Predictive Evidence Threshold Scaling:
does the evidence meet a confirmatory standard?

Beat Neuenschwander, Novartis

InSPiRE Conference, Warwick, UK, 26-28 April 2017

Joint work with Satrajit Roychoudhury and Michael Branson

Acknowledgments: Simon Wandel, Heinz Schmidli, and David Spiegelhalter

NOVARTIS

Abstract

Making better use of evidence is one of the tenets of modern drug development. This calls for an understanding of the evidential strength of non-confirmatory evidence relative to a confirmatory standard. *Predictive evidence threshold scaling (PETS)* provides a framework to do so. Under *PETS*, the evidence meets a confirmatory standard if the predictive probability of a positive effect reaches the predictive evidence threshold from hypothetical confirmatory data. Obtaining these probabilities requires hierarchical models with plausible heterogeneity and bias assumptions. After introducing the methodology, I will discuss two examples. The first is childhood Guillain-Barré syndrome, with sparse children data enriched with adult data. The second is breakthrough designation, illustrated by a recent FDA approval of Crizotinib for non-small-cell-lung-cancer based on phase I and II data. The examples suggest that the evidential strength of non-confirmatory data can meet a confirmatory standard. This is reassuring for modern drug development, which exploits various types of evidence to inform licensing decisions.

Outline

- Scope & Objective
- Predictive Evidence Threshold Scaling (PETS)
  - Idea
  - Methodology
- Examples
  1) Crizotinib for NSCLC
  2) Plasmapheresis for childhood Guillain-Barré syndrome
- Conclusions

Scope and Objective

Scope & Objective

Problem statement

- Problem
  - For a treatment effect parameter θ, we want to compare the evidential strength of two data sources Y_E and Y_C
  - Which one provides more evidence for a treatment effect?
- Question:
  - Why? If we have two relevant data sources, why don't we combine them to inform θ?
- Answer:
  - Only one is observed (Y_E), the other (Y_C) is hypothetical

Scope & Objective

Example 1: breakthrough therapy designation

- Breakthrough therapy
  - an FDA designation that expedites drug development (FDA Safety and Innovation Act, July 9, 2012)
  - unmet clinical
  - *real world evidence (RWE),* data outside well-controlled clinical trials, can be used
  - effect sizes are large
- How does RWE (Y_E) compare to a confirmatory standard (Y_C)?

Scope & Objective

Example 1: Crizotinib in non-small-cell lung cancer (NSCLC)

- Promising NSCLC progression-free survival (PFS) data: median 8-9 months

Trial	median (95%-CI)	y (s)*
PROFILE 1001	9.7 (7.7, 12.8)	2.272 (0.130)
PROFILE 1005	8.1 (6.8, 9.7)	2.092 (0.091)

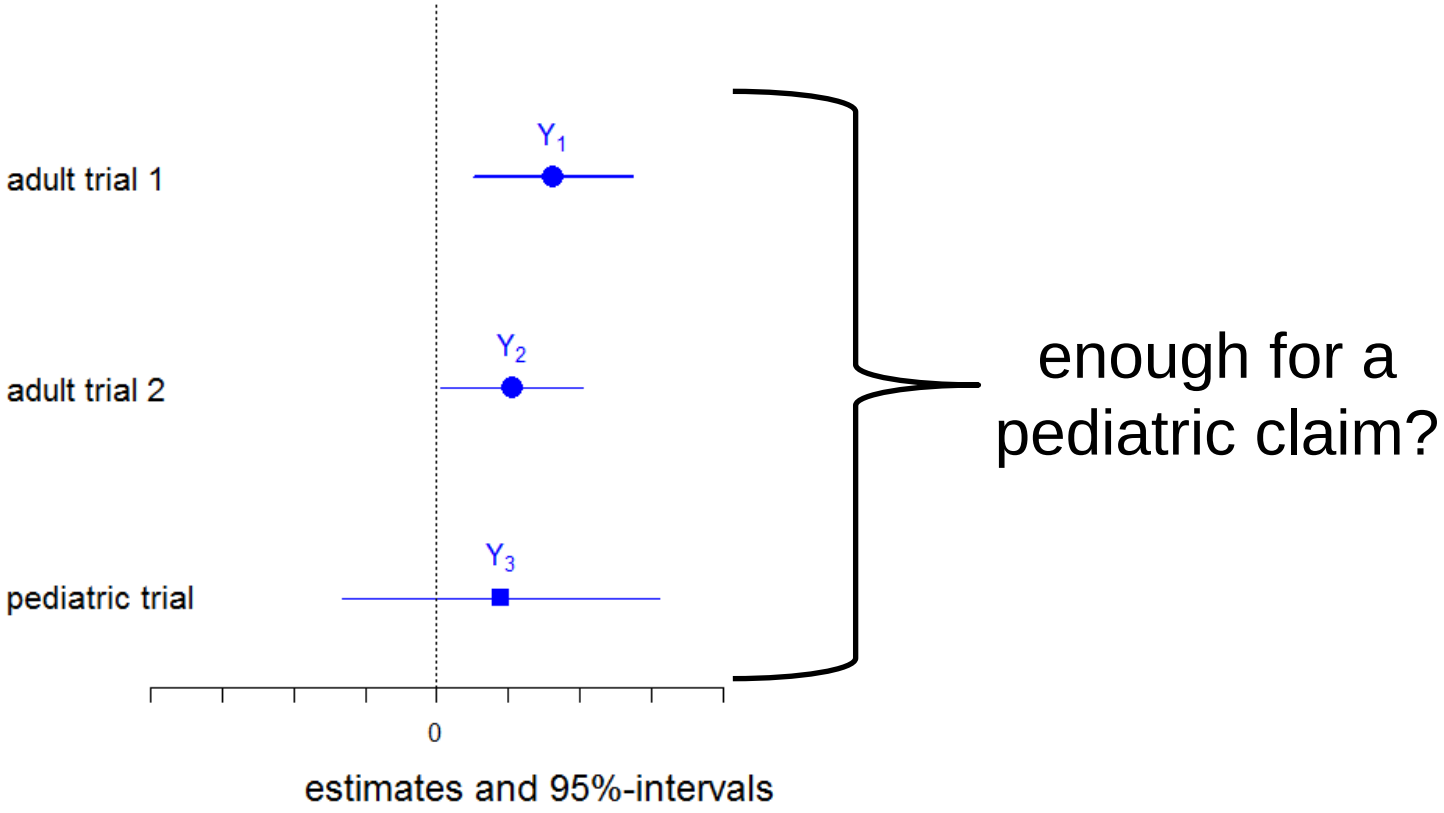
* normal approximation: est (se) of log-median PFS

