Evidentiary Considerations for Integration of Biomarkers in Drug Development

Statistical Considerations for Clinical Safety Biomarkers

Robin Mogg
rmogg@its.jnj.com
Scientific Director, Statistical Modeling
Janssen Research and Development

August 21, 2015

Potential biomarker panel for drug-induced pancreatic injury: Hypothetical example COU 1

Potential biomarkers:

~~1. MiR-216a~~

~~2. MiR-375~~

3. Protein RA1609

4. Protein RT2864

~~5. Trypsinogen-1~~

~~6. Trypsinogen-2~~

7. Trypsinogen-3

Context of Use (COU 1):

Claim: Qualified biomarkers to be used together with conventional biomarkers, in early clinical drug development (in HV) to **support conclusions as to whether a drug is likely or unlikely to have caused a mild injury response in the pancreas at the tested dose and duration**.

Research use: To make decisions in real time on individual or dose cohort based on **changes in biomarker concentrations (from baseline)**, complementing the use of standard biomarkers

Supportive studies: Two prospective case/control studies in patients using medications that have potential to cause pancreatic injury:

1. Azathioprine in Crohn's disease patients
2. Mesalazine in ulcerative colitis patients with normal pancreas function

✓ Show greater diagnostic predictivity compared to amylase and lipase with a formal adjudication procedure and a **predefined statistical evaluation**

Hypothetical example for drug-induced pancreatic injury COU 1 (cont.)

- **Learn and confirm approach: ample learning completed at this stage**
  - COU 1 clearly defined (support conclusions related to pancreatic injury response)
  - Objectives of confirmatory studies defined (greater diagnostic predictivity)
  - Biomarker panel chosen (though not clear from COU 1 how panel will be used, e.g., individual biomarkers or combination)
  - Measure of biomarker identified (e.g., dynamic change from baseline instead of single timepoint concentration)
- **Predefined statistical evaluation of two prospective studies**
  - Study results must support defined COU 1

Predefined statistical evaluation: study results must support defined COU 1

- **Clear hypotheses regarding how biomarkers are to be considered for use** (relevant null and alternative):
  - E.g., using biomarkers + conventional markers relative to conventional markers alone will **improve the sensitivity (or specificity)** to identify patients treated (not treated) with medications known to potentially cause pancreatic injury
- **Individual analysis to support each hypothesis**
  - Lower bound 95% CI on difference > 0 (is 0 good enough?)
- **But, how to identify patients as having potential injury response?**
  - Signal in **any 1** biomarker, signal in **2 of 3**, signal in **ALL**, signal in a measure that combines and reduces 3 biomarker measures into 1 **composite measure**?
  - And, what is a "signal"? Predictive of injury? Predictive of exposure? Outside variation of HV? Is there a pseudo or true gold standard?

True gold standard vs “pseudo-gold standard”

- **Gold standard (e.g., histopathology)**
  - May be unavailable, too invasive, too expensive
  - If exists, new biomarker performance can be assessed through standard methods (e.g., ROC analysis) to show “comparability” to gold standard
- **“Pseudo-gold standard” often inadequate (e.g., amylase / lipase in pancreatic injury lack specificity)**
  - Comparing new biomarker using pseudo-gold standard as reference is unlikely to show improvement
  - Using **treatment (exposure) as a reference** possible to show improvement

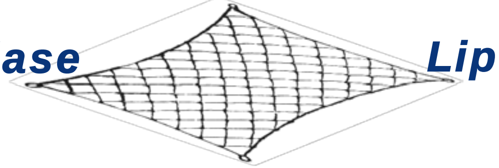
		Conventional markers only		
		Assessed as exposed	Assessed as NOT exposed	Total
Biomarkers+ Conventional markers	Assessed as exposed	A	B	A + B
	Assessed as NOT exposed	C	D	C + D
	Total	A + C	B + D	# controls

