Cele próby klinicznej

- Cele próby
- Kryteria oceny skuteczności leczenia
- Miary skuteczności leczenia
- Wnioskowanie statystyczne (testowanie hipotez)
- Próby nadrzędności i równoważności
- Wielokrotne testowanie
- Istotność statystyczna a kliniczna

Cele próby klinicznej

- Koncentrujemy się na próbach III fazy
- Próba powinna udzielić jednoznacznej odpowiedzi na jedno podstawowe pytanie
  - Czy nowa metoda leczenia E jest skuteczniejsza, biorąc pod uwagę kryterium oceny skuteczności K, od kontrolnej metody C dla pacjentów z chorobą X?
- Dodatkowe pytania
  - prawie zawsze są dołączane (np. analiza podgrup)
  - wymagają ostrożności w interpretacji

Rodzaje prób ze względu na cel

- Nadrzędność (superiority)
  - Skuteczność E jest większa niż C
  - C może być placebo lub aktywnym leczeniem kotrolnym
- Nie-podrzędność (non-inferiority)
  - Z pewną tolerancją, skuteczność E nie jest gorsza niż C.
- Równoważność (equivalence)
  - Z pewną tolerancją, skuteczność E jest taka sama jak C.

Oceny efektu leczenia

- Zwykle używane jest jedno pierwszoplanowe kryterium (primary endpoint), najlepiej odpowiadające celowi próby z punktu widzenia oceny skuteczności leczenia
- Bezpośrednio związane ze stanem zdrowia chorego
- Powinno być wiarygodne, mierzalne bez obciążeń
- Wrażliwe na zmiany stanu zdrowia
  - krótkoterminowe
  - długoterminowe

Przykład: choroby śmiertelne

- W tym przypadku interesuje nas na ogół efekt leczenia na śmiertelność
- Możliwe kryteria oceny skuteczności leczenia
  - zgon w ustalonym okresie (np., w ciągu 1 roku)
  - czas przeżycia
  - czas do zgonu wynikającego z choroby (time to disease-related death)

Pomiary kliniczne

- Dane pozwalające na ocenę efektu leczenia mogą być uzyskiwane w różny sposób dla różnych chorób
  - badania laboratoryjne (np., parametry krwi)
  - badania obrazowe (X-ray, CT, NMR,...)
  - skale oceny objawów (psychiatria)
  - kwestionariusze wypełniane przez chorego (jakość życia)
  - obserwacja zdarzeń klinicznych i/lub objawów
  - ocena przez lekarza

Dodatkowe kryteria oceny efektu leczenia

- Drugoplanowe kryteria oceny skuteczności (secondary endpoints)
  - wspomagające interpretację pierwszoplanowego, lub
  - związane z drugoplanowymi celami próby
- Kryteria ekonomiczne
  - dane dotyczące używania pomocy medycznej przez pacjentów, w powiązaniu z informacją o kosztach
  - Pozwalają na analizę efektywności ekonomicznej leczenia (cost-effectiveness analyses)

Dane dotyczące stosowania się do wymogów leczenia

- Niestosowanie się do wymogów leczenia (non-compliance) oznacza odstępstwo od sposobu leczenia przewidzianego w protokole
- Może mieć duży wpływ na ocenę skuteczności leczenia
- Często trudne do oceny
  - Jak sprawdzić, czy chory wziął tabletkę w domu?

Dane dotyczące toksyczności

- Monitorowanie toksyczności leczenia i efektów niepożądanych jest obligatoryjne
- Najczęściej występujące rodzaje toksyczności powinny być określone w próbach fazy I i II, ale rzadsze mogą ujawniać się w próbach fazy III
- Ocena wymaga użycia jednorodnych kryteriów i terminów (MedDRA, NCI-CTC)
- Ważne jest obserwacja krótko- i długoterminowa

Rodzaje kryteriów oceny skuteczności leczenia

- Binarne (odpowiedź na leczenie, wyleczenie, ...)
- Kategoryzowane (odpowiedź guza na leczenie: pełna, częściowa, stabilizacja, progresja)
- Ciągłe (parametry krwi, waga, ...)
- Czas do wstąpienia zdarzenia (czas przeżycia)
- Pomiary powtarzane (repeated measurements)
- Wielowymiarowe/wielokrotne (jakość życia, kryteria złożone)

Miary efektu leczenia

- Miara efektu leczenia, tzn. numeryczne ujęcie różnicy w kryterium oceny skuteczności leczenia między randomizowanymi grupami, zależy od typu kryterium.

- Ogólnie, efekt leczenia może być mierzony na skali względnej lub bezwzględnej

Binarne kryteria oceny skuteczności: różnica ryzyka

- π_C = Prob(response | control)
- π_E = Prob(response | experimental)

- Różnica ryzyka (risk difference): $RD = \pi_E - \pi_C$
- Problem: porównanie efektów ubocznych 2 leków

  a) 0.410 vs. 0.401

  b) 0.010 vs. 0.001

  - *RD=0.009*, ale dla b) ryzyko dla jednego z leków jest 10 razy wyższe niż dla drugiego !

