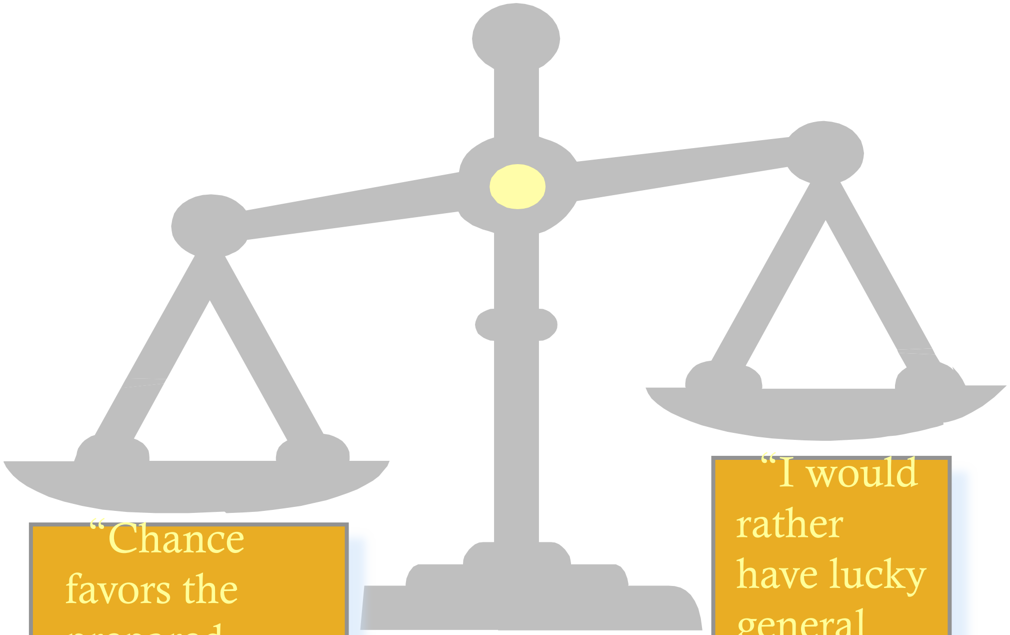
Biomarker Discovery and Validation

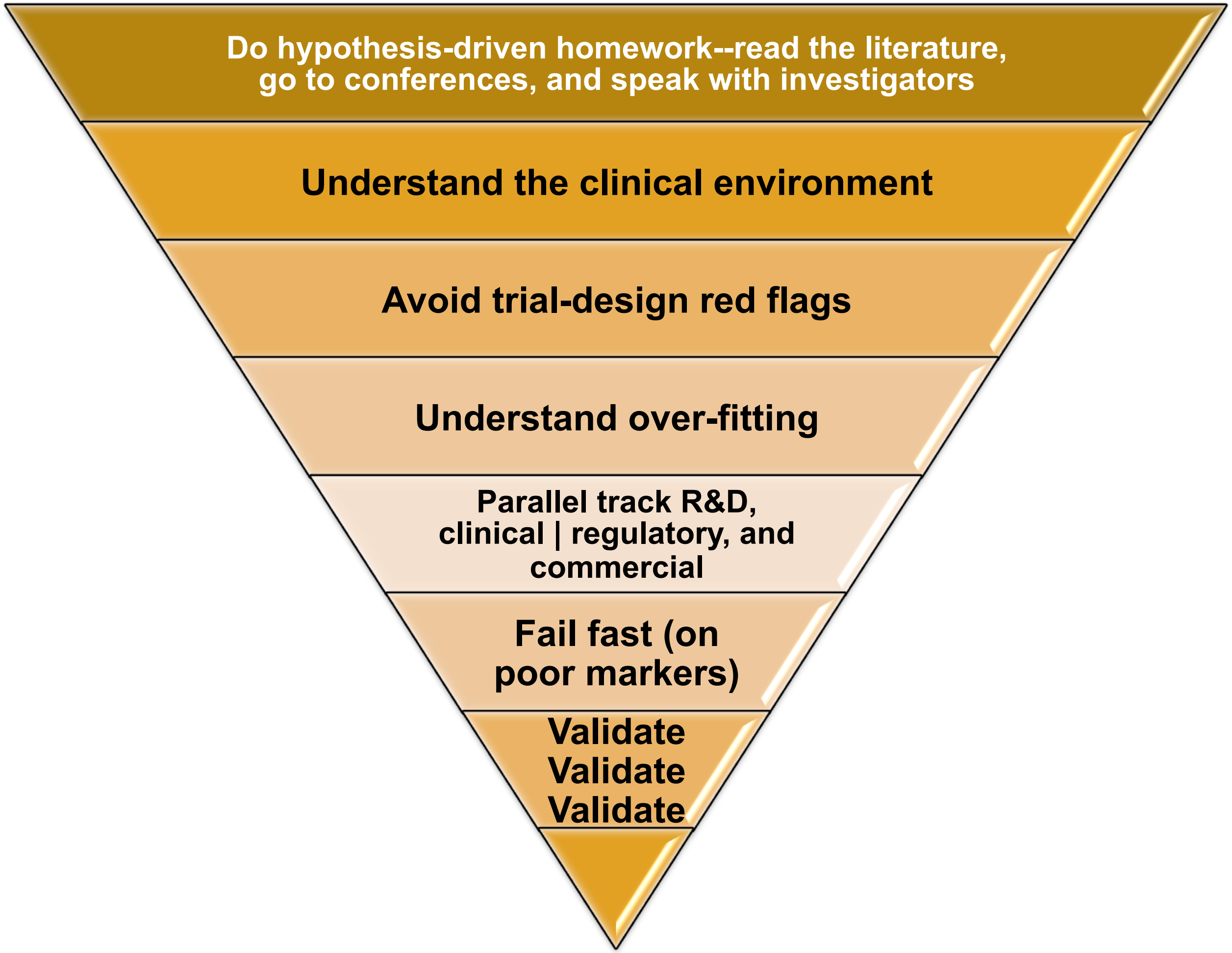
Setting the Stage



"Chance favors the prepared mind."
Louis Pasteur

"The greatest general is he who makes the fewest mistakes."
Napoleon

"I would rather have lucky general than a good one."
Napoleon



Do hypothesis-driven homework--read the literature, go to conferences, and speak with investigators