Specificity of conventional markers can be compared to that of biomarkers+ conventional markers to show improvement (e.g., 95% CI LB > 0)

What is the risk if the biomarker(s) lack predictive accuracy: Type I vs Type II error

Novel Safety Biomarkers

Gold Standard



Amylase

Lipase

Safety Net

Safety of Individuals in Clinical Trials

Type I error: qualify biomarkers that do not predict toxicity

Type II error: reject biomarkers that do predict toxicity

Which is worse? Depends on intended use and current standard practice

- **Intended use:** to expand testing new drug when conventional biomarkers alone are considered inadequate (i.e., too risky) ⬄ ensure biomarkers predict outcome **(Type I error)**

- **Intended use:** to conclude new drug is unsafe if biomarkers or conventional markers indicate it unsafe when conventional biomarkers alone are considered adequate ⬄ ensure identify potential injury **(Type II error)**

Predefined statistical evaluation: agreement of analytical plan

- Pre-defined statistical analysis plan to address:
  - How to combine data from **multiple studies** (pooling, meta-analysis)
  - How to handle **missing data** (ignore/remove, LOCF, imputation)
  - What are important **sensitivity analyses**?

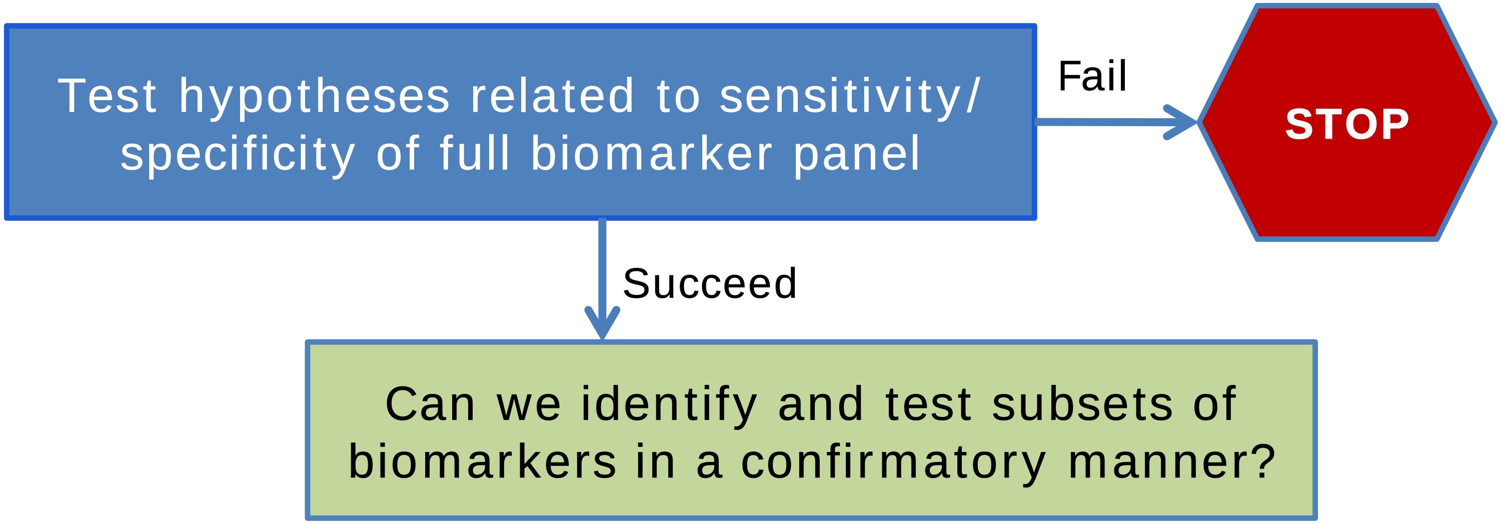
Additional considerations: adaptive strategy to continue learning while confirming?