Binarne kryteria oceny skuteczności: ryzyko względne

- *RR* jest ilorazem dwóch wartości ryzyka (prawdopodobieństwa):

$$RR = \frac{\text{Risk for E}}{\text{Risk for C}} = \frac{\pi_E}{\pi_C}$$

- Dla przykładu z efektami ubocznmi, *RR* wynosi 0.401/0.400=1.0025 and 0.01/0.001=10

- Jeśli *RR*=1, mamy *RD*=0, tzn. brak różnicy.

Binarne kryteria oceny skuteczności: iloraz szans

- Iloraz szans (odds ratio) :

$$OR = \frac{\text{Odds of response for E}}{\text{Odds of response for C}}$$

$$= \frac{\dfrac{\pi_E}{1-\pi_E}}{\dfrac{\pi_C}{1-\pi_C}} = \frac{\pi_E(1-\pi_C)}{\pi_C(1-\pi_E)}$$

Iloraz szans, ryzyko względne

- Zakres *OR i RR* to (0,+∞)
- *OR* = 1 lub *RR* = 1⇒ brak różnicy (związku)
- *OR* ≈ *RR* dla rzadkich zdarzeń
- *OR* i *RR* przyjmują wartości w tym samym kierunku (>1 lub <1), ale *OR* jest zawsze dalej od 1 !

Miary efektu leczenia

- Bezwzględna redukcja ryzyka (Absolute Risk Reduction, *ARR*)
  ARR = *RD*
- Względna redukcja ryzyka (Relative Risk Reduction, *RRR*)
  RRR = 1 – *RR*
- Względna redukcja szans (Relative Odds Reduction, *ORR*)
  ORR = 1 – *OR*

- *ORR* ≥ *RRR* ≥ *ARR*

Kryteria kategoryzowane

- Kłopot z uzyskaniem sumarycznej miary efektu
- *OR* oparte na modelu
  - model „proporcjonalnych szans” (proportional odds) dla danych porządkowych (ordinal data)

$$OR = OR_j = \frac{\frac{\sum_{k=1}^{j} \pi_k^E}{1-\sum_{k=1}^{j} \pi_k^E}}{\frac{\sum_{k=1}^{j} \pi_k^C}{1-\sum_{k=1}^{j} \pi_k^C}} \quad \forall j$$

Czas do wystąpienia zdarzenia

- *S(t)* = prawdopodobieństwo przeżycia czasu *t*

- Funkcja hazardu: „chwilowa intensywność zdarzeń”
  - $\lambda(t)$, liczba zdarzeń na jednostkę czasu
  - „Prędkość” występowania zdarzeń
  - $\lambda(t)\Delta t = P(\text{ event in } (t, t+\Delta t], \text{ given no event until time } t\ )$

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

Czas do wystąpienia zdarzenia

- Iloraz hazardu (hazard ratio):

$$HR(t) = \frac{\lambda_{\mathrm{E}}(t)}{\lambda_{\mathrm{C}}(t)}$$

  - Często zakłada się *HR(t)* = *HR* = const.
  - Czyli model proporcjonalnych hazardów

Pomiary powtarzane

- Miara efektu leczenia na podstawie modelu
  - $E(Y_{ij}| C) = \mu + \varphi \times t_{ij}$
  - $E(Y_{ij}| E) = \mu + \varphi \times t_{ij} + \theta$

  lub
  - $E(Y_{ij}| E) = \mu + \varphi \times t_{ij} + \theta + \psi \times t_{ij}$

Kryteria wielowymiarowe/wielokrotne

- Na ogół kłopot ze zdefiniowaniem sumarycznej miary efektu leczenia.
- Poszczególne wymiary często analizowane osobno.
- Poprawka na wielokrotne testowanie
  - Bonferroni
  - Sumaryczny test

Testowanie hipotez

- Próby fazy III są porównawcze
- Wnioskowanie na ogół na podstawie testów istotności statystycznej

- Dwie hipotezy:
  - zerowa (np. „brak różnicy w skuteczności leczenia”)
  - alternatywna (np. „różnica w skuteczności leczenia”)

- p-value = P(Test statistic > observed | H_0) < α
  - α = 0.05

Przykład testowania hipotez

- Ciągłe kryterium oceny skuteczności: zmiana DBP
- Rozkład zmian DBP zgodny z normalnym
  - rozkład różnic średnich również
- Hipoteza zerowa: brak różnicy dla E i C
- Obliczenia
  - Dla E: średnia różnica = -13.4 mmHg (SEM = 2.0)
  - Dla C: średnia różnica = -9.4 mmHg (SEM = 2.0)
  - Statystyka testowa Z = {(-13.4) - (-9.4)}/2.0 = 2.0
  - p-value = P(|N(0,1)| > 2.0) = 0.034 < 0.05
  - Wniosek : E jest statystycznie istotnie skuteczniejsze w redukcji DBP niż C