Understand the clinical environment

Avoid trial-design red flags

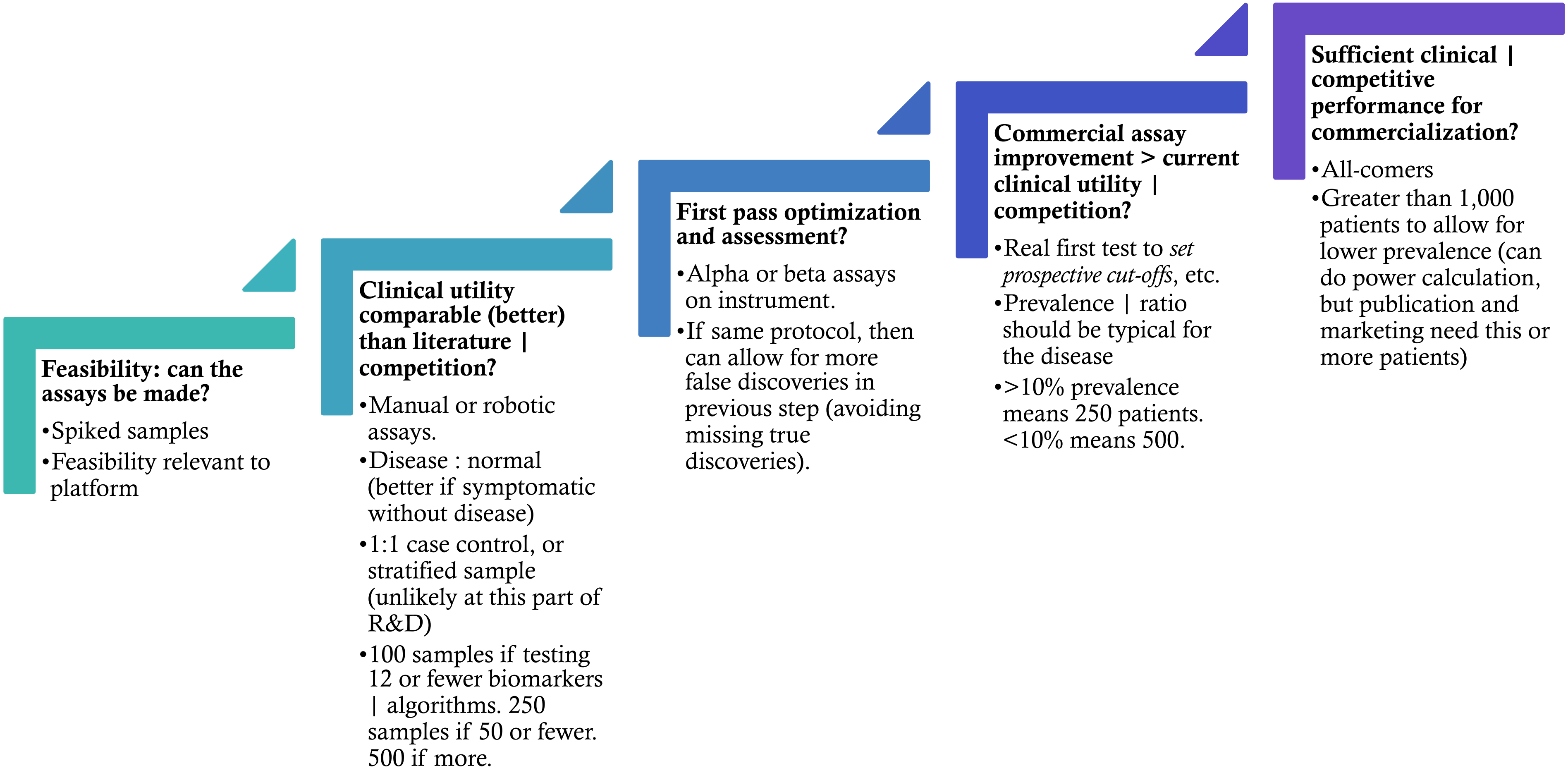
Understand over-fitting

Parallel track R&D, clinical | regulatory, and commercial

Fail fast (on poor markers)

Validate
Validate
Validate

Concurrent assay development and clinical testing



Feasibility: can the assays be made?

- Spiked samples
- Feasibility relevant to platform

Clinical utility comparable (better) than literature | competition?

- Manual or robotic assays.
- Disease : normal (better if symptomatic without disease)
- 1:1 case control, or stratified sample (unlikely at this part of R&D)
- 100 samples if testing 12 or fewer biomarkers | algorithms. 250 samples if 50 or fewer. 500 if more.

First pass optimization and assessment?

- Alpha or beta assays on instrument.
- If same protocol, then can allow for more false discoveries in previous step (avoiding missing true discoveries).