Interim Analysis	Timing of Interim Analysis	Purpose of Interim Analysis	Example Rule
1 (IA 1)	After completion of ~ first 25% of all study data (first ~25% from each prospective studies)	• Assess initial performance to with respect to sensitivity/ specificity hypotheses • Potential to modify biomarker rules to identify "signal" • Potential to increase sample size	• If observed specificity < 80%, modify biomarker rules. Exclude data from IA 1 in final analysis, increase overall sample size so final analysis is fully powered • If observe specificity ≥ 80% continue to final analysis
2 (perform only if modify rules at IA 1)	After completion of ~ second 25% of all study data (second ~25% from each prospective studies)	• Assess initial performance of modified rules with respect to sensitivity/specificity hypotheses • Potential to stop prospective studies for futility	• If observed specificity < 80%, stop studies for futility • If observe specificity ≥ 80% continue to final analysis

What is impact on Type I /Type II error?

Simulations are useful

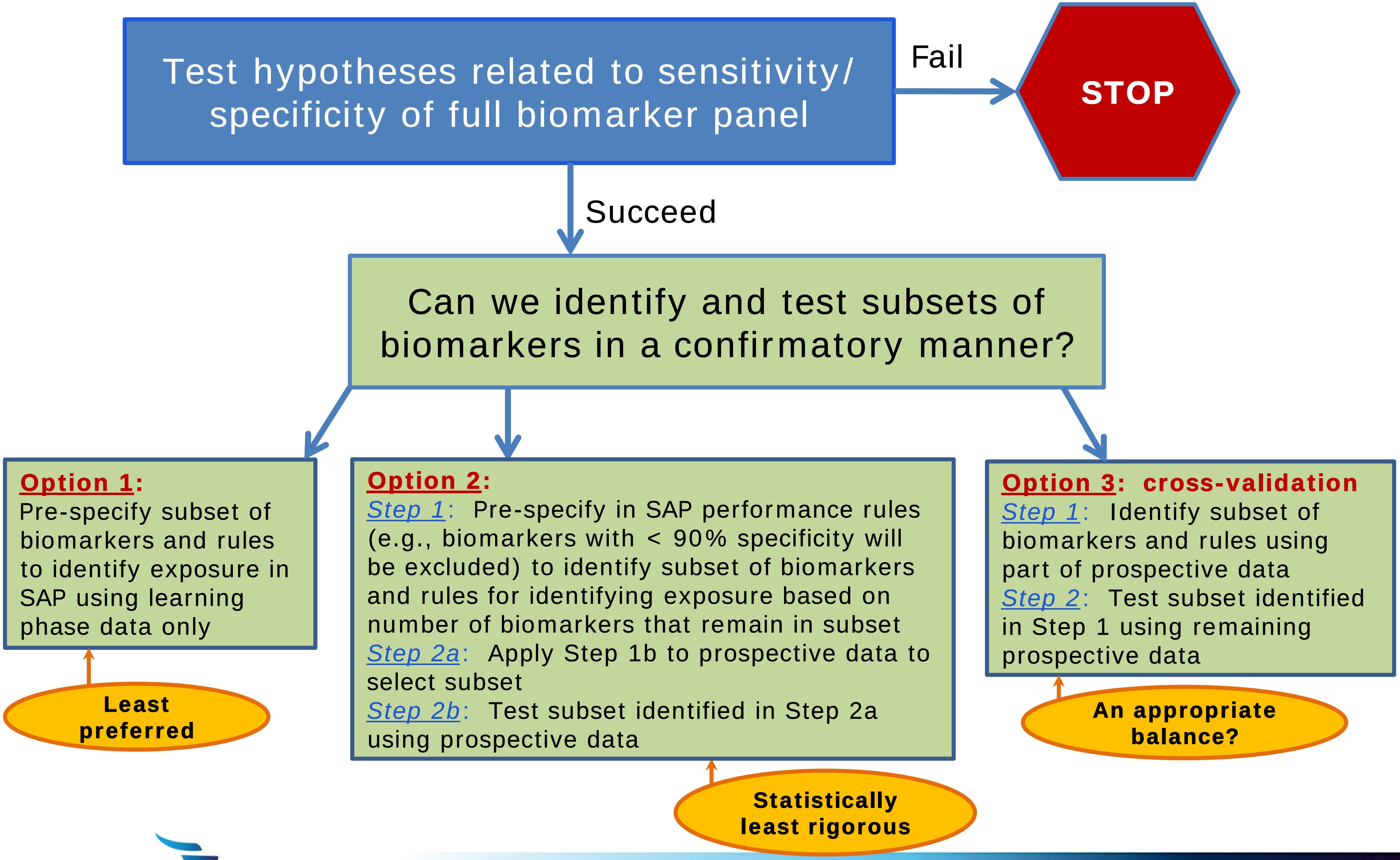
Additional considerations: can we explore biomarker subsets while confirming?