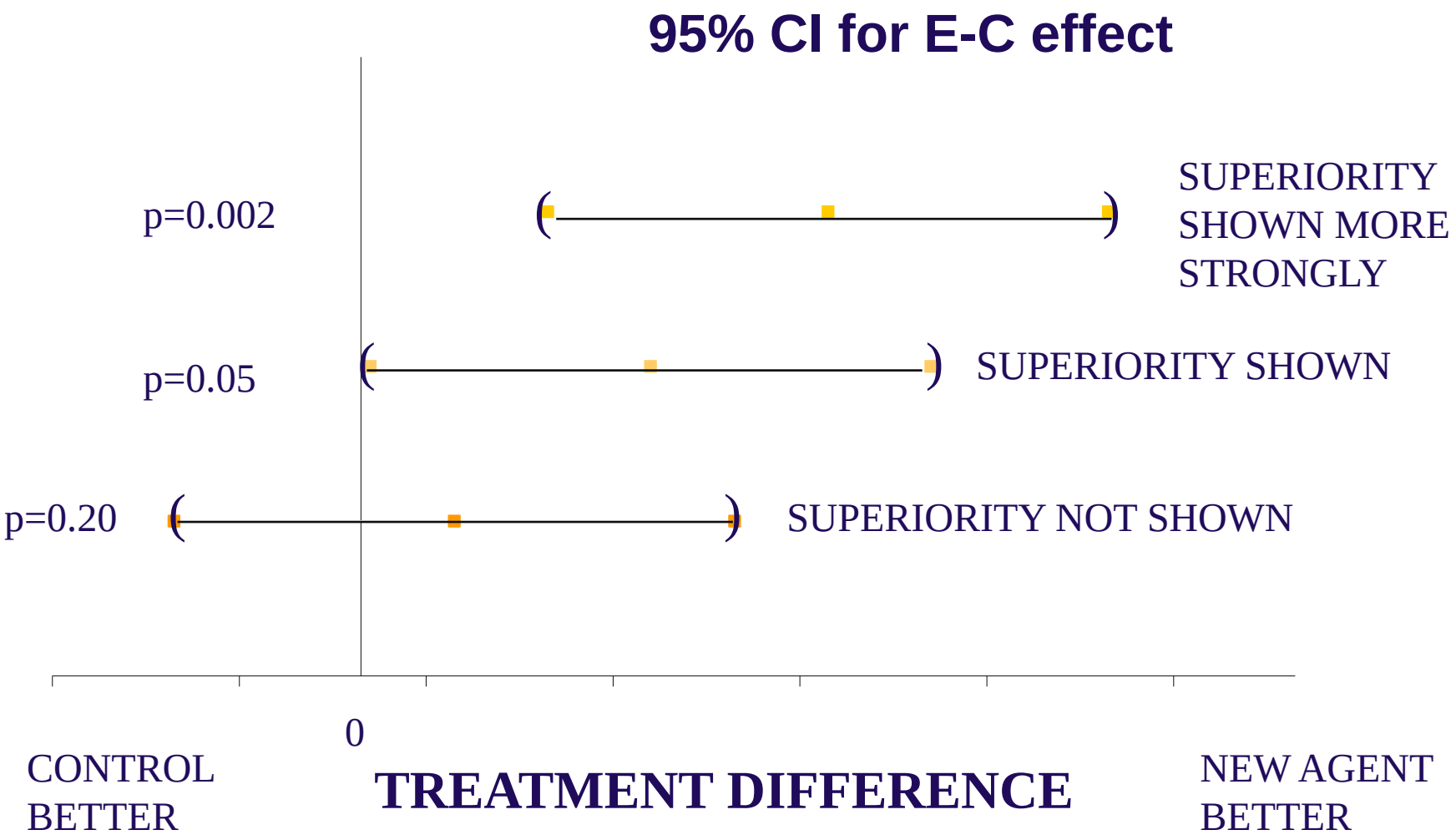
Przedziały ufności

- Oprócz (lub wręcz zamiast) poziomu krytycznego testu *p*, konieczne jest podanie przedziału ufności dla miary efektu leczenia.
- (1-α)100% CI (często) odpowiada testowi istotności dla poziomu istotności α …
- … ale podaje dodatkową informację
  - oszacowanie punktowe wielkości efektu leczenia
  - precyzję oszacowania

Różnica/korzyść

- *Różnica (difference)*: wyniki dla E i C się różnią
  - H_0: $\mu_E - \mu_C = 0$ (brak różnicy) H_A: $\mu_E - \mu_C \neq 0$ (różnica)
- *Korzyść (benefit)*: wyniki dla E są lepsze niż dla C
  - H_0: $\mu_E - \mu_C \leq 0$ (E gorsze) H_A: $\mu_E - \mu_C > 0$ (E lepsze)