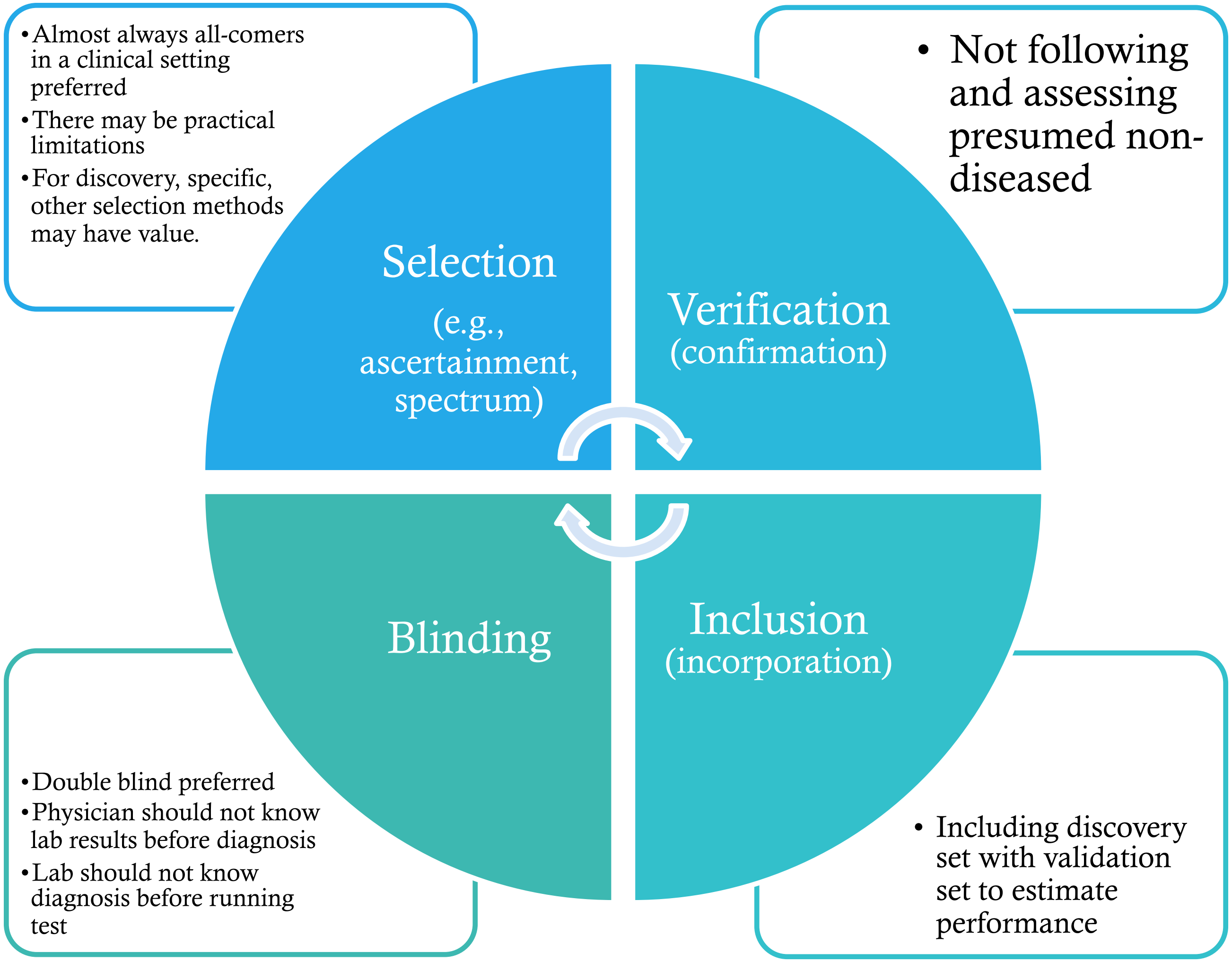
Commercial assay improvement > current clinical utility | competition?

- Real first test to *set prospective cut-offs*, etc.
- Prevalence | ratio should be typical for the disease
- >10% prevalence means 250 patients. <10% means 500.

Sufficient clinical | competitive performance for commercialization?

- All-comers
- Greater than 1,000 patients to allow for lower prevalence (can do power calculation, but publication and marketing need this or more patients)

Big Four Biases



- Almost always all-comers in a clinical setting preferred
- There may be practical limitations
- For discovery, specific, other selection methods may have value.

Selection

(e.g., ascertainment, spectrum)

- Not following and assessing presumed non-diseased

Verification
(confirmation)

