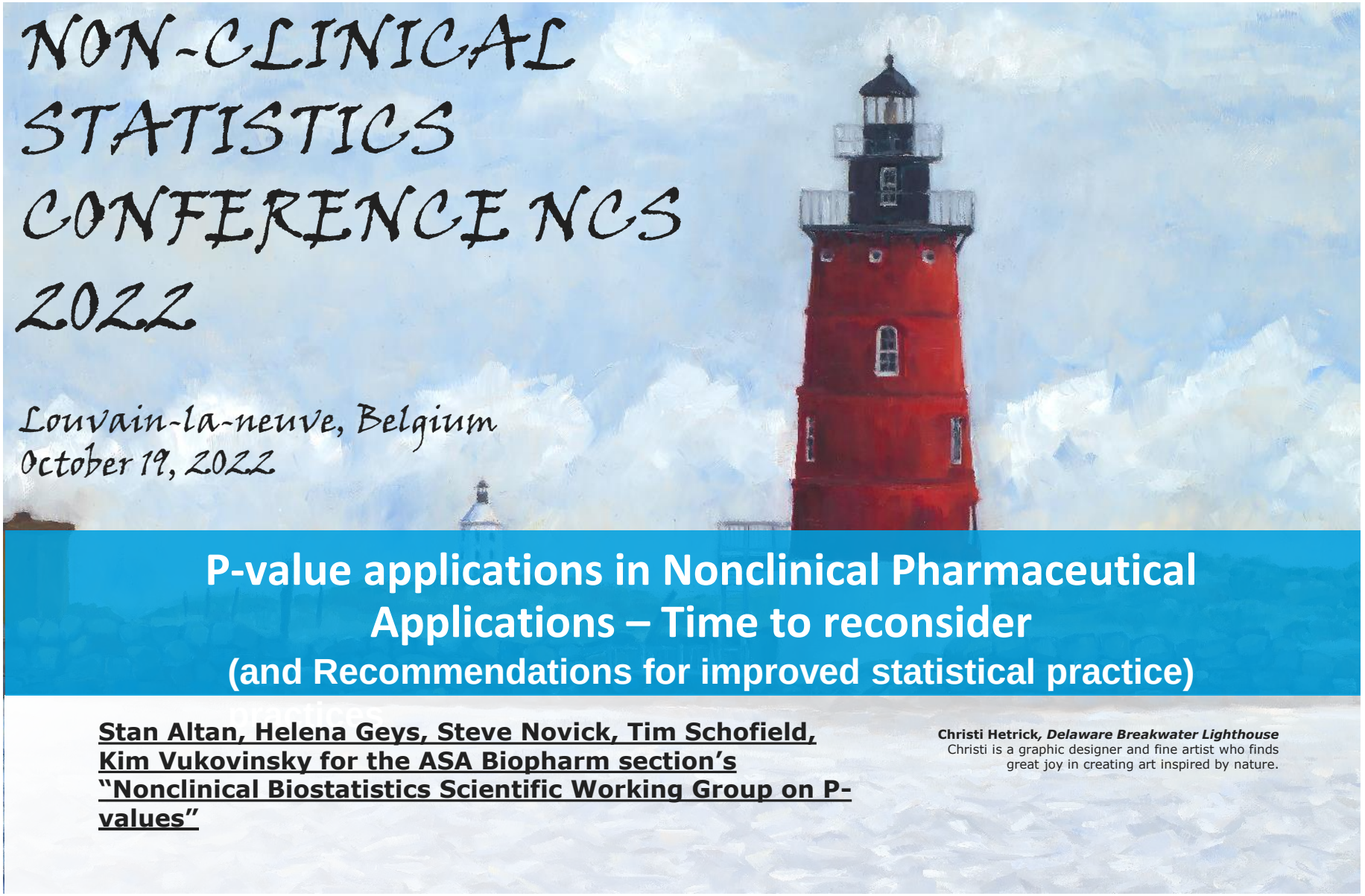


P-value applications in Nonclinical Pharmaceutical Applications – Time to reconsider

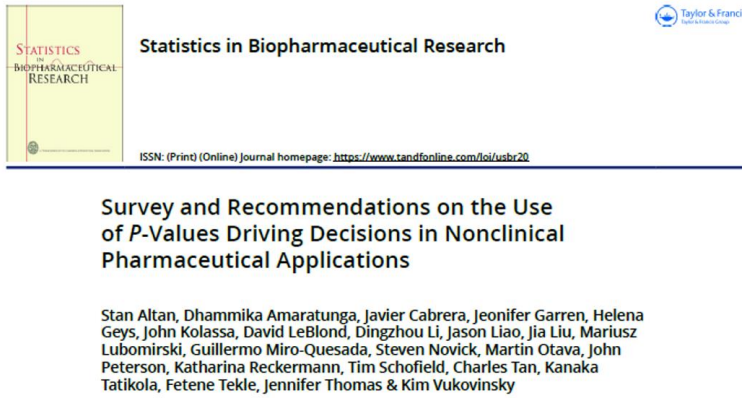
(and Recommendations for improved statistical practice)

Stan Altan, Helena Geys, Steve Novick, Tim Schofield, Kim Vukovinsky for the ASA Biopharm section's "Nonclinical Biostatistics Scientific Working Group on P-values"

Christi Hetrick, *Delaware Breakwater Lighthouse*
Christi is a graphic designer and fine artist who finds great joy in creating art inspired by nature.

