Assessing treatment effect heterogeneity in observational studies

Epid 765: Intro to Pharmacoepidemiology

Slides adapted from Michele Jonsson Funk, PhD

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

© 2019 by Til Stürmer. All rights reserved.

1

Outline

- Motivation
- Common approaches to assessing HTE
- Cautionary tale regarding use of interaction terms
- Absolute vs relative measures
- Conclusions

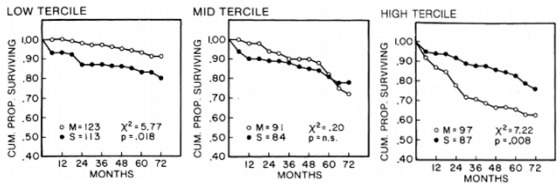
2

The search for treatment effect heterogeneity

- Identify patients in whom the expected benefit and/or harm of treatment are meaningfully different than in the average patient
- Broadest: indication/contraindication
- Within these subgroups often defined by:
  - Sex, age, race/ethnicity
  - Concurrent medications, comorbid conditions, stage of disease, background risk of the outcome
  - Genetic variation
- Increased interest in machine learning/AI: precision medicine, personalized medicine

3

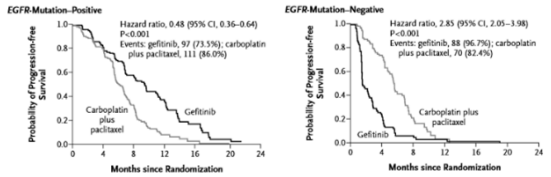
Qualitative Heterogeneity: Medical vs. Surgical Management of Coronary Artery Disease by Tertiles of Predicted Risk of Mortality



Detre et al, Circulation 1981

4

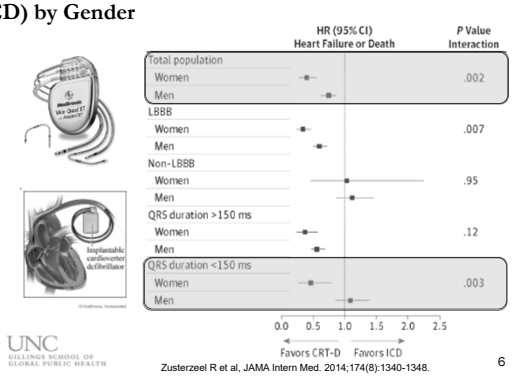
Qualitative Heterogeneity: Getifinib vs. Carboplatin plus Paclitaxel in Pts. With Advanced Lung Cancer by Genetic Factor



Mok TS et al, NEJM 2009

5

Quantitative Heterogeneity: Cardiac Resynchronization Therapy (CRT) vs. Implantable Cardioverter Defibrillator (ICD) by Gender



Zusterzeel R et al, JAMA Intern Med. 2014;174(8):1340-1348.

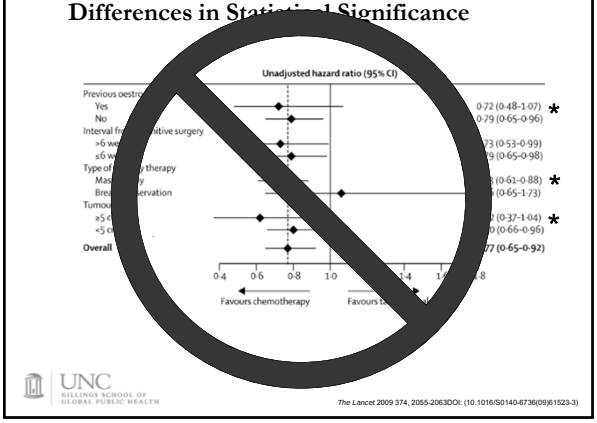
6

Common Approaches to Assessing HTE

7

Figure 3

Differences in Statistical Significance

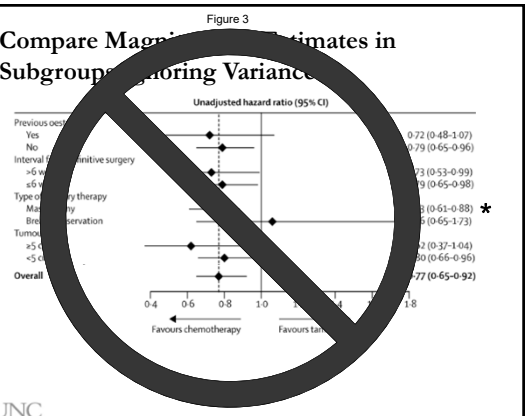


The Lancet 2009 374, 2055-2063DOI: (10.1016/S0140-6736(09)61523-3)

8

Figure 3

Compare Magni timates in Subgroups oring Variance



The Lancet 2009 374, 2055-2063DOI: (10.1016/S0140-6736(09)61523-3)