Test hypotheses related to sensitivity/ specificity of full biomarker panel

Can we identify and test subsets of biomarkers in a confirmatory manner?

Additional considerations: can we explore biomarker subsets while confirming?



Test hypotheses related to sensitivity/ specificity of full biomarker panel

Fail

STOP

Succeed

Can we identify and test subsets of biomarkers in a confirmatory manner?

Option 1:
Pre-specify subset of biomarkers and rules to identify exposure in SAP using learning phase data only

Least preferred

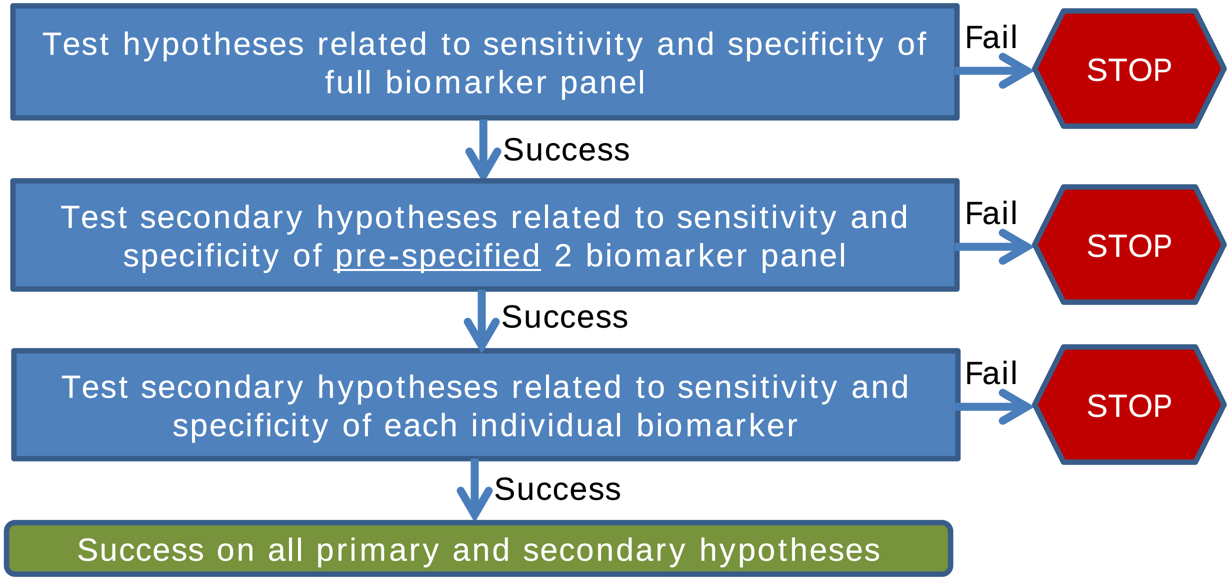
Option 2:
Step 1: Pre-specify in SAP performance rules (e.g., biomarkers with < 90% specificity will be excluded) to identify subset of biomarkers and rules for identifying exposure based on number of biomarkers that remain in subset
Step 2a: Apply Step 1b to prospective data to select subset
Step 2b: Test subset identified in Step 2a using prospective data

Statistically least rigorous

Option 3: cross-validation
Step 1: Identify subset of biomarkers and rules using part of prospective data
Step 2: Test subset identified in Step 1 using remaining prospective data

An appropriate balance?