Outline

- History – Quick tour
- Overview Nonclinical Statistics in Drug Development
- Survey
  - Discovery/-Omics
  - Preclinical Safety/Toxicity
  - Chemistry. Manufacturing and Controls
- Summary, Final Remarks



Statistics in Biopharmaceutical Research
DOI: 10.1080/19466315.2022.2038258

Published online: 21 Mar 2022

Quick tour of our p-value journey

1900 Karl Pearson

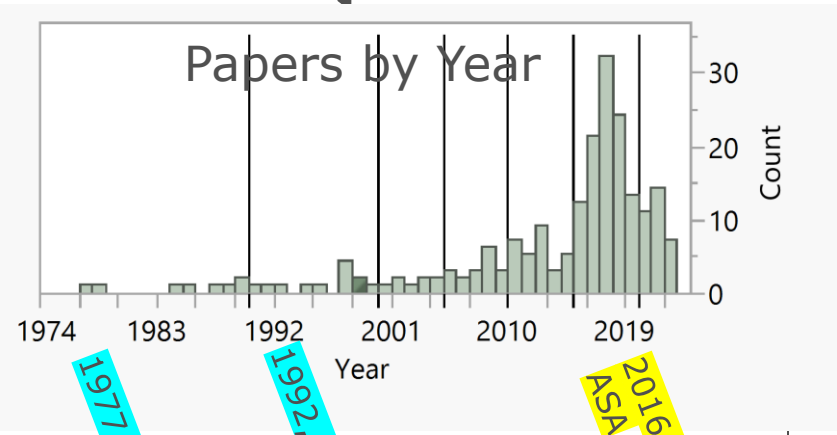
1925 RA Fisher

1961 H. Jeffreys

Wide Acceptance by Experimenters 1925-1950-1995

Stakeholders

A. Mathematical/Foundational
B. Institutional, Scientific Journals
B. Practicioners
   1. Academic researchers
   2. Regulatory Agencies
   3. Industry Researchers



1977 Jekel

1992 Goodman

2016, 2019 ASA

Consensus Convergence ?

Impetus to form an ASA BIOP section Nonclinical Scientific Working Group on P-Values

- ASA position statement on p-values by Wasserstein and Lazar (2016).
- Editorial by Wasserstein, Schirm and Lazar (2019) "Moving to a World Beyond "$p < 0.05$"
- Seen as a call to rethinking the use of the p-value by the larger statistical community, specific domain areas
- Proposal to form a nonclinical scientific working group to align with ASA position

Scientific Working Group Members

- Three focus groups representing the three broad subject areas **of** nonclinical research, development, and lifecycle management in the pharmaceutical industry.
- 22 members representing 8 pharmaceutical companies and one university, collectively expressing the experiences of many decades of industry and academic practice.
- GOALS (Improve the Practice)
  - Survey applications where the p-value is used for decision making
  - Identify areas for improvement
  - Find consensus
  - Harmonize with ASA position

Focus	Name	Affiliation
Discovery/ -Omics	D. Amaratunga	Consultant
	J. Cabrera	Rutgers
	J. Garren	Pfizer
	M. Lubomirski	Amgen
	S. Novick *	AstraZeneca
	C. Tan	Pfizer
Safety/ Tox/ Biomarkers	H. Geys *	Janssen
	J. Kolassa	Rutgers
	D. Li	Pfizer
	K. Tatikola	Janssen
	F. Tekle	Janssen
	J. Thomas	Covance
	S. Altan	Janssen
CMC	D. LeBlond	Consultant
	J. Liao	Merck
	J. Liu	Pfizer
	G. Miro-Quesada	AstraZeneca
	M. Otava	Janssen
	J. Peterson	GSK
	K. Reckermanns	Roche
	T. Schofield *	CMC Sciences
	K. Vukovinsky *	Pfizer

* Focus Group Chair

Nonclinical Statistics in Drug Development

Drug Discovery/ -Omics

Target ID & Screening

- Disease mechanisms
- Compound Identification Go/No-go decision making

Drug Safety Toxicology

Preclinical Studies

- Selection of the “best” compound
- Efficacy
- Safety/Tox

Chemistry, Manufacturing and Control (CMC)