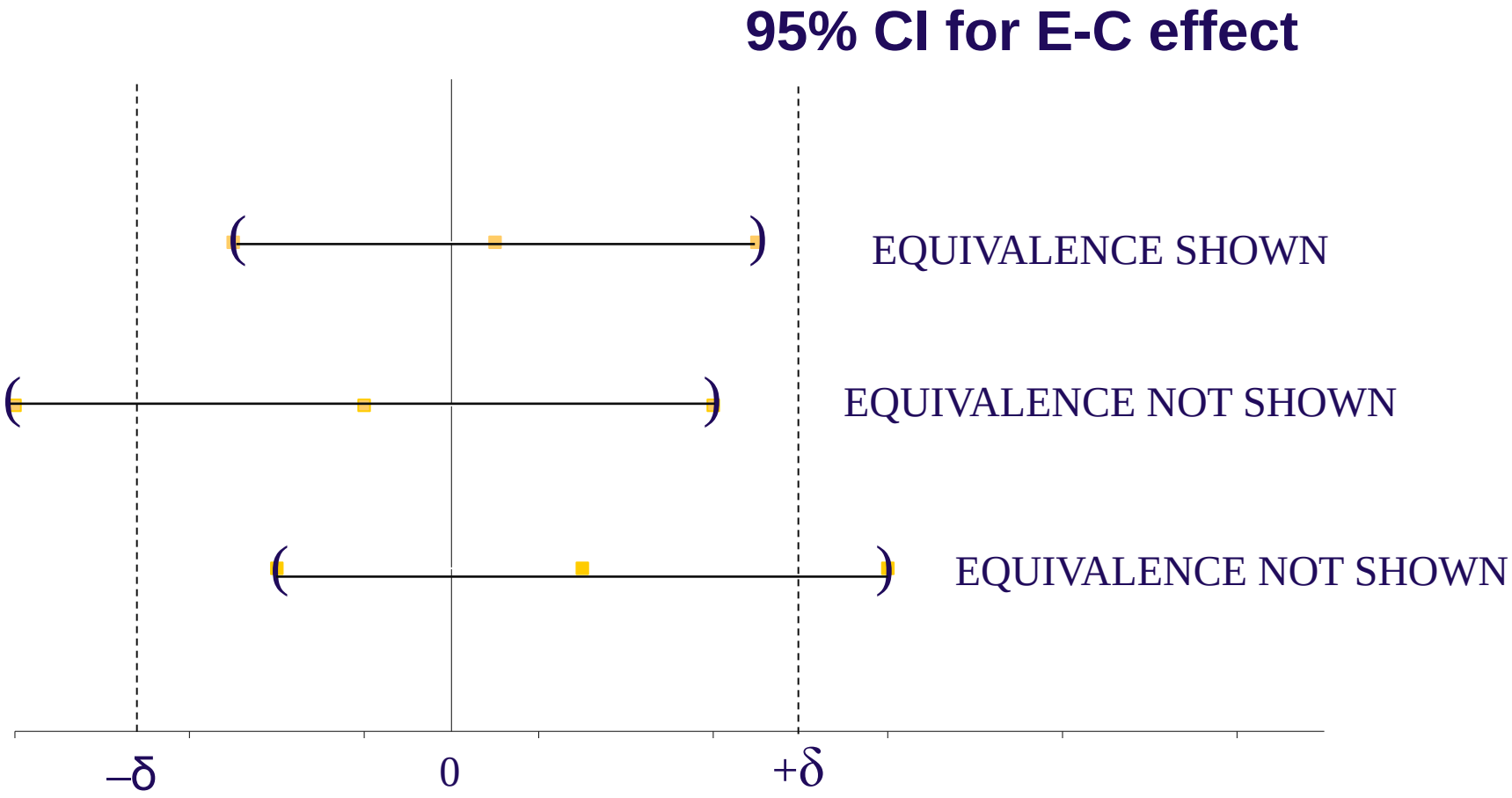
Nadrzędność



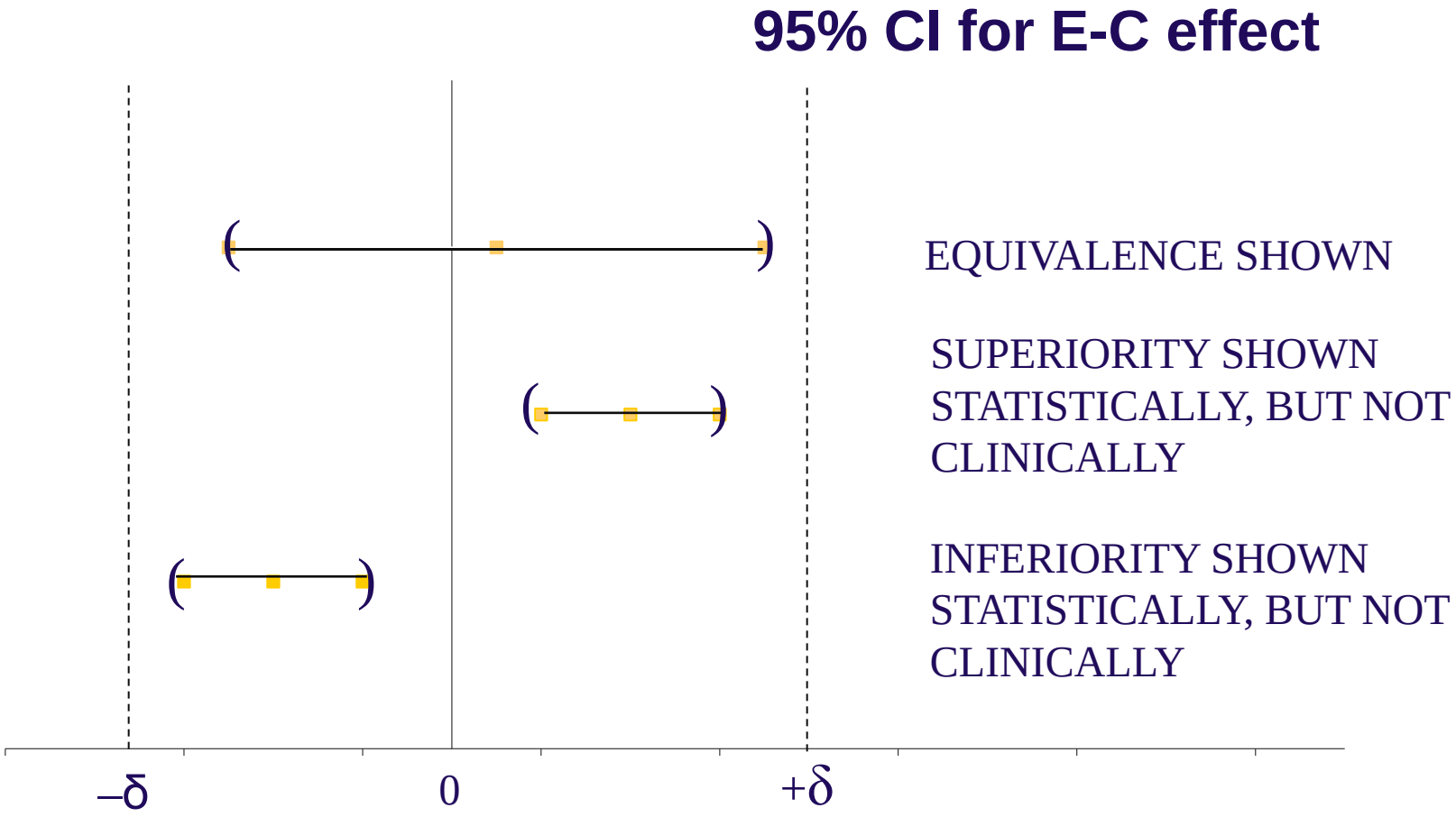
Równoważność/Nie-podrzędność

- *Equivalence*: wyniki dla E i C różnią się, ale wielkością nieistotną klinicznie
  - H_0: $|\mu_E - \mu_C| \geq \delta$ (nie równoważne) H_A: $|\mu_E - \mu_C| < \delta$ (równoważne)
- *Non-inferiority*: wyniki dla E nie są klinicznie istotnie gorsze niż dla C
  - H_0: $\mu_E - \mu_C \leq -\delta$ (E gorsze) H_A: $\mu_E - \mu_C > -\delta$ (E nie gorsze)
- δ ujmuje klinicznie istotną róznicę
  - δ zwykle mniejsza niż różnica w próbie nadrzędności