  - phase I expansion (PROFILE 1001) and phase II single-arm trial (PROFILE 1005)
  - typical (control) median-PFS is 3 to 4 months
  - FDA granted breakthrough designation
  - How do these data compare to a confirmatory standard?

Scope & Objective

Example 2: extrapolation from adults to pediatrics

Assume with have promising adult evidence for a treatment effect. How much pediatric data is needed?



NOVARTIS

Scope & Objective

Quantifying Real World Evidence

- Setting: actual **RWE Y_E** for a clinical endpoint
- Objective: to propose a quantitative approach that
  - allows comparing the actual evidence **Y_E** to a confirmatory standard
  - complements and improves qualitative decisions
- Disclaimer: what follows
  - is not meant to replace the standard confirmatory approach
  - is meant to complement it

Predictive Evidence Threshold Scaling

Idea

Predictive Evidence Threshold Scaling (PETS)

Three requirements

Three requirements

1. a *confirmatory standard:* (hypothetical) data $\mathbf{Y_{(C)}}$
2. a *metric* to compare $\mathbf{Y_E}$ to $\mathbf{Y_{(C)}}$
3. a *rule* to decide whether the non-confirmatory data is sufficiently strong

PETS

Hierarchical structure

- Actual, non-confirmatory data $\mathbf{Y_E}$ from J sources
  - estimates $Y_1, Y_2, \ldots Y_J$
  - standard errors $s_1, s_2, \ldots s_J$
  - **parameters** $\boldsymbol{\theta_1, \theta_2, \ldots \theta_J}$
- Hypothetical (minimal) confirmatory data $\mathbf{Y_{(C)}}$
  - e.g., two significant trials; or one in Oncology
  - estimates $Y_{(1)}, Y_{(2)}$
  - standard errors $s_{(1)}, s_{(2)}$
  - **parameters** $\boldsymbol{\theta_{(1)}, \theta_{(2)}}$
- The effect parameters differ (heterogeneity!)

PETS

Metric

- Metric to compare actual and hypothetical confirmatory evidence
  - metric should be trial-independent—not the effect parameter of one of the trials in the database!
  - choice: probability of a «positive» effect θ_P in a new trial

$$pr(\theta_P > 0 \mid data)$$

  - note: inequality cutoff may be non-zero (e.g. NI trials)

PETS

Three heterogeneities

- Heterogeneities: deviations from mean value μ
  - for effect parameters in actual trials τ_E
  - for effect parameters in confirmatory trials τ_C
  - for effect parameter in new trial τ_P

$$\theta_1, \theta_2, \ldots, \theta_J \approx \theta_P \approx \theta_{(1)}, \theta_{(2)}$$

$$\Downarrow$$

If parameters are similar, the actual evidence Y_E will have higher confirmatory relevance

If parameters differ considerably, the evidence will be discounted due to larger heterogeneity

PETS

Predictive Evidence Probability (PEP) and Threshold (PET)

- Scaling of Y_E vs. $Y_{(C)}$
- For the actual evidence Y_E

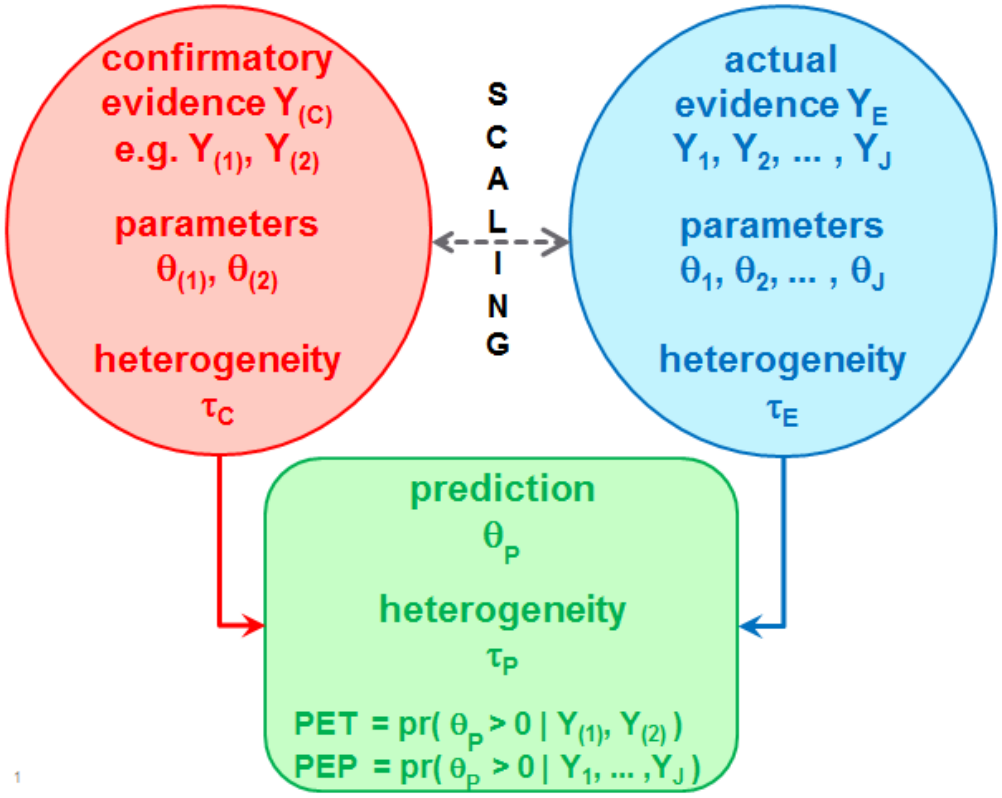
  PEP (predictive evidence probability) $\quad pr(\theta_P > 0 \mid Y_E)$