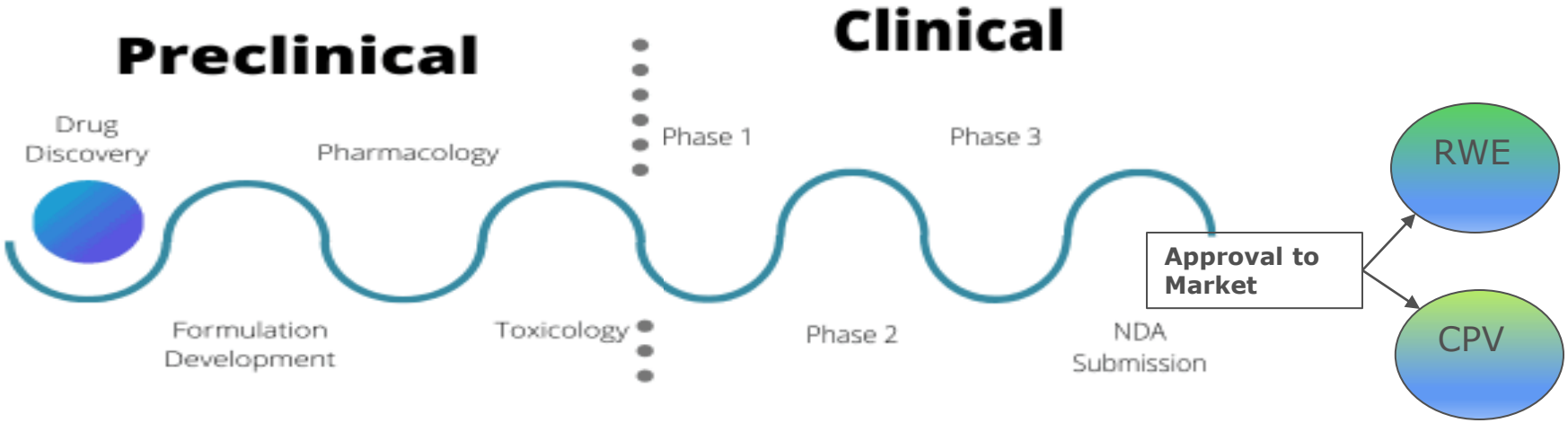
Manufac-turability

- Analytical Tools
- Formulation
- Process Development & Validation

Supply Chain Post-Marketing

Commercial-ization

- Lifecycle Quality Management

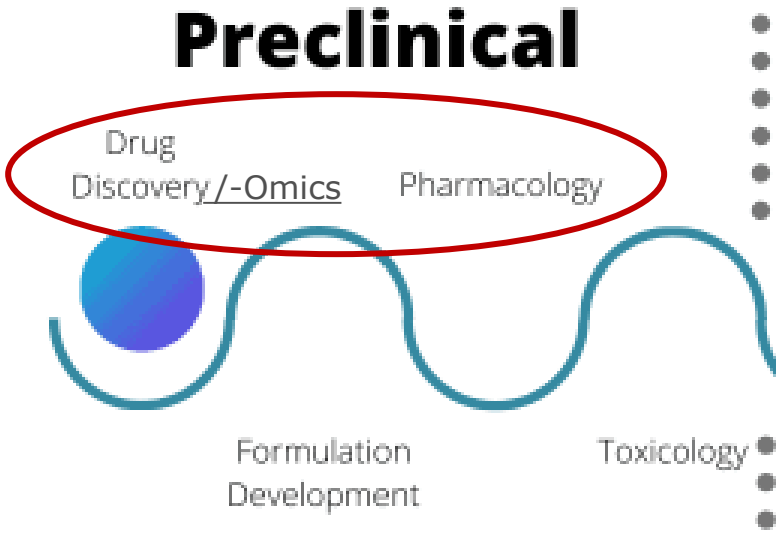


Discovery / -Omics

Viral exacerbation at 40x magnification

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Drug Discovery, -Omics Studies



- Discover compounds exhibiting potentially important therapeutic effects (target disease pathways of interest) with acceptable safety profiles
- 50% late-stage clinical development fail in efficacy, safety, or both (Hwang et al, 2016)
- Minimal regulatory guidances governing the discovery statistical practices

Study Features Impacting on p-value interpretation

- Small, exploratory; powered to detect "gross" differences
- Multiple potentially active entities
- New technologies, requires venturing into unknown areas
- Hypothesis generation distinct from concept validation (PoP)

- Operational intention : Go/no-go decisions are made with high confidence

Examples

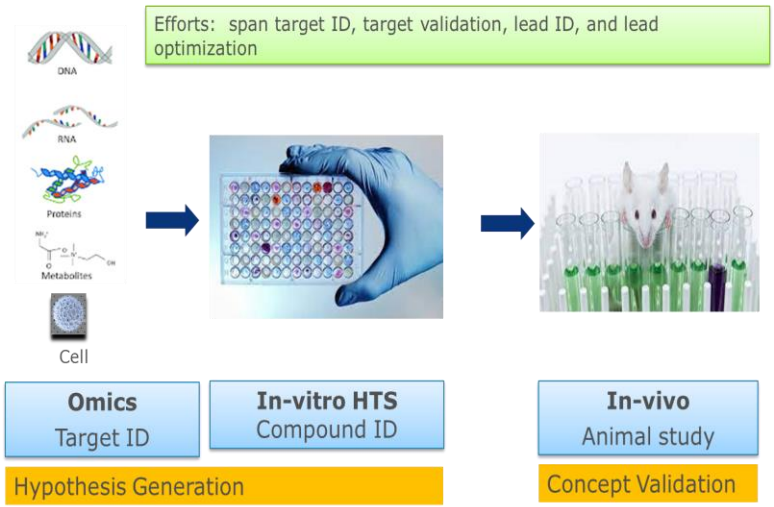
- Difference Testing to identify and validate druggable disease targets.

$$H_0: \mu_{Trt} - \mu_C = 0$$
$$H_1: \mu_{Trt} - \mu_C \neq 0$$

- Mean comparisons via linear (mixed) models,
- log-odds comparisons from large-sample logistic regression modeling,
- Survival-curve comparisons (Kaplan-Meier or Cox proportional hazards modeling).

The Discovery/-Omics Perspective

- P-value alone does not measure effect size or importance
- It's a descriptive statistic or data summary, not conclusive by itself



- P-value should not be blamed for lack of reproducibility
- The p-value is a useful tool for inference, but not sufficient for a go/no-go decision.