Additional considerations: Option 1 to explore biomarker subsets



Test hypotheses related to sensitivity and specificity of full biomarker panel

Fail → STOP

Success

Test secondary hypotheses related to sensitivity and specificity of pre-specified 2 biomarker panel

Fail → STOP

Success

Test secondary hypotheses related to sensitivity and specificity of each individual biomarker

Fail → STOP

Success

Success on all primary and secondary hypotheses

A hierarchical testing strategy was proposed to protect the overall Type I error at ≤ 2.5% (1-sided)

- Both sensitivity and specificity tested at each level, success on both must be met to proceed to the next level
- Within final level of the hierarchy, the sensitivity and specificity of the 3 individual BmXs can be tested using appropriate multiplicity adjustment (e.g., Hochberg)

May be difficult to pre-specify and identify subsets when the number of biomarkers in the panel is > 3

Potential biomarker panel for drug-induced pancreatic injury: Hypothetical example COU 2

Potential biomarkers:

1. **Protein RA1609**
2. **Protein RT2864**
3. **Trypsinogen-3**

Context of Use (COU 2):

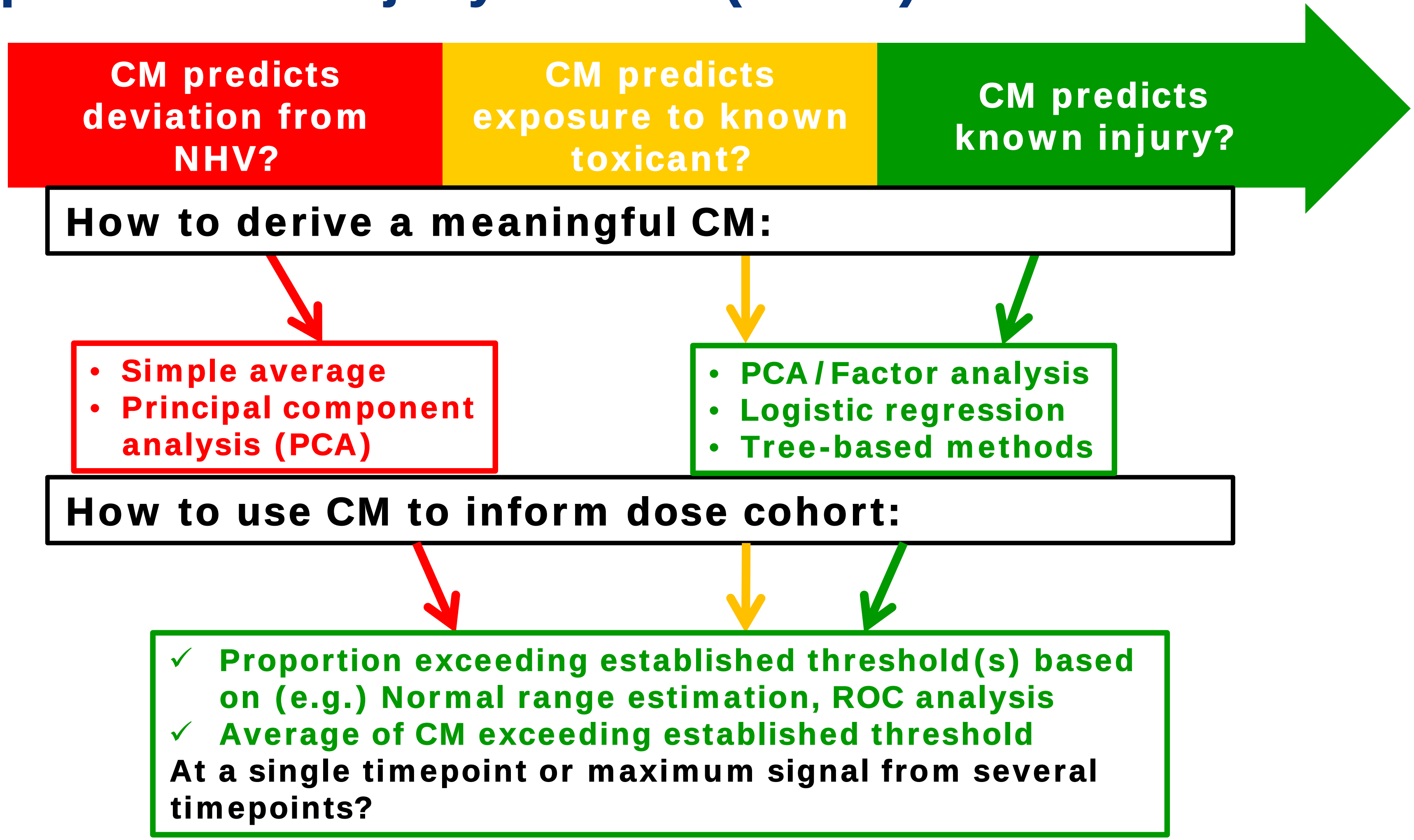
Claim: A composite measure (CM) of the qualified biomarkers to be used together with conventional biomarkers, in normal healthy volunteer trials supporting early clinical drug development