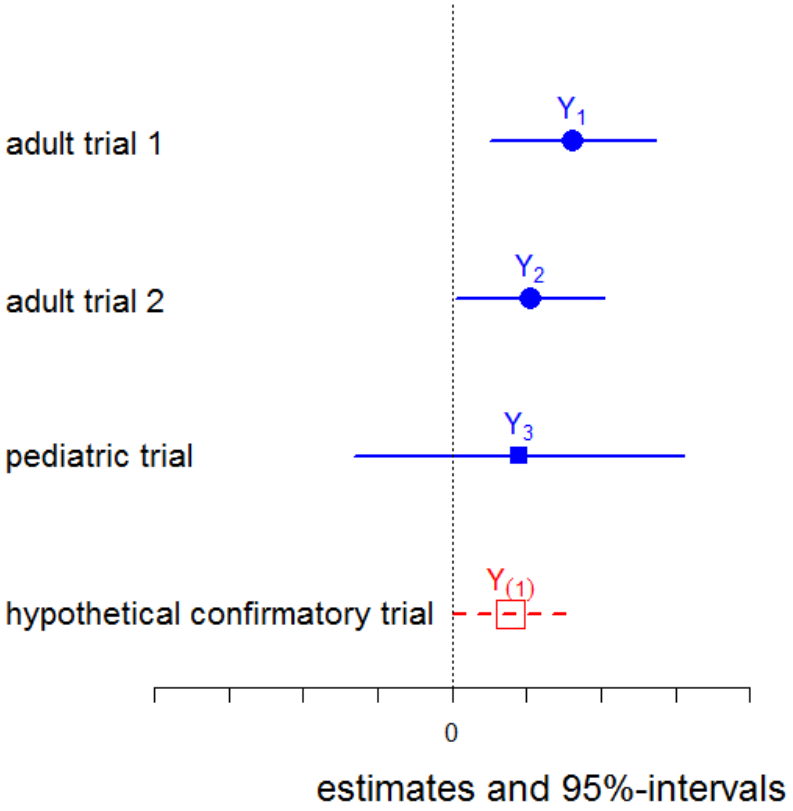
  - predictive probability of a «positive effect» (in a new trial)
- For the (hypothetical) confirmatory evidence $Y_{(C)}$

  PET (predictive evidence threshold) $\quad pr(\theta_P > 0 \mid Y_{(C)})$
- How large are PEP and PET? Is PEP $\geq$ PET

PETS

PETS framework: summary





NOVARTIS

Predictive Evidence Threshold Scaling

Methodology

PETS

Methodology: hierarchical model

- The normal-normal hierarchical model (NNHM)
  - (approximately) normally distributed estimates Y
  - normally distributed parameters θ
- Heterogeneity parameters τ_C, τ_P, τ_E
  - similar (or equal) small *confirmatory and predictive heterogeneity*, $\tau_C \approx \tau_P$, since confirmatory setting is more revevant
  - two approaches
    - assumed parameters $\rightarrow$ sensitivity analyses for plausible scenarios
    - or uncertain parameters, with prior distributions on parameters
    - choices must be sensible (context-specific)

PETS

Normal-normal hierarchical model and predicted effect θ_P

- NNHM with differential heterogeneity
  - data model $\boldsymbol{Y_k|\theta_k, s_k^2 \sim N(\theta_k, s_k^2)}$
  - parameter model $\boldsymbol{\theta_k|\mu, \tau_k^2 \sim N(\mu, \tau_k^2)}$
  - $\tau_k = \boldsymbol{\tau_E}$ for actual, $\tau_k = \boldsymbol{\tau_C}$ for confirmatory evidence
  - prediction $\boldsymbol{\theta_p|\mu, \tau_p^2 \sim N(\mu, \tau_p^2)}$
  - note: standard meta-analysis uses a common τ
- Two calculations with NNHM: PET and PEP
  - PET: pr($\boldsymbol{\theta_p > 0}$ | confirmatory data $\mathbf{Y_{(c)}}$)
  - PEP: pr($\boldsymbol{\theta_p > 0}$ | actual data $\mathbf{Y_E}$)

PETS

NNHM PET and PEP calculations for fixed heteogeneities

- PET and PEP calculation for fixed τ parameters
  - Bayesian with flat prior for μ

$$\boldsymbol{\theta_P | Y_1, \ldots \sim N(\hat{\mu}, \frac{1}{w_+} + \tau_p^2)}$$

$$\hat{\mu} = \sum_k w_k Y_k \, / w_+$$

$$w_k = \frac{1}{s_k^2 + \tau_k^2} \quad \text{(precisions)}$$

$$w_+ = \sum_k w_k \quad \text{(total precision)}$$

  - «equivalent» classical result: $\boldsymbol{\hat{\theta}_P = \hat{\mu}, \quad \widehat{se}^2 = \frac{1}{w_+} + \tau_p^2}$