P-value usage recommendations

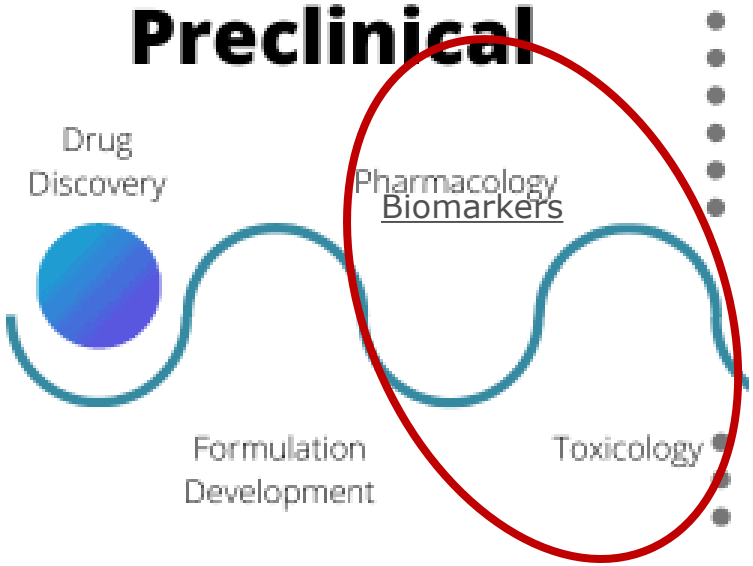
- Consider relevant prior experiments
- Include a confidence or credible interval of the effect size as part of the decision making process
- Bayesian methods can be employed to support data interpretation
- Information Criteria can play a useful role
- Base decisions on consistent signals across multiple experiments combined with statistical, chemical and biological considerations
- Used properly and with totality of evidence, p-value is a reasonable go/no-go decision gate keeper for discovery/omics when combined with scientific judgment.

Preclinical Safety/Toxicology/ Biomarkers

Viral exacerbation at 40x magnification

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Preclinical Safety/Toxicology



- Characterize toxic effects with respect to target organs, dose dependencies, relationship to exposure and potential reversibility
  - Estimate starting dose and dose range for human trials
  - Identify safety parameters for clinical monitoring
- Heavily regulated, numerous guidances

Types of Studies reviewed

Study Type	Statistical issues impacting on p-value
Short term In Vivo Toxicology studies	Small sample sizes (due to ethical considerations) Many endpoints rather than one or two primary endpoints
Long term large-scale Carcinogenicity studies	Low number of endpoint events
Genetic Toxicology studies	Unadjusted pairwise comparisons
Safety Pharmacology studies	Network causality approaches

Significance Testing in Regulatory Toxicology Studies

- Often outsourced to CROs using large IT systems to capture, report and analyse many parameters
- The scale of the IT systems and size of CROs make change difficult as cost and time involved for a (small) improvement is often prohibitive
- **DATA ANALYSIS**
  - Typically analysed in an automated way through decision trees
  - Statistical procedure determined by variability in a sample rather than pre-determined through the data-generating process
  - No adjustment for false positive rates when multiple tests applied

Significance Testing in Regulatory Toxicology Studies

- Common in toxicology studies of all types
- Different styles to present "significant" testing, "stars" commonly used
- Proof of hazard versus Proof of safety?
- Challenges to the statistician
  - Multiplicity
  - Statistical power
  - Interpretation of statistical significance tests

How is Drug Safety assessed?

Proof of Hazard	Proof of Safety
Detect a possible effect	Prove Harmlessness of a Drug

- If p-value > 5% ➡ declare compound harmless
- If p-value < 5% ➡ declare compound harmful
  - Only if the subsequent assessment of "biological relevance" agrees with the statistical conclusion.
- Hothorn (2014) advocates for alternative approach

Absence of Evidence vs. Evidence of Absence

Proof of Safety

- No consensus on relevant thresholds (many endpoints, many species, ages, sex,..!)
- Alternatives
  - **Confidence intervals** assessments post-hoc contained within a safety range
    - Treatment effect size and uncertainty
    - Explicitly interpretable in terms of biologically meaningfulness
  - Informal "proof of safety" assessment through **historical control data**, e.g. if the combined sample distribution of the treated groups fall within the historical control sampling distribution

Proof of Hazard

- No consensus has emerged to suggest standard practice
- No regulatory guidance

- Some argue not to use any multiplicity adjustment (neither against several doses, nor against multiple endpoints)
  - Desirable to accept an increased risk of false positives over false negatives (safety context!)
  - Multiplicity adjustments will reduce power (critical as most sample sizes not calculated based on power consideration)

- **ALTERNATIVE**
  - False Discovery Rate (Benjamini and Hochberg (1995)

Preclinical/Tox Perspective

- Current approaches satisfy current regulatory requirements. Lack of regulatory progress is a barrier to improved practice.
- **Data management and analysis systems are common in the industry, governed by complex rules that make changes difficult.**
  - Raise awareness of limitations of statistical analyses
  - Practice of "reporting p-values with stars" is likely to continue but should be qualifed as a starting point only.
  - Further investigations may be warranted
  - 'lack of a star' does not signify 'safety'
  - **Training is KEY!**
- Data-driven prior information from historical control data could be used to justify the choice of priors in **Bayesian approaches** for early drug safety assessments in many instances
- Statistical significance does not equate to biological importance.