Research use: to make decisions in real time on **dose cohort** using group average of CM, based on **changes in biomarker concentrations (from baseline)**, complementing the use of standard biomarkers

Supportive Data: Learning phase data to support objectives for COU 1 One study in healthy subjects at 2 visits, and one study in patients with known pancreatic injury

✓ Characterize expected variability of CM in NHV and show association of CM with known injury

Hypothetical example for drug-induced pancreatic injury COU 2 (cont.)



CM predicts deviation from NHV?

CM predicts exposure to known toxicant?

CM predicts known injury?

How to derive a meaningful CM:

- Simple average
- Principal component analysis (PCA)

- PCA / Factor analysis
- Logistic regression
- Tree-based methods

How to use CM to inform dose cohort:

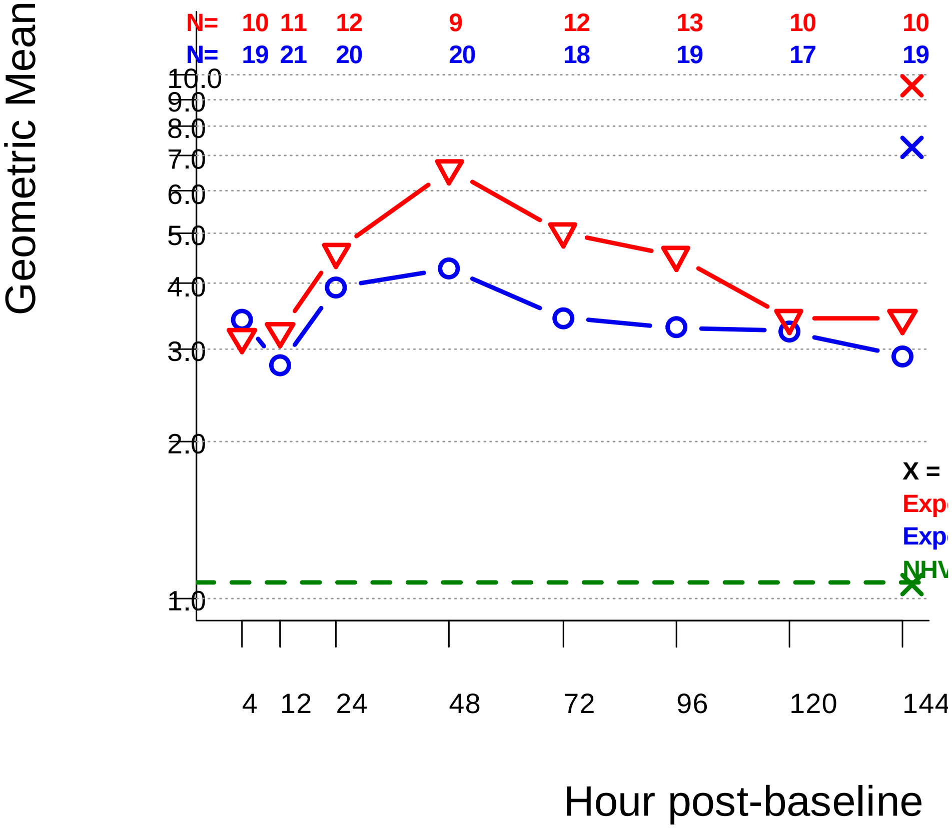
✓ Proportion exceeding established threshold(s) based on (e.g.) Normal range estimation, ROC analysis