PETS

Extensions

- Other sampling models
- Analyses with uncertainty for τ
- Inclusion of covariates
- Individual patient data
- ...
- *Systematic biases*

PETS

Extensions: systematic biases

- So far: no systematic biases assumed.
  All distributions centered at μ
- (Sensitivity) analyses with systematic biases
  - allow for trial-specific biases $\boldsymbol{\delta_k}$
  - require judgement about plausible bias scenarios
  - simple model extension

$$\boldsymbol{\theta_k | \mu, \tau_k^2, \delta_k \sim N(\mu + \delta_k, \tau_k^2)}$$

  - biases
    - can be fixed (scenarios) or uncertain (priors)
    - but must be plausible

Example 1:

Breakthrough Designation

Crizotinib for NSCLC

Crizotinib

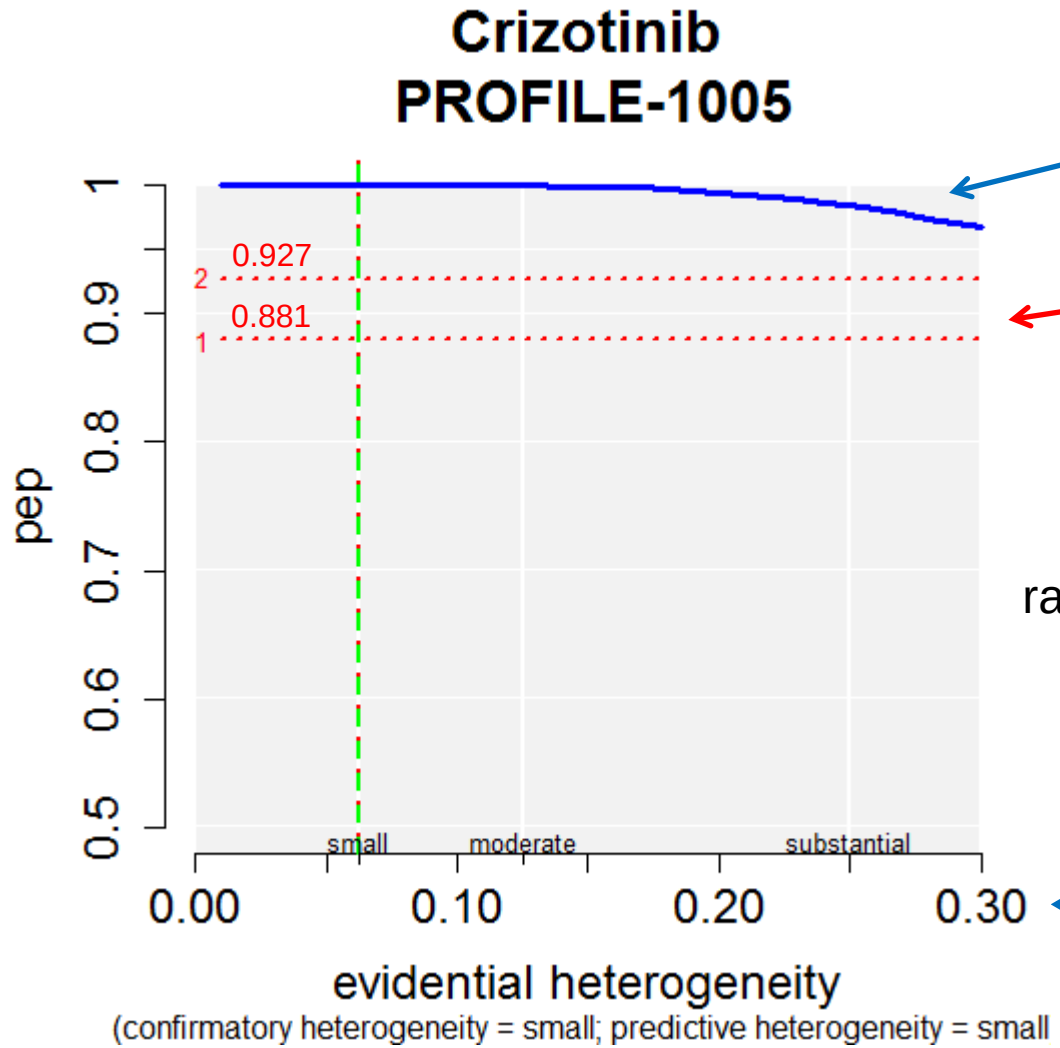
Data

Trial	median (95%-int)	y (s)
actual data		
PROFILE 1001	**9.7 (7.7, 12.8)**	**2.272 (0.130)**
PROFILE 1005	**8.1 (6.8, 9.7)**	**2.092 (0.091)**
hypothetical confirmatory data (one trial)		
CONF*	5.12 (4.5, 5.83)	1.635 (0.066)

* one confirmatory trial with 225 events;
H_0: θ = log(4.5 months); σ=1; one-sided p-value = 0.025.

Crizotinib

PETS graph: PEP vs. PET (single-arm analyses)



PEP

pr($\theta_P > 0 \mid Y_E$)

PET

(for 1 or 2 conf. trials)

pr($\theta_P > 0 \mid Y_{(C)}$)