Specific Recommendations

Carcinogenicity Studies

- Recognize the widespread use of a *Proof of Hazard* approach
- Proof of Safety approach represents a real “break from tradition”
- Strongly support Hothorn and Hasler (2008) recommendation: use of confidence intervals allowing for both approaches in their paper
  - Recommend wider adoption by statisticians supporting toxicology studies

Specific Recommendations

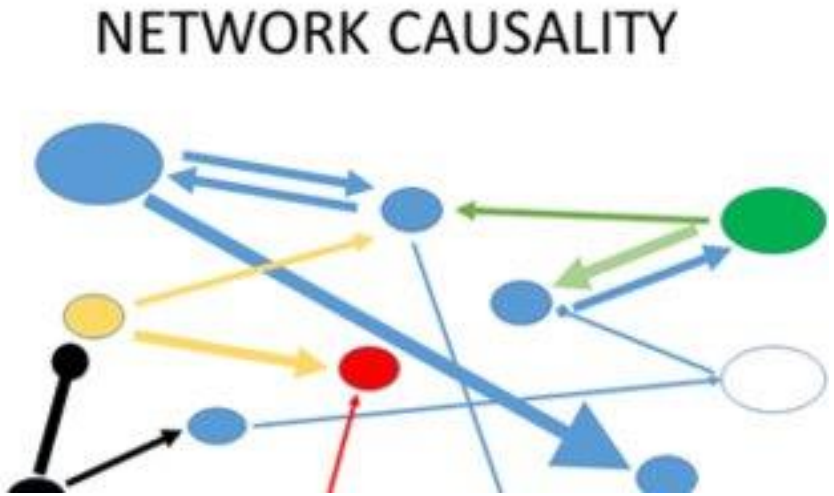
Genetic Toxicology Studies

- Current guidelines provide details on how to classify compound as negative or positive
- They do not cover additional information required for risk assessment
  - Dose-response fitting
  - Point of Departure (MacGregor et al. 2014)
- Adds to weight of information available on compound and aids the evaluation of genotoxicity
- Guidance Document on Revisions to OECD Genetic Toxicology Test Guidelines
  - Recommendations discourage over-reliance on p-values associated with the statistical significance of differences found by pair-wise comparisons.
  - Statistical significance based upon a particular p-value is relevant, but is only one of the criteria used to decide whether to categorize a result as positive or negative.

Specific Recommendations

Safety Pharmacology Studies

- No strong objection to challenge the $p<0.05$ rule provided cutoff is used to address appropriate question
- Confidence intervals and power analysis should accompany the p-values
- Causality and network analyses (DaSilva et al. 2019, Lazic et al. 2020)
  - could bring useful insights to hypothesis formulation and testing
  - Complement current approach to statistical significance

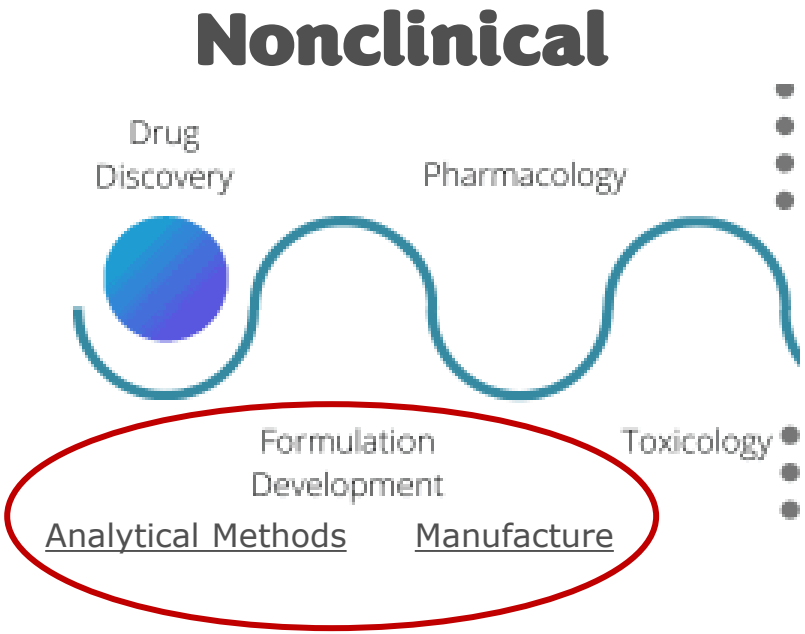


Chemistry, Manufacture & Control (CMC)

Viral exacerbation at 40x magnification

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

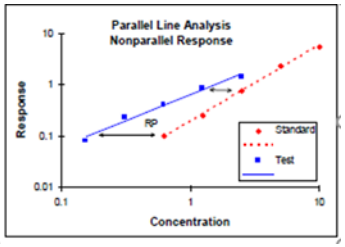
Chemistry, Manufacture & Control

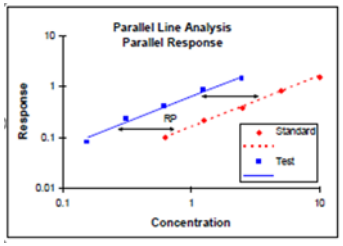


- **Formulate** a shelf stable drug product, with consistent bioavailability
- **Develop analytical methods** to track its chemical and physical properties to permit accurate and precise quality management
- **Engineering studies to permit large scale manufacture** (continuous manufacture is revolutionizing the industry)