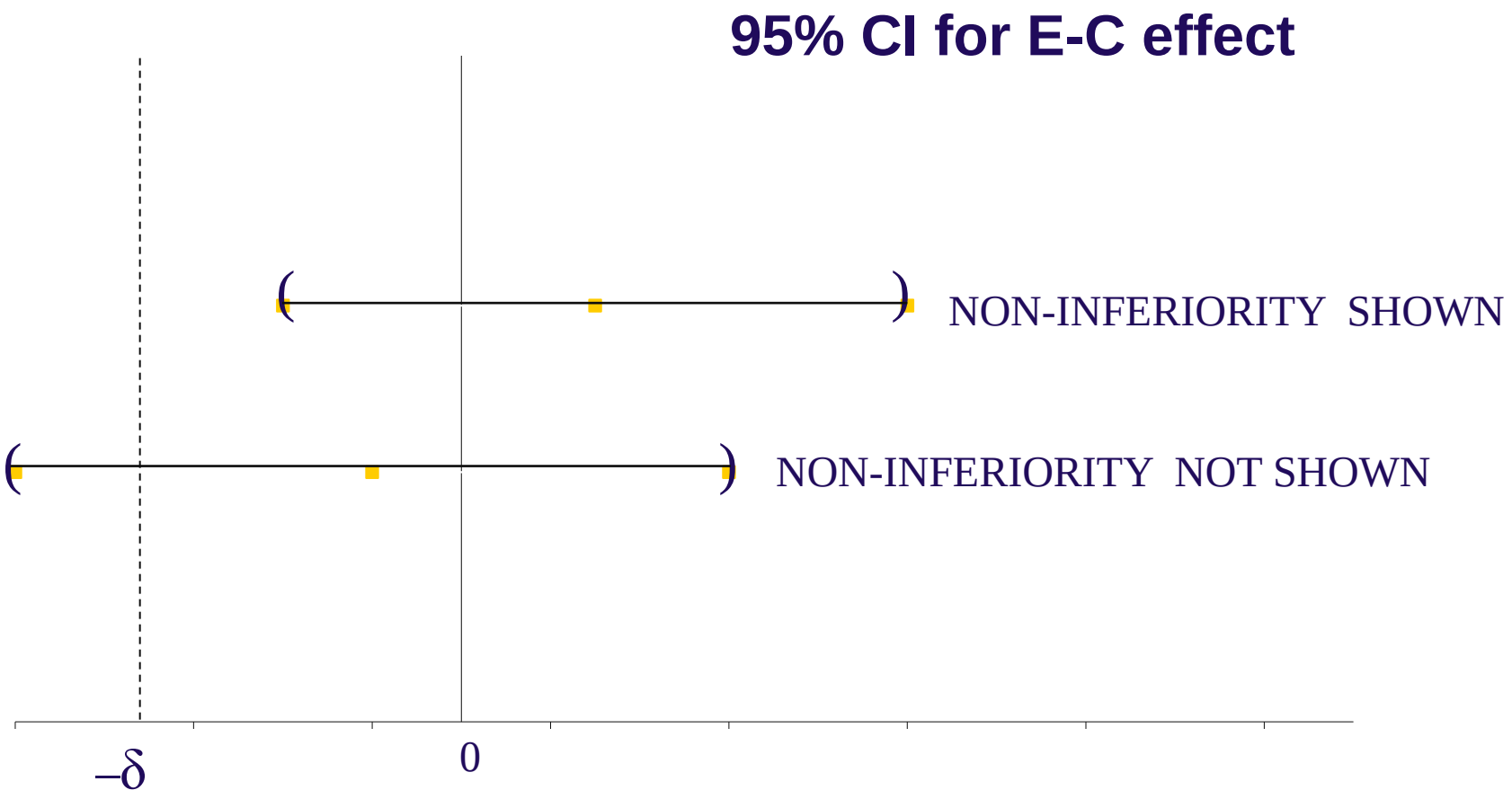
Równoważność



Równoważność założona
Różnica zaobserwowana !



Nie-podrzędność



Po co próby nie-podrzędności?

- Coraz trudniej wykazać nadrzędność w próbach klinicznych dla pewnych chorób
  - np., istnieją **bardzo skuteczne** antybiotyki
- Nie-podrzędność może być interesująca, jeśli nowa metoda leczenia oferuje dodatkowe korzyści (np. jest bezpieczniejsza, łatwiejsza do zastosowania, tańsza)
  - np. chemioterapia z porównywalną skutecznością lecz mniej toksyczna, lub w formie doustnej (capecitabine) zamiast dożylnej (5-fluorouracil)

Błędy wnioskowania

	Truth (unknown)	
Test result (known)	Null hypothesis	Alternative
Significant ($p < \alpha$) (reject the null)	**TYPE I ERROR** ***(α)***	**OK** ***(1 - β)***
Non-significant ($p > \alpha$) (accept the null)	**OK** ***(1 - α)***	**TYPE II ERROR** ***(β)***

- *Wynik testu może zawsze być błędny !*
- Możemy tylko kotrolować p-stwo błędu

Konsekwencje błędów

- Błąd I rodzaju: błędne odrzucenie hipotezy zerowej
  - Błędne uznanie nadrzędności/równoważności
    - Problem w próbach nadrzędności, jeśli E wiąże się z wyższą toksycznością
    - Lek może być wycofany np. po próbach fazy IV
- Błąd II rodzaju: błędne przyjęcie hipotezy zerowej
  - Błędne przyjęcie braku efektu lub nierównoważności
    - Superiority trial: skuteczne leczenie może zostać porzucone
    - Equivalence trial: mniej toksyczny lek może zostać porzucony
    - Konsekwencje finansowe

Kontrola prawdopodobieństwa błędu I rodzaju

- Poprzez przyjęcie (i kontrolę) odpowiedniego poziomu istotności testu α
  - Problem: wielokrotne testowanie