✓ Average of CM exceeding established threshold

At a single timepoint or maximum signal from several timepoints?

- What are the limitations of the learning data?

Some potential limitations of learning data

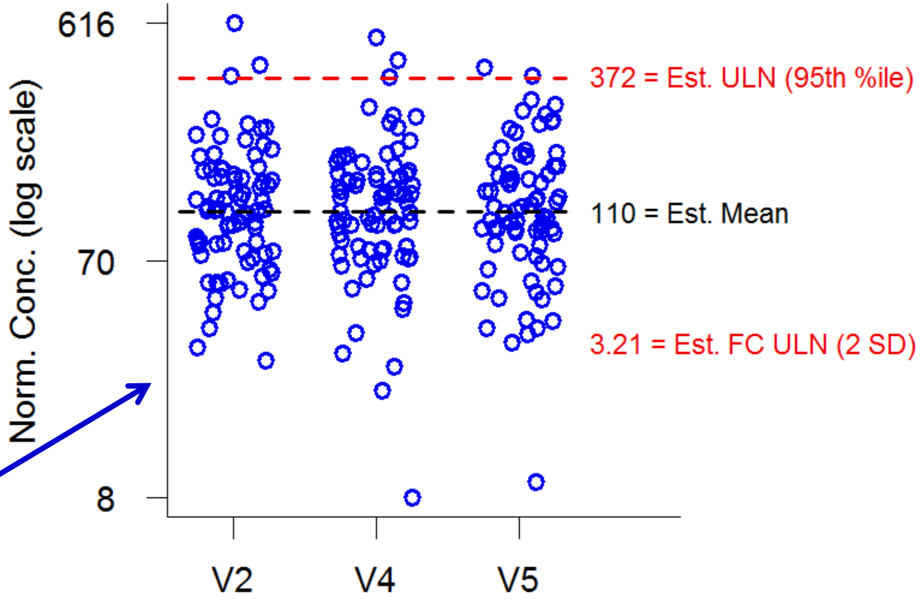


Individual Patient CM = GM of 3 BmX FC from BL
GM CM = GM of Individual Patient CMs

- May only confidently use to predict deviation from NHV
- Multiple timepoints for exposed patients, limited timepoints for NHV
- Signal much larger using maximum across all timepoints
- Association ≠ Causation
- How can we derive thresholds?
  - Bootstrap, but only for single timepoint
  - Modeling and simulation, with assumptions