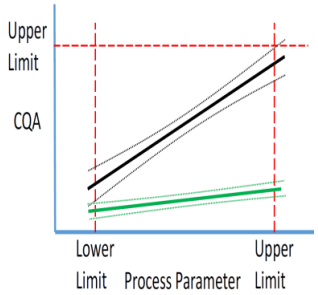
Four Applications Reviewed

1. Similarity testing in bioassay
   - Parallelism between dose response curves is required for a valid RP estimate

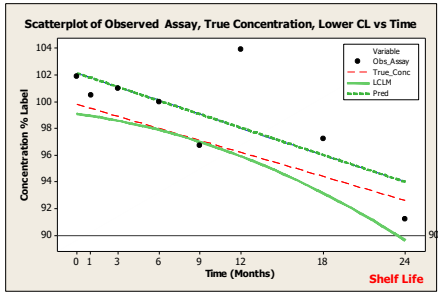




2. Critical Process Parameter (CPP) determination
   - Manufacturing Experiment (DoE) to identify CPPs unambiguously



3. Stability Modeling and Pooling tests (ICHQ1E)
   - Determination of shelf life
   - Complicated regulatory rules for achieving a final model



4. QbD Design Space Construction

Issues and Recommendation

Similarity Testing in Bioassays

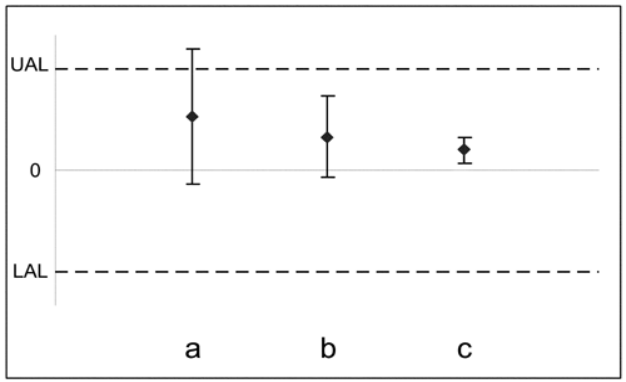
Traditional Approach

- $NHST\ \ H_0: \Delta = 0\ vs\ \ H_1: \Delta \neq 0$
- Δ : Difference in slopes
- Decision rule : IF p-value < 0.05 Reject H_0

- Issues
  - Penalizes labs for improved precision
  - Rewards labs with poor precision

Recommendation

- Test of Equivalence
  - Encourages better design
  - Equivalence margin based on a meaningful criterion.



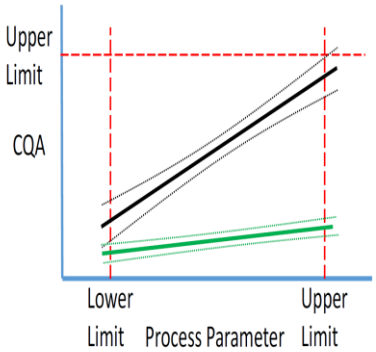
- Bayesian approach
  - Posterior probability of similarity

Issues and Recommendation

Critical Process Parameter Identification

Traditional Approach

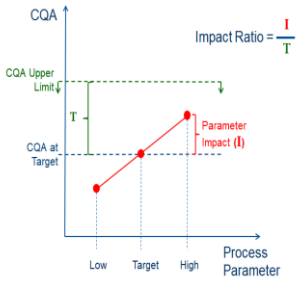
- $NHST\ \ H_0: \Delta = 0\ vs\ \ H_1: \Delta \neq 0$
- Δ : Effect of varying factor



- Issues
  - Design Space in QbD
    - P-value vs Quality Impact
    - Operating Region
- P-value based decision rule not always useful to determine CPPs

Recommendation

- Practical Significance (Wang et al. 2016, Hakemeyer et al. 2016)
  - Impact ratio's
  - Z-scores



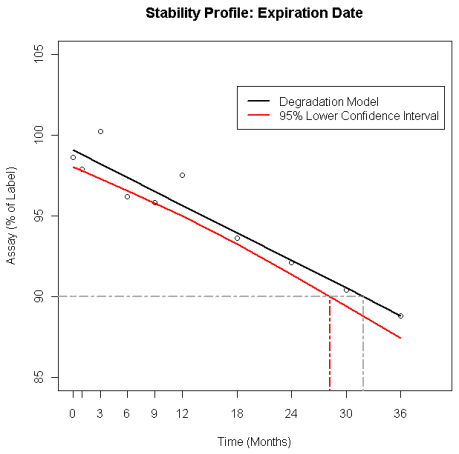
  - Bayesian approach
    - Calculate the posterior probability (quality metric) that variation in each CPP will result in a quality failure.
    - Loss function approach