Kontrola prawdopodobieństwa błędu II rodzaju

- Prawdopodobieństwo (β) zależy od
  - poziomu istotności (α);
  - założonej różnicy w skuteczności leczenia Δ;
  - zmienność oszcowania miary efektu leczenia;
  - **liczebności próbki**.
- 1- β = moc
  - P-stwo odrzucenia hipotezy zerowej jeśli jest fałszywa = P-stwo „wykrycia” hipotezy alternatywnej

Próby nie-podrzędności: problemy

- Wybór marginesu tolerancji klinicznej (non-inferiority margin, δ)
- Brak „wewnętrznego” dowodu na wrażliwość próby (assay sensitivity)
- Brak konerwatywnej strategii analizy wyników
- „Zaślepianie” nie oferuje pełnej ochrony przed obciążeniem

Wybór marginesu tolerancji klinicznej – ICH E9

- Wiele metod wyboru ! (w efekcie, trudności w uzyskaniu zgody co do wartości δ)

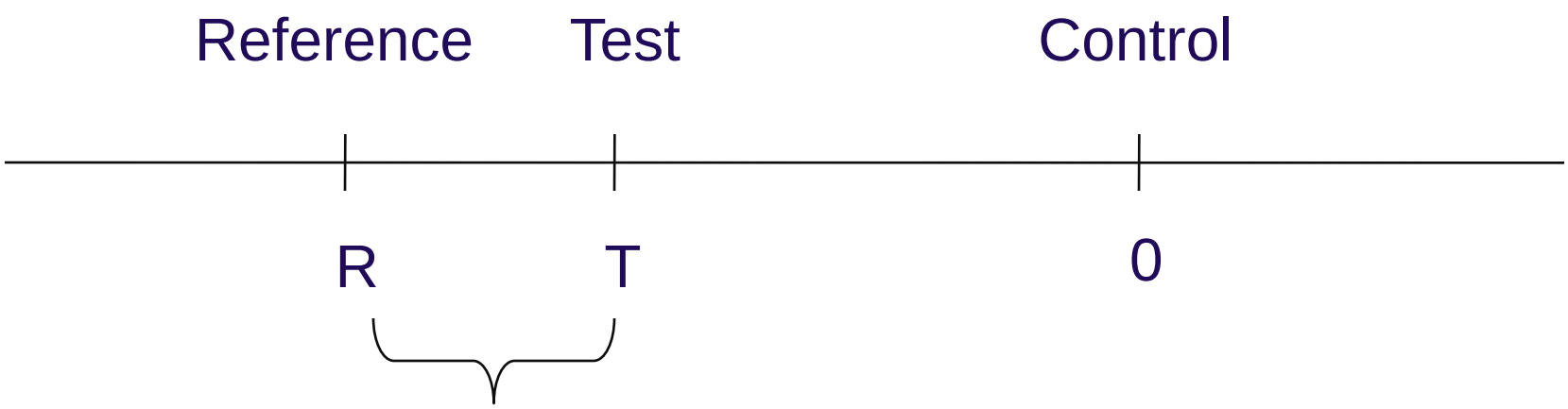
- " …largest difference that is judged to be clinically acceptable, and should be smaller than differences observed in superiority trials of the active comparator..."

  (np. δ = ½ różnicy uważanej za sensowną klinicznie dla prób nadrzędności)

Wybór marginesu tolerancji klinicznej – EMA

- Najlepszy układ doświadczalny z trzema grupami: leczoną (test), odniesienia (reference), i kontrolną. Celem jest wykazanie:
  - Nadrzędności dla testowej vs. kontrolnej
  - Nie-podrzędności dla testowej i odniesienia
- Może nie być możliwy jeśli użycie grupy kontrolnej (np. placebo) jest nieetyczne; wówczas próba porównuje grupę testową z odniesienia, a porównanie odniesienia z kontorlną wnioskowane na podstawie danych historycznych.

Wybór marginesu tolerancji klinicznej – EMA



R – T = observed difference between Reference and Test

R = historical effect of Reference

T = indirect effect of Test

Wybór marginesu tolerancji klinicznej – FDA

- Przyjmijmy, że dla grupy odniesienia (R) wykazano, że HR = 0.76 (95% C.I. 0.66 – 0.89)
- M1 jest najmniejszym efektem R w przeciwnym kierunku. W przykładzie: M1 = 1/0.89 = 1.12
- M2 jest częścią (np. 50%) oszacowania efektu R w przeciwnym kierunku. To tzw. „zachowana część efektu („percentage of effect retained”). W przykładzie

  log HR = log (0.76) = -.274

  ½ log HR = -.137

  M2 = exp(.137) = 1.15

Wrażliwość próby?