<u>Conclusion</u>

PEP dominates PET for a wide range of evidential heterogeneities

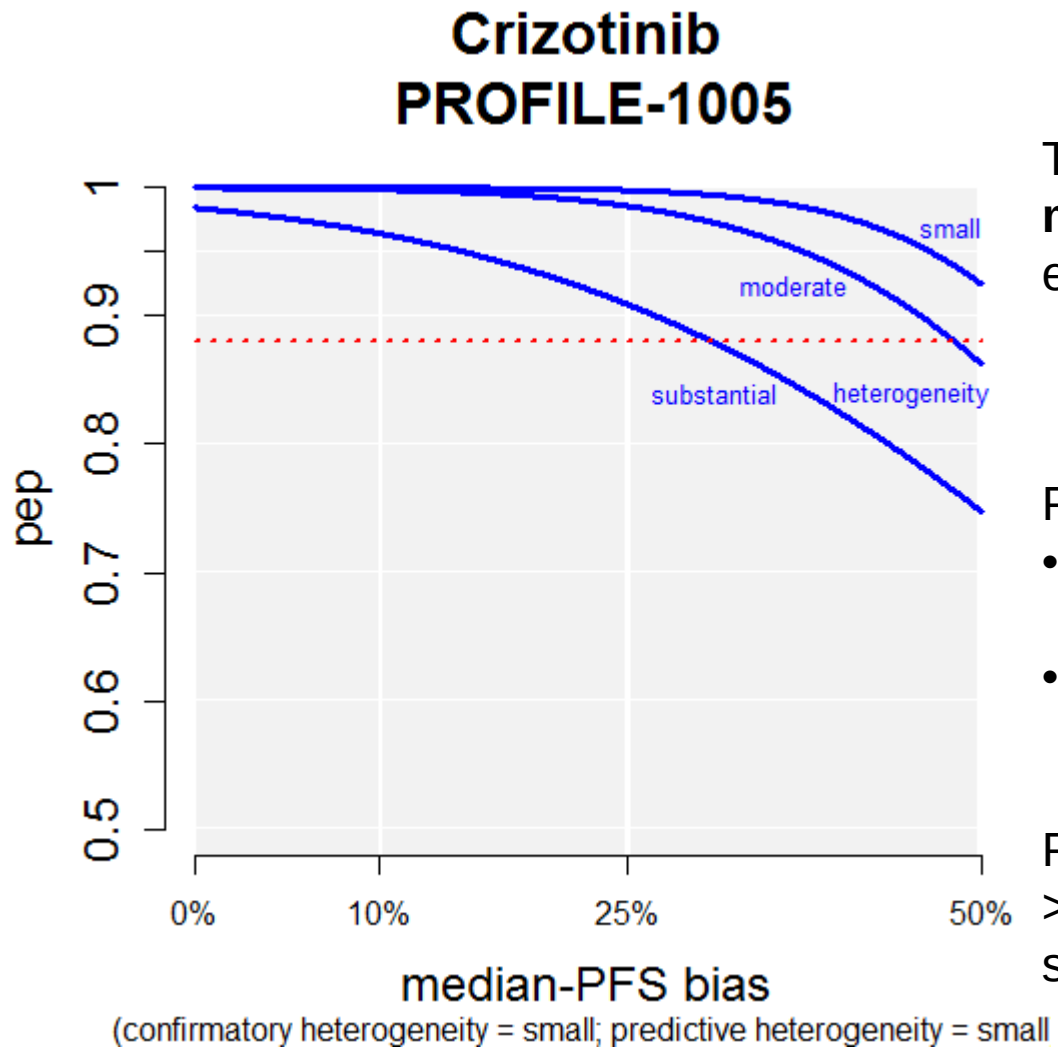
evidential heterogeneity τ_E «no to substantial»

fixed $\tau_C = \tau_P = 0.0625$

NOVARTIS

Crizotinib

Phase II trial: systematic bias sensitivity analyses



Heterogeneities

Three blue lines are for **small**, **moderate**, and **substantial** evidential heterogeneity (τ_E)

Conclusions

PEP dominates PET

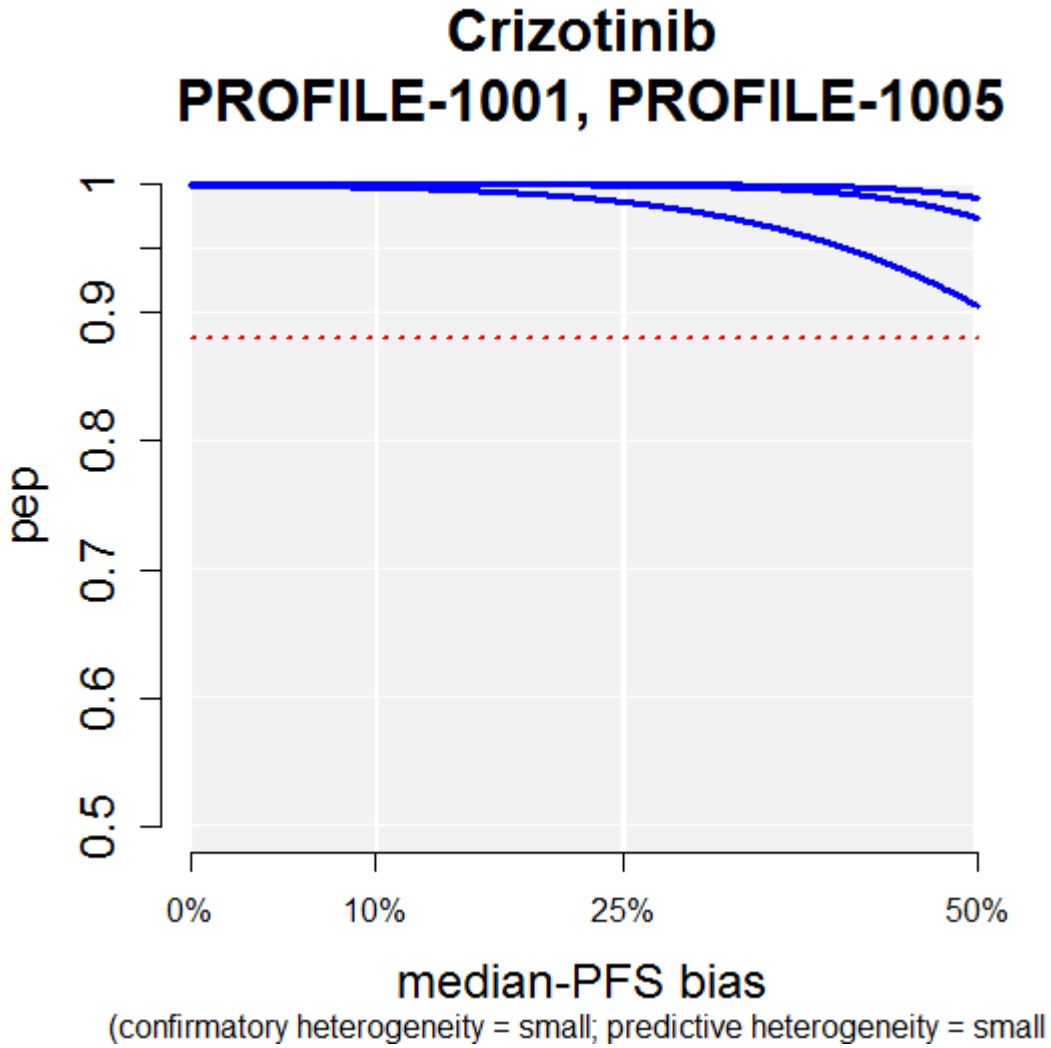
- for small to substantial heterogeneity if bias is <25%
- for small to moderate heterogeneity if bias <50%