Issues and Recommendation

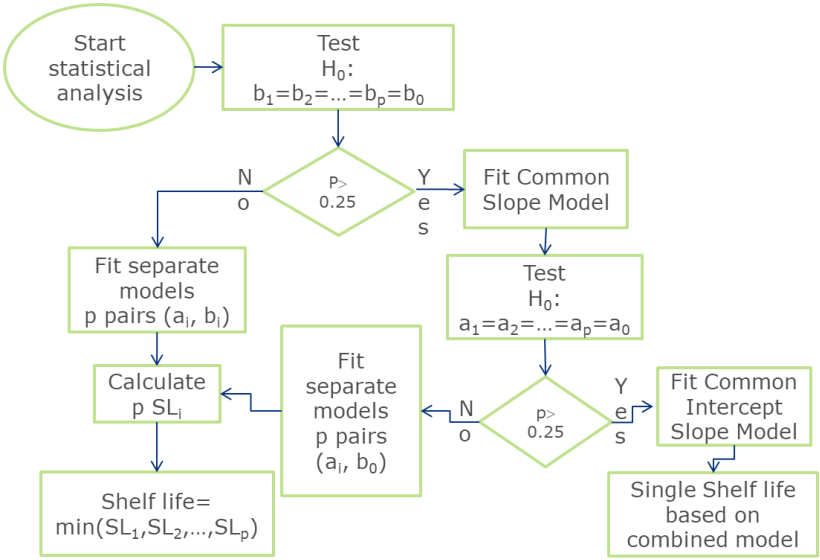
ICH Q1E Stability Modeling

Regulatory Approach

- Stability design
  - 3 batches
  - Linear model $y_{ij} = A_i + b_i \cdot t_{ij} + e_{ij}$
  - Complicated Pooling Rules



- Issues
- Batches are not identical at release
- Chemistry is independent of batch
- Residual error term for pooling intercepts
- P=0.25 is a disincentive to good control of process and analytical variability
- Cannot power a stability study design emphasis is on limited resources



Issues and Recommendation

ICH Q1E Stability Modeling

Recommendations

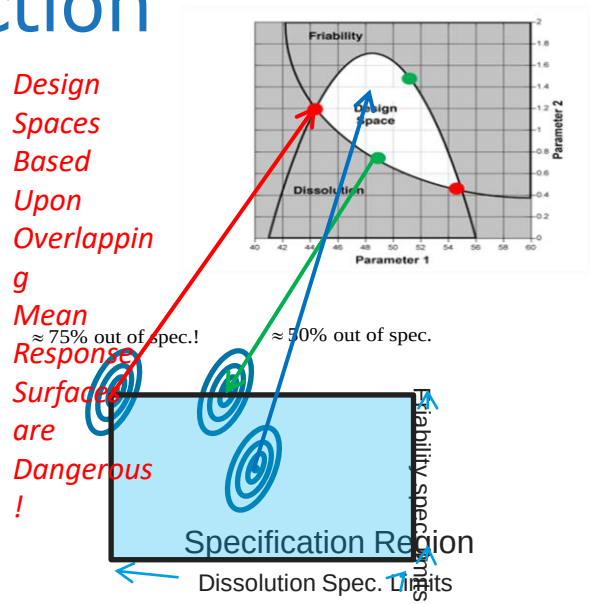
- ICHQ1E poolability rules impose undue burden on the industry
- QbD says exploit the science –the ICHQ1E pooling rules should be qualified light of current scientific knowledge and technology
- A Mixed Effects model is a more natural representation of a fixed manufacturing process
  - –Bayesian framework can incorporate process engineering and scientific judgment

Issues and Recommendation

QbD Design Space Construction

Traditional Approach

- Overlapping means approach, *p*-values, to construct a multidimensional combination of critical material attributes and process parameters that assure quality.



- Issues
  - Computational difficulties and multiplicities .
  - *p*-value and confidence limit approaches do not provide a risk region interpretable as a probability.

Recommendation

- Bayesian approach
  - A risk formulation* based on a posterior-predictive probability measure overcomes the frequentist shortcomings.

Summary

Recurrent Themes

- The p-value is only one piece of a large puzzle and should be interpreted in relation to scientific judgment and prior knowledge.
- Rules that simply rely on the textbook p-value alone for decision making will not be optimal for determining practical significance
- Determination of practical significance is a judgment that should combine experimental results with external information and is best done by statisticians and domain knowledge experts working together
- Bayesian methods and decision making based on posterior probability calculations should be used more widely