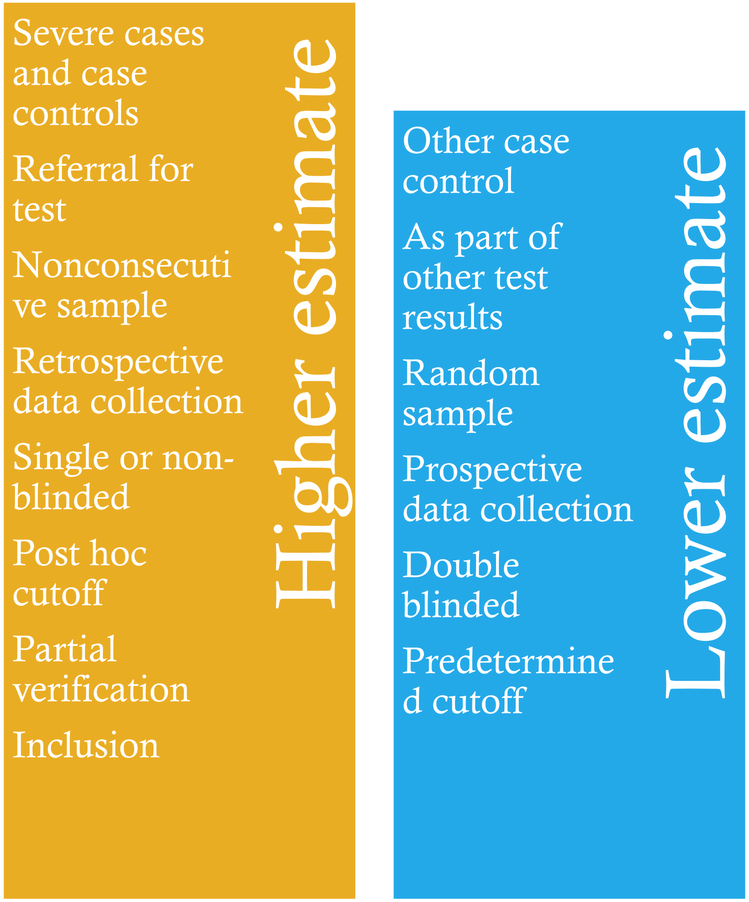
Blinding

- Double blind preferred
- Physician should not know lab results before diagnosis
- Lab should not know diagnosis before running test

Inclusion
(incorporation)

- Including discovery set with validation set to estimate performance

Biased Design Effects on Relative Estimate of Diagnostic Performance



Higher estimate

- Severe cases and case controls
- Referral for test
- Nonconsecutive sample
- Retrospective data collection
- Single or non-blinded
- Post hoc cutoff
- Partial verification
- Inclusion

Lower estimate

- Other case control
- As part of other test results
- Random sample
- Prospective data collection
- Double blinded
- Predetermined cutoff

Source: CMAJ, JAMA

False Discovery Assessments

Bonferroni correction (most conservative)

- Divide the desired p value (probability of true discovery) by the number of biomarkers or algorithms tested.
- This establishes the new p value that any biomarker or algorithm must pass.

False discovery rate

- Similar to Bonferroni for assessment of the biomarker | algorithm with the best p value.
- For subsequent, the desired p value is divided by the number of biomarkers | algorithms remaining to be assessed (i.e., the correction gets easier if some biomarkers | algorithms pass)

Important lessons on proportions

- Don’t use less than 250 patients even when assessing only a few markers
- Start to beware retrospective individual marker discovery at 50 potential markers, in the context above
- For multi-marker indices, the beware starting at 25 potential markers
- When prevalence below 12%, then use more than 1,000 patients
- If using 500 to 1,000 patients with prevalence greater than 12%, relatively good even up to 100 markers.

Further Notes on the Discovery Simulation for Estimates of Samples Sizes

- Degrees of freedom can dramatically affect retrospective biomarker analysis.
  - Simulations run tested whether as the number of markers investigated increases, and either the prevalence, or number of patients decrease, the higher the risk for perceived but random positive results in marker mining.
  - In order to assess the likely outcome of this effect within the realm of marker mining, an experiment was run using random data sets, and varying the quantities of the three variables just listed.
- False AUCs (c-statistics) can be quite high
  - Average experimental AUC for random single markers was 0.62, with the highest a whopping 0.97
  - Average experimental AUC for random multi-marker indices was 0.65, with the highest 1.00
- Sample size (number of patients), prevalence, and number of markers mined are important variables to assess against random results
  - The major danger zone appear to be characterized by patient sizes less than 250 (for essentially all prevalence values, and even if mining only a few markers)
  - Additionally, when mining 25 or more markers, a prevalence below 12% raises concerns, even with patient sizes up to 1,000
  - The converse of this seems to indicate that patient sizes of 500 to 1,000 appear to obviate positive random results even when mining 100 markers as long as the prevalence is greater than 12%

- Blog
  - http://www.wingibbons.wordpress.com
- LinkedIn
  - http://www.linkedin.com/in/wintongibbons/
- SlideShare
  - http://www.slideshare.net/wingibbons
- Twitter
  - @wingibbons