PEP is not sufficient if bias is > 25% and heterogeneity is substantial (plausible?)

NOVARTIS

Crizotinib

*Bias sensitivity analyses **using both trials***



Heterogeneities

Three blue lines are for **small**, **moderate**, and **substantial** evidential heterogeneity (τ_E)

NOVARTIS

Crizotinib

Later confirmatory data were consistent with earlier data

Median-PFS results		
Trial	median (95%-int)	y (s)
actual non-confirmatory data		
PROFILE 1001	**9.7 (7.7, 12.8)**	**2.272 (0.130)**
PROFILE 1005	**8.1 (6.8, 9.7)**	**2.092 (0.091)**
hypothetical confirmatory data		
CONF	5.12 (4.5, 5.83)	1.633 (0.066)
later confirmatory data		
PROFILE 1007*	**7.7 (6.0, 8.8)**	

* Randomized phase 3 trial PROFILE-1007 with standard 2nd line chemotherapy (pemetrexed or taxotere) confirmed the effect of Crizotinib. **Median PFS for chemotherapy: 3 (2.6, 4.3)**

Crizotinib

Complication: single-arm vs. randomized

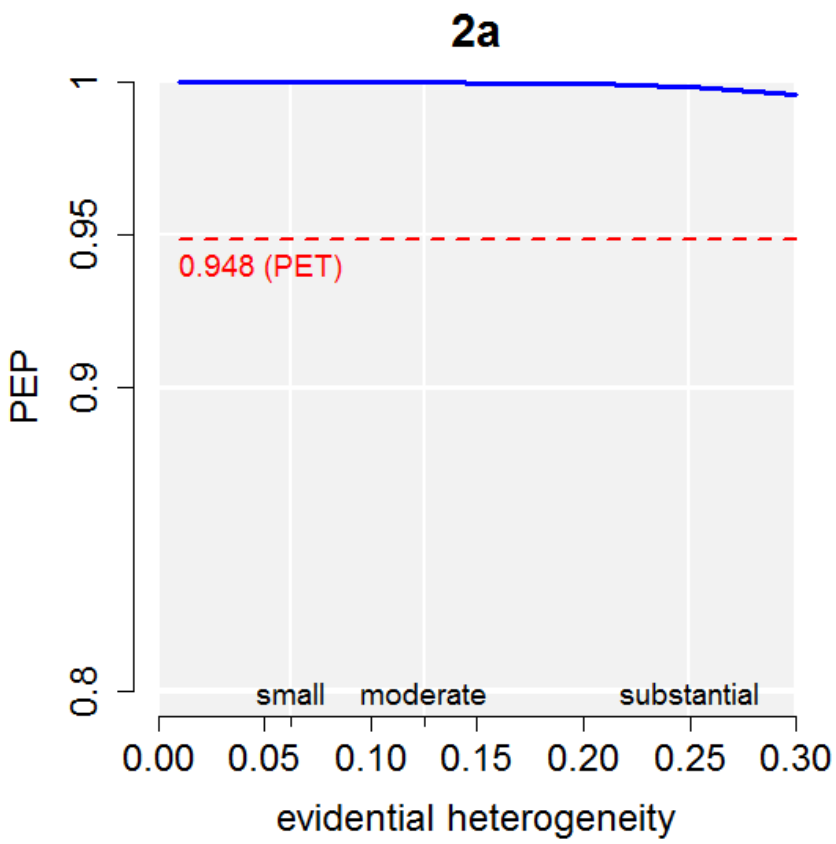
- Phase I and II trials were single-arm
- PETS analyses compared Crizotinib to a fixed control median of 4.5 months
- What about the randomized setting?
  - hazard-ratio Critzotinib vs. SoC. Two scenarios
    1. assuming a fixed control effect: median = 4.5 months
    2. **assuming uncetain control effect: median = 4.5 months (worth ~ 50 events)**
  - results qualitatively similar to single-arm PETS analyses