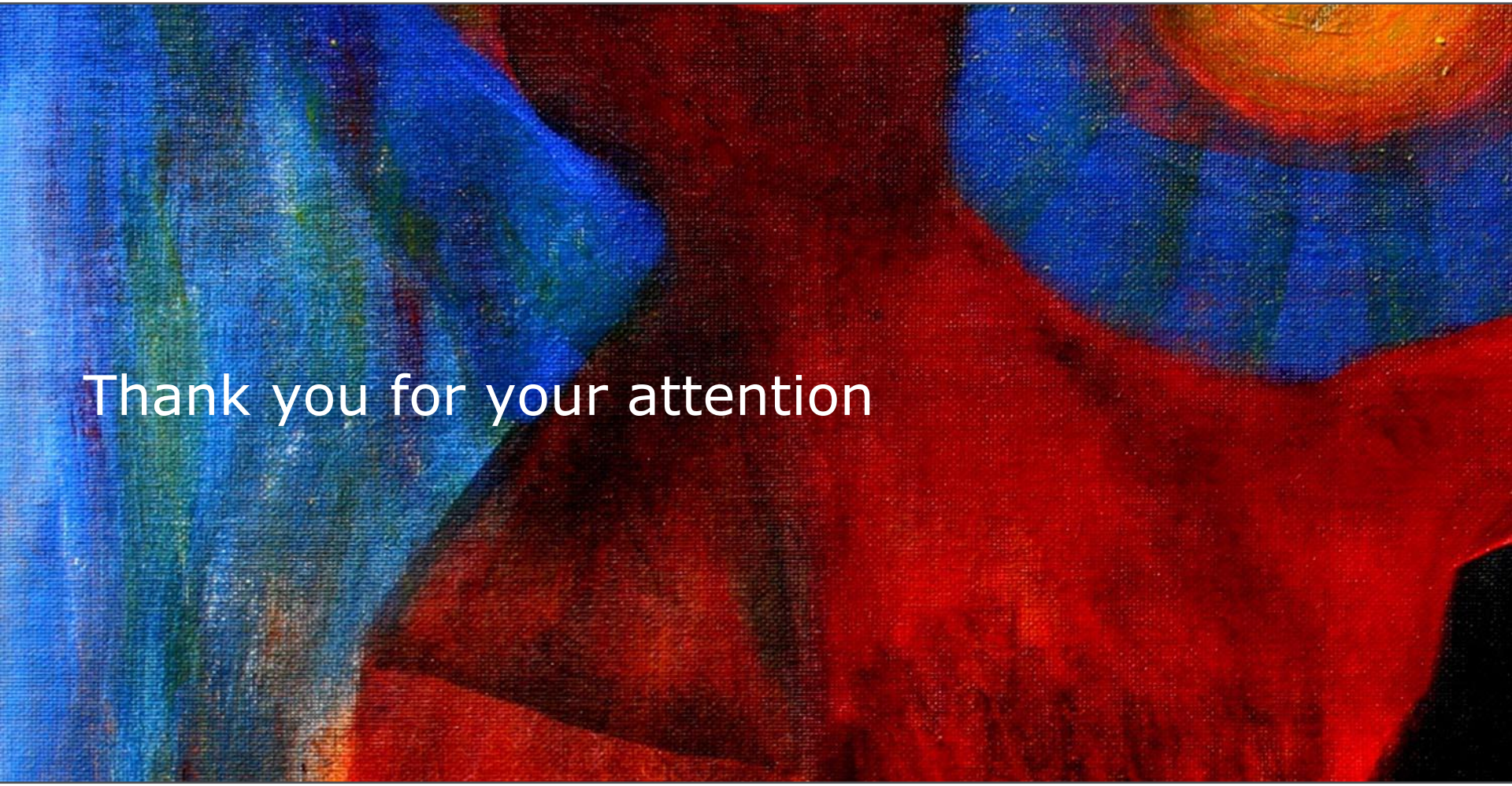
Specific recommendations by application area

- P-values can be useful in providing a go/no-go decision metric for discovery/-Omics
- No strong objection to challenge the $p<0.05$ rule provided cutoff is used to address appropriate question in preclinical toxicology studies
  - Data-driven prior information from historical control data could be used to justify the choice of priors in preclinical safety studies
- P-values or posterior probability calculations for An equivalence approach is recommended for CMC applications

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Final Remarks

- A review of statistical practice of three subject matter areas was carried out by a group of experienced statisticians focusing on decision making having commercial implications, as opposed to the more general concerns related to reproducibility in academic clinical and epidemiological research.
- The experience has shown the potential benefits to the profession of comprehensive discussions and consensus building
  - Not all applications were reviewed, or reviewable due to some confidentiality concerns
  - It's likely that as technology evolves, the regulatory landscape changes, similar groups may be convened to improve practice
  - Data Science practices were not included in this cycle of reviews



janssen | PHARMACEUTICAL COMPANIES OF Johnson&Johnson

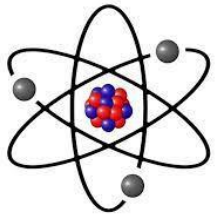
Sample Papers and their titles and years

Should we stop using the P value in descriptive studies?	1977
We should stop misuse of P value	1978
What is the p value and what is it worth?	1984
Misconception concerning the ubiquitous p value	1985
Power of the P value	1987
The Enduring Evolution of the P Value	2016
The P-value and the problem of multiple testing	2016
Are the fallacies of the P value finally ended?	2016
A clash of cultures in discussions of the P value	2016
P Value: Significance Is Not All Black and White	2016
Replication: Do not trust your p-value, be it small or large	2021
The frequent insignificance of a "significant" p-value	2022
Beyond “p<0.05 was considered statistically significant" and other cut-and-paste statistical methods	2022
An observational analysis of the trope "A p-value of < 0.05	2022
The evidence contained in the P-value is context dependent	2022

ASA's Statement on P-values (Wasserstein & Lazar, 2016)

- *P*-values can indicate how incompatible the data are with a specified statistical model.
- *P*-values do not measure the probability that the studied hypothesis is true or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based solely on whether a *p*-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A *p*-value does not measure the size of an effect or the importance of a result.
- By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

"Moving to a World Beyond "p < 0.05" (Wasserstein, Schirm & Lazar, 2019)



Don't Say "Statistically Significant":

- Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless.

Embrace an appropriate attitude regarding the role of Statistics in research:

- "**A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest.
- Remember "**ATOM**"


janssen
PHARMACEUTICAL COMPANIES OF
Johnson & Johnson