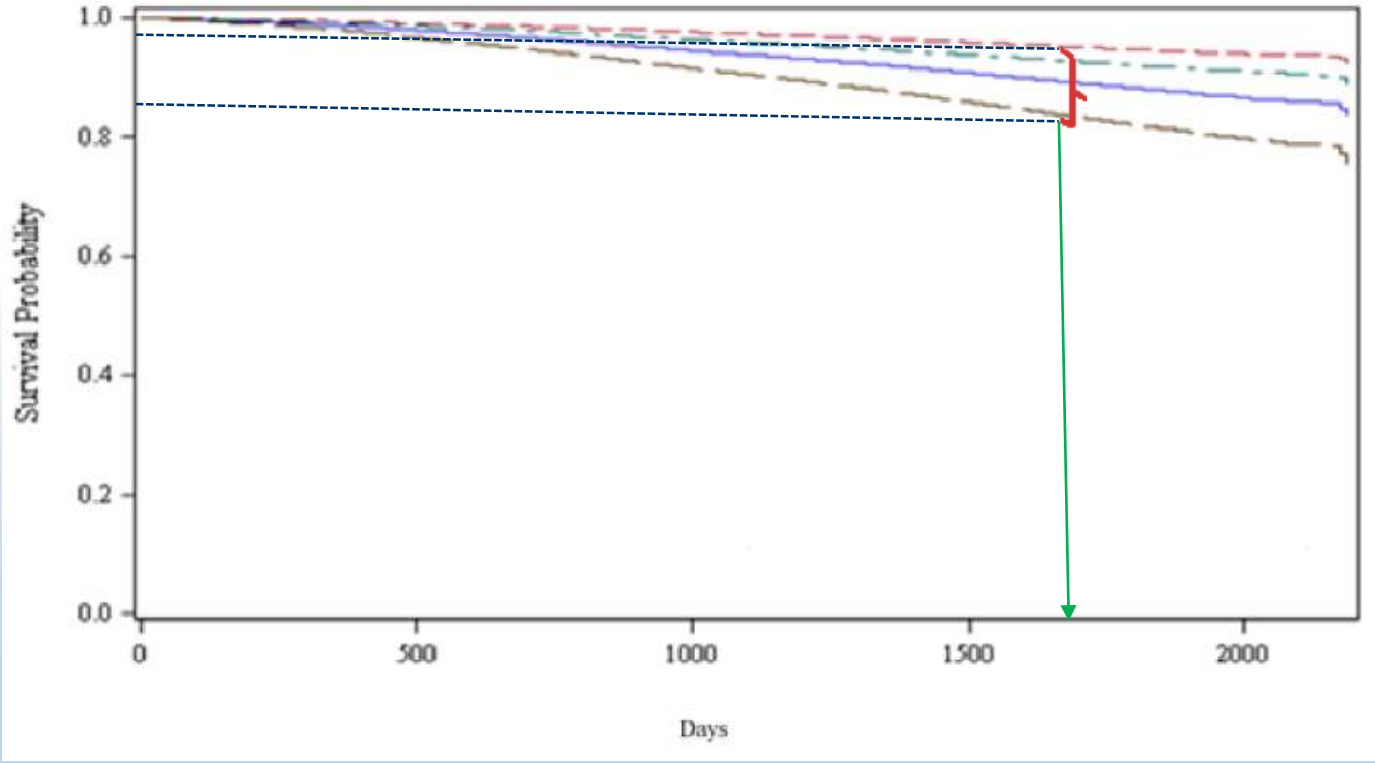
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
LOG_ADEFF1		1	<.0001	0.579	0.491	0.682
AGE		1	<.0001	1.015	1.009	1.020
SEX	2	1	<.0001	1.511	1.373	1.662
BMI		1	0.0002	1.017	1.008	1.025
BADEFF2		1	<.0001	1.027	1.025	1.028

After grouping

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
LOG_ADEFF1		1	<.0001	0.579	0.492	0.683
AGE		1	<.0001	1.015	1.009	1.020
SEX	2	1	<.0001	1.513	1.375	1.665
BMI		1	0.0002	1.016	1.008	1.025
BADEFF2		1	<.0001	1.027	1.025	1.028

Explore the correlation between disease progression and occurrence of clinical endpoints

Adjusted survival for all cause death



- Low risk
- Moderate risk
- High risk
- Very high risk

- Cox model includes selected covariates

Stay Compliant

While bridging data to solution…

https://www.pinterest.com/pin/64457838386560598/

https://freebeacon.com/issues/several-killed-foot-bridge-collapses-florida-university/

Stay Compliant

- Evaluate the objectives of the secondary use of data
- Grant access to limited team members only
- Risk assessment on extended data usage
- Consult with legal and compliance department
- Stay in line with the laws

Questions?