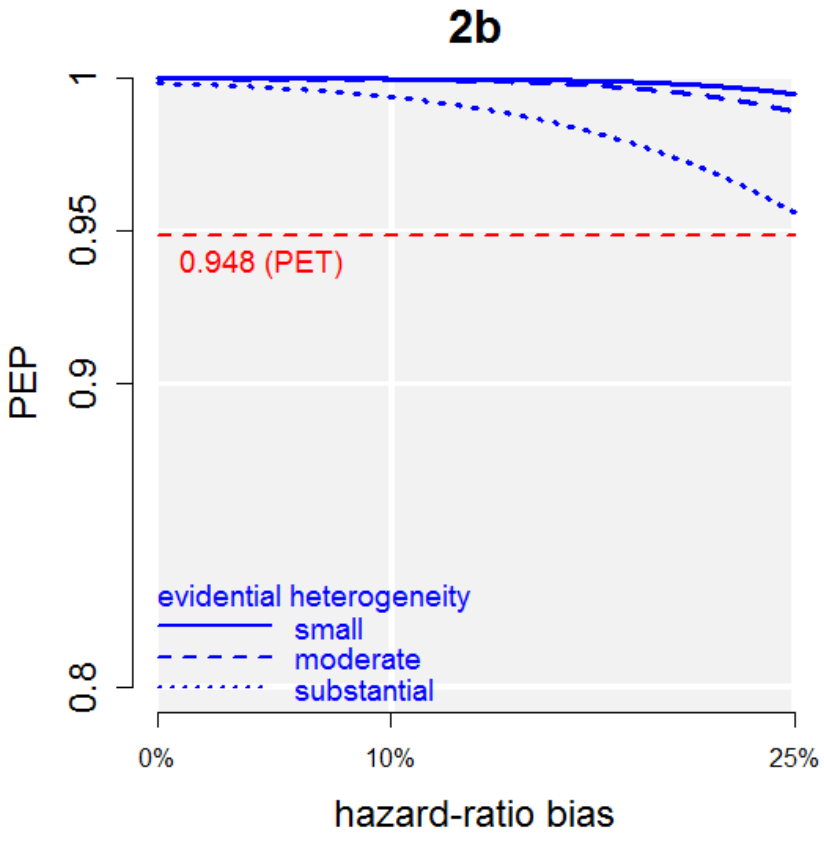
Crizotinib

PETS for hazard-ratio scale (uncertain control median)

without bias

with bias
(3 heterogeneity scenarios)





NOVARTIS

Example 2:

Extrapolation from Adults to Pediatrics

Plasmapheresis for Guillain-Barré Syndrome (GBS)

Source: Goodman & Sladky (2004)

Plasmapheresis for childhood GBS

Introduction

- Guillain-Barré syndrome
  - a rare neurologic disease
  - affects all age groups, but is more common in children
  - main treatments:
    - ***plasmapheresis (plasma exchange, PE)***
    - *intraveneous immune globulin (IVIg)*
  - Both treatments were shown to be effective in adults and then used in children off-label
- Here, we apply PETS to ***PE***
  - to predict efficacy in children, using adult data (and sparse children data)

Plasmapheresis for childhood GBS

Data

- A1-2: 2 trials in adults in the 1980s
- C1-C4: 4 small trials in children in the 1990s
- Endpoint: time to independent walking
- Does the evidence from these trials meet a confirmatory standard? For example, for trials A1, A2, and C1

	HR	95%-CI	y	s
A1. McKhann 1985	0.62	(0.46-0.84)	-0.472	0.153
A2. Raphael 1987	0.63	(0.47-0.84)	-0.461	0.149
C1. Epstein 1990	0.4	(0.17-0.94)	-0.916	0.434
C2. Lamont 1991	0.4	(0.16-1.03)	-0.916	0.481
C3. Jansen 1993	0.55	(0.23-1.34)	-0.598	0.455
C4. Graf 1999	1.52	(0.54-4.29)	0.419	0.529

Plasmapheresis for childhood GBS

PETS scenario analyses: assumptions

- Scenario assumptions for heterogeneities/biases
  - Y_E: actual trials
    - **3 heterogeneity scenarios** for adult/children trials:
      - moderate/small, substantial/moderate, large/substantial
    - **3 bias scenarios** for children trials
      - 0% (no bias), 10% bias, 25% bias
  - Y_C: one confirmatory children trial (1-sided p-value=0.025)
    - 200 events
    - confirmatory heterogeneity = small

    $\Rightarrow$ PET = 0.95
  - predictive heterogeneity = small

Plasmapheresis for childhood GBS

PETS scenario analyses: results

	heterogeneity: adult/children		
bias (trials C1-C4)	moderate/small	substantial/moderate	large/substantial
	adult trials		
	0.999	0.985	0.894
	adult trials + children trial 1		
no	1	0.997	0.98
10%	1	0.996	0.974
25%	1	0.994	0.958

- Extrapolation based on adult data only is insufficient if heterogeneity is large/substantial: PEP = 0.894
- With 1st children trial (C1), PEP > PET for all scenarios
- Conclusion: strong adult data combined with sparse pediatric data provides sufficient evidence