- *Assay Sensitivity (AS)*: zdolność do wykrycia różnicy, jeśli ta istnieje.
- Jeśli próba wykazuje nadrzędność, to dowód na AS.
- Jeśli próba wykazuje nie-podrzędność, nie mamy dowodu na AS:
  - próba była OK, nie-podrzędność jest rzeczywista, lub
  - nie wykazaliśmy różnicy, bo próba była źle wykonana
- AS musi być wywiedziona z innych informacji:
  - historycznych (podobne próby wykazywały różnice)
  - jakośći wykonania i grupy kontrolnej

„Zaślepianie”

- W próbach nadrzędności, „zaślepienie” uniemożliwia lekarzowi faworyzowanie w systematyczny sposób jednej z metod leczenia.
  - Musiałby znać metodę przydzieloną choremu.

- W próbach nie-podrzędności, „zaślepiony” lekarz mógłby oceniać wyniki leczenia tak samo dla wszystkich chorych.
  - Niwelując różnicę w wynikach leczenia.

Protokół próby nie-podrzędności

- Musi zawierać
  - uzasadnienie wyboru i dawek dla grupy kontrolnej
  - margines tolerancji klinicznej
  - uzasadnienie wrażliwości próby (AS)
  - opis metody analizy statystycznej nie-podrzędności w oparciu o przedziały ufności
  - opis populacji użytej do analizy: użycie intencji leczenia nie jest konserwatywną strategią (ICH E9)

Wykazanie nadrzędności - równoważności – nie-podrzędności

- Superiority: przedział ufności (np. 95%) dla różnicy E-C wyklucza 0 lub różnice na korzyść C
- Equivalence: przedział ufności leży całkowicie w obszarze równoważności klinicznej
- Non-inferiority: dolna granica przedziału ufności jest większa od marginesu tolerancji klinicznej
- Możliwe jest wykazanie nadrzędności/równoważności w tej samej próbie
  - δ musi być ustalona *a priori*, wykonanie i liczebność próby muszą być OK.

Ryzyko i koszt prób równoważności/nie-podrzędności

- Próby te często wymagają większej liczebności niż próby nadrzędności
- Wybór marginesu tolerancji klinicznej δ nie jest łatwy
- Problemy z wykazaniem wrażliwości próby (AS)
- „Przełączenie” z nadrzędności na nie-podrzędność niemożliwe, jeśli nie zaplanowane *a priori*

Wielokrotne testowanie

- „Idealna” próba kliniczna ma tylko jeden cel i, w efekcie, testuje tylko jedną hipotezę.
- P-stwo błędu I rodzaju dla całej próby rośniw błyskawicznie ze wzrostem liczby testów dla wielu kryteriów oceny skuteczności mierzonych w wielu punktach czasowych, i/lub wielu metod leczenia, i/lub analizy wielu podgrup chorych.
- Podstawową ideą jest kontrola p-stwa błędu I rodzaju na poziomie $\alpha = 0.05$, niezależnie od liczby testów.

Wzrost p-stwa błędu I rodzaju

- P-stwo błędu dla całego doświadczenia (experiment-wise Type I error) to p-stwo błędnego odrzucenia przynajmniej jednej hipotezy zerowej

- Experiment-wise Type I error rate = $1 - (1 - \alpha)^K$

  α = comparison-wise significance level,

  K = number of comparisons performed

Experiment-wise Type I Error

# of Comparisons (K)	Experimentwise Type I Error
1	0.05
2	0.0975
3	0.1426
5	0.2262
10	0.4013
14	0.5123

Wiele kryteriów oceny skuteczności leczenia

- Jedno pierwszoplanowe (primary) kryterium
- K pierwszoplanowych (co-primary) kryteriów
  - Bonferroni
    - K=2: 0.05 experiment-wise Type I error
      could take 0.025 for the two endpoints or
      0.04 for the 1st endpoint + 0.01 for the 2nd endpoint
  - Multiple Comparison Strategies
    - Sequentially rejective procedure, hierarchical testing, …
  - Overall Test Statistic
    - E.g., weighted averages (O'Brien 1984, Pocock et al. 1987, …)
  - Summary Measures

Drugoplanowe kryteria oceny skuteczności leczenia

- Rekomendacje FDA: kryteria drugoplanowe brane pod uwagę tylko jeśli
  - dla pierwszoplanowego uzyskano wynik statystycznie istotny oraz ....
  - ... p-stwo błędu I rodzaju dla kryteriów drugoplanowych jest kontrolowane na tym samym poziomie istotności co kryterium pierwszoplanowe.

Istotność statystyczna a kliniczna

- Dla odpowiednio dużej liczebności próbki, nawet mała różnica może dać statystycznie istotny wynik.
- To nie oznacza, że różnica ta jest klinicznie znacząca
  A. redukcja ryzyka zgonu z 60% do 40% (RR=33%)
  B. redukcja ryzyka zgonu z 6% do 4% (RR=33%)
  - Jeśli choroba jest rzadka, B nie jest interesujące z punktu widzenia zdrowia publicznego (choć na pewno dla pojedyńczego chorego).
  - Ale jeśli choroba jest powszechna, B będzie interesujące zarówno z punktu widzenia zdrowia publicznego, jak i pojedyńczych chorych.