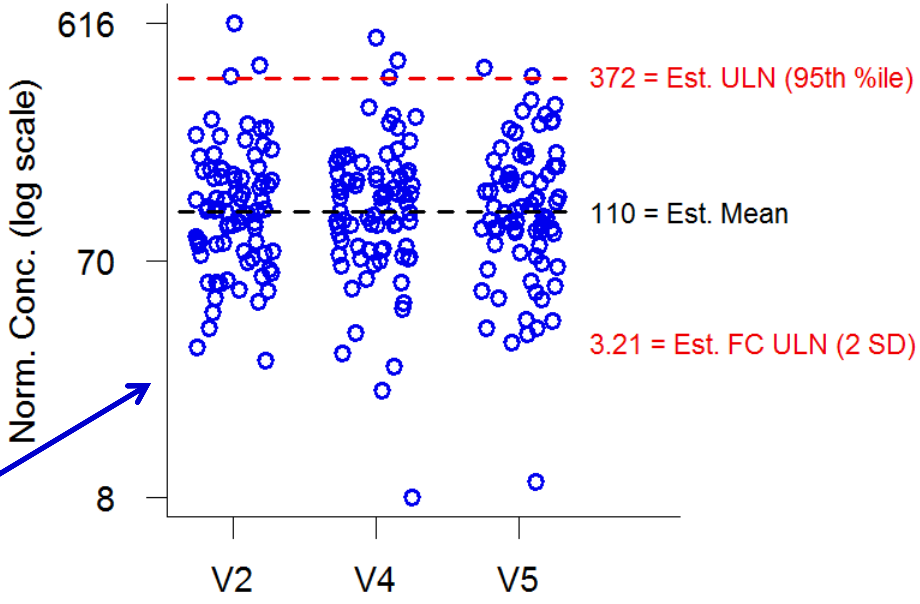
Other relevant statistical considerations **before** COU 1 / COU 2

- What is the right biomarker measure?
  - Raw concentrations, normalized concentrations, change from baseline (absolute or fold-change)
- How to estimate normal ranges (i.e., **in NHV**)?
  - "robust" (Horne and Pesce) method, non-parametric bootstrap, assumptions of normality (can transform)



Other relevant statistical considerations **before** COU 1 / COU 2

- What is the right biomarker measure?
  - Raw concentrations, normalized concentrations, change from baseline (absolute or fold-change)
- How to estimate normal ranges (i.e., **in NHV**)?
  - “robust” (Horne and Pesce) method, non-parametric bootstrap, **assumptions of normality** (can transform)
- **Potential effects of covariates**



Convenient

Can estimate within (σ_W^2) and between (σ_B^2) subject variability

If $\sigma_W^2 << \sigma_B^2$ ⇔ change

If $\sigma_B^2 >> \sigma_W^2$ ⇔ absolute measure

Other relevant statistical considerations before COU 1/COU 2 (cont.)

- **Selection of biomarkers**
  - **Many statistical methods:** regression (traditional, ridge, LASSO), classification/ROC, tree-based methods
    - Multiplicity concerns can be mitigated using false discovery rate methods and cross-validation
  - Selecting a few among potentially many typically goes beyond statistics

Biomarker	Performance in Learning Studies	Biological Interpretation	Assay Availability and Confidence – e.g., LLOQ / Analyte Stablility / No Special Buffer needs	Translatability	Cost
1					
2					
…					

Concluding remarks

- Defining universal evidentiary standards for safety biomarker qualification is difficult
  - Significant diversity in potential context of use
- Appropriate evidentiary standards rely on core statistical principles
  - Some may mimic traditional evidentiary standards associated with drug development (Clear hypotheses, analyses, multiplicity, missing data, ...)
  - Some may not (Settings in safety qualification where Type II error may be important, integrating more than one study for final analysis, ...)
- **Key beyond statistics:** cooperative efforts (consortium), regulatory interactions, patience

Acknowledgements

- Xavier Benain
- Aloka Chakravarty
- Irene Nunes
- John-Michael Sauer
- Matthew Schipper
- Frank Sistare
- PSTC/FNIH Biomarkers Consortium Kidney Safety Project Team Members