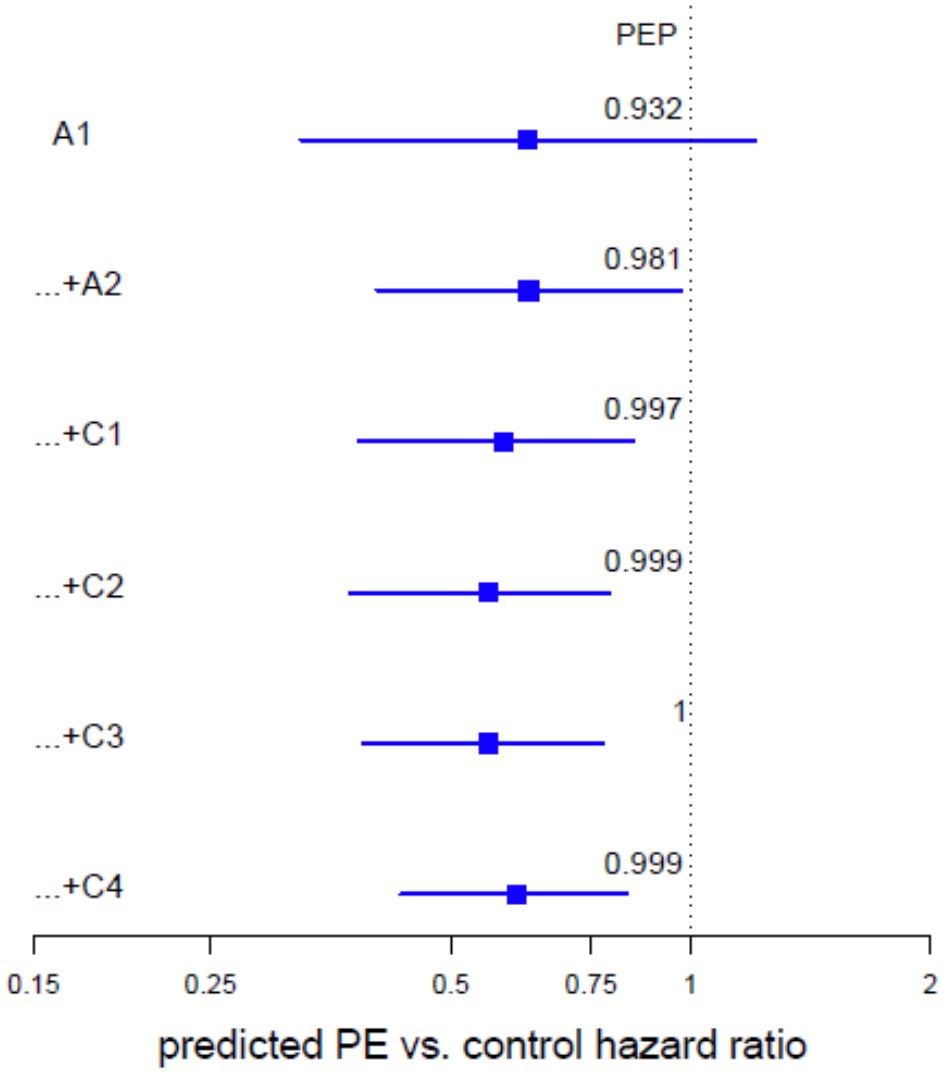
Plasmapheresis for childhood GBS

Bayesian PETS analyses

- Alternative to fixed scenarios: prior distributions on
  - heterogeneities: log-normal priors on τ parameters
  - biases: normal priors on δ parameters
- PETS results
  - are similar if priors cover the range of the fixed scenarios used previously
  - are shown cumulatively on next slide:
    trial A1, trials A1+A2, trials A1+A2+C1, etc.

Plasmapheresis for childhood GBS

Bayesian PETS analyses: cumulative results



Note:
PET = 0.95

NOVARTIS

Conclusions

- Increasing pressure on the pharmaceutical industry
  - scope for innovation is broad (for policy and science)
  - one aspect is to better use the evidence, which includes real-world evidence (21st Century Cures Act)
  - this is challenging and requires that
    1) data are accessibe
    2) data quality is understood
    3) data are properly analyzed (hierarchical modeling)
    4) results of the analysis are properly interpreted
  - PETS contributes to the inferential 3) and 4)
  - NNHM: the basic model (extensions possible)

Conclusions

- PETS has limitations
  - although quantitative, PETS requires contextual judgement about plausible *heterogeneities* and *biases*
  - for these, robustness of PEP>PET is needed
- Examples
  1. robust PETS results for Crizotinib, which clearly support FDA's breakthrough designation
  2. PETS supports intuition that strong adult data combined with sparse children data suffices for a pediatric claim
- NNHM can be easily implemented in *R* (for fixed scenarios) and *WinBUGS/JAGS/Stan* (for priors)

References

- 21st Century Cures Act., May 19, 2015. http://docs.house.gov/meetings/IF/IF00/20150519/103516/BILLS-1146ih.pdf.
- A Blueprint for Breakthrough: Exploring Utility of Real World Evidence (RWE). Friends of Cancer Reserach, Washington DC, June 16, 2016. http://www.focr.org/sites/default/files/pdf/RWE%20-%20Project%20PRE-MEETING%20DRAFT.pdf
- The Food and Drug Administration Safety and Innovation Act (FDASIA), July 9, 2012. http://www.fda.gov/RegulatoryInformation/Legislation/SignificantAmendmentstotheFDCAct/FDASIA/ucm20027187.htm
- Camidge et al. (2012). Activity and safety of crizotinib in patients with ALK-positive non-small-cell-lung cancer: updated results from a phase I study. Lancet.
- Eichler et al. (2015). From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. Clin Pharm & Ther
- Goodman, Sladky (2004). A Bayesian approach to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barré syndrome. Clinical Trials
- Kesselheim, Darrow (2015). FDA designations for therapeutics and their impact on drug development and regulatory review outcomes. Clin Pharm & Ther

NOVARTIS

References

- Kim et al. (2012). Results of a global phase II study with crizotinib in advanced ALSK-posiitive non-small-cell lung cancer (NSCLC). J Clin Oncol.
- Oye et al. (2015). The next frontier: fostering innovation by improving health data access and utilization. Clin Pharm & Ther
- Neuenschwander, Roychoudhury, Schmidli (2016). Using co-data in clinical trials. Statistics in Biopharmaceutical Research.
- Neuenschwander, Roychoudhury, Branson. Predictive evidence threshold scaling: does the evidence meet a confirmatory standard? Submitted to Statistics in Biopharmaceutical Research.
- Schneeweiss et al. (2011). Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. Clin Pharm & Ther
- Spiegelhalter, Abrams, Myles (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluations. Wiley.
- Woodcock (2012). Evidence vs. Access: can twenty-first-century drug regulation refine the tradeoffs? Clin Pharm & Ther