9

Evaluating HTE With Interaction Terms

- Primary analysis
  - Main effects only
- Considering possible heterogeneity
  - Interaction term between main effect of subgroup (e.g. gender) and the treatment (e.g. statin)
  - If p≥0.05 for interaction term, no evidence of subgroup difference
  - If p<0.05 for interaction term, then subgroups *are* different and no meaningful summary effect estimate

10

10

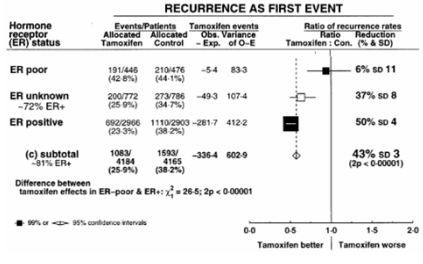
Errors in Assessment of HTE

- False Positive (Type I error)
  - Finding evidence of heterogeneity when effects are homogeneous
  - Historically of greatest concern
  - Tests of interaction have expected Type I error rate (5% given alpha=0.05) (Brooks et al, J Clin Epi 2004)
- False Negative (Type II error)
  - Failing to detect heterogeneity that is present

11

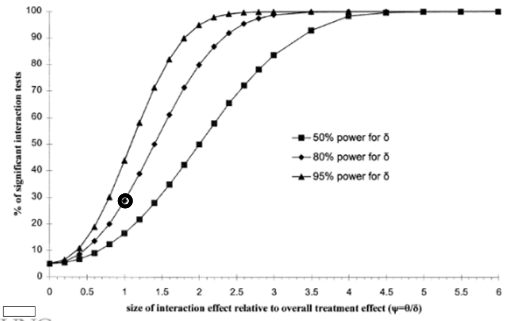
11

Tamoxifen Effect in ER 'Poor' vs ER Positive



12

Power for Tests of Interaction



Brookes et al, J Clin Epi, 2004

13

13

Statistical Significance of Interaction Terms

- A function of sample size and effect size
  - Small differences significant in large samples
  - Large differences non-significant in small samples
- Low power relative to overall study
  - Report all estimates with CIs to contribute to the eventual meta-analysis

14

14

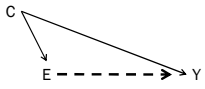
Interaction Terms and Potential Residual Confounding

15

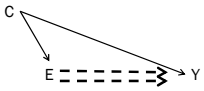
15

DAGs for primary analysis and interaction by Z

Primary analysis
Estimate effect of E on Y, adjusted for C



HTE analysis
Interaction term to allow the effect of E on Y to differ by levels of Z (subgroup, a potential effect modifier)

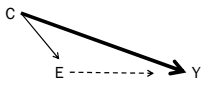


16

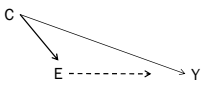
16

Effects of Confounder May Differ by Z

Among Z=0
C is a strong risk factor for Y



Among Z=1
C is a weak risk factor for Y

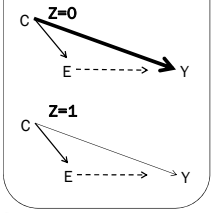


17

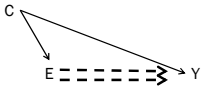
17

Are Subgroup Effect Estimates Adequately Adjusted for C?

True effects of C on Y within strata of Z



Subgroup estimates based on outcome regression model with Z*E interaction term



18

18

Illustration with simulated data

	Overall	Z=0	Z=1
Crude	3.07 (3.00, 3.14)	3.28 (3.17, 3.39)	2.89 (2.80, 2.99)
Adjusted for Z	0.99 (0.95, 1.04)		
Z*E interaction term		1.16 (1.10, 1.23)	0.86 (0.81, 0.91)
Stratified by Z		1.01 (0.95, 1.07)	0.98 (0.92, 1.04)
† All adjusted for main effects of baseline covariates			

© M Jonsson Funk, 2017

19

Recommendations

- Check covariate balance within subgroups
- Check stability of effect estimates between interaction-based assessment vs subgroup specific models
  - Examine parameter estimates from outcome regression or PS models for clues
- Not limited to non-experimental studies
  - Small subgroups in experimental studies may not be exchangeable and therefore confounded

20

Relative vs Absolute Effects

21

Conventional Wisdom: RD is Inherently More Heterogeneous Than RR (OR)

TABLE 1. Results of RD and OR Homogeneity Tests in Three Meta-meta-analyses

	Engels et al.[15]	Sterne and Egger[16]	Deeks[17]
Number of meta-analyses	125	78	551
Homogeneity test alpha (%)	5	5	10
RD homogeneity rejected (%)	47	35	34
OR homogeneity rejected (%)	−35	−29	−22
Difference (%)	12	6	12
Only RD homogeneity rejected (%)	14	3[a]	13
Only OR homogeneity rejected (%)	2	8[a]	1
Both rejected (%)	33	27[a]	21
Neither rejected (%)	+50	+63[a]	+65
Overall test concordance (%)	83	90[a]	86

Poole, Shrier, VanderWeele, Epidemiology, 2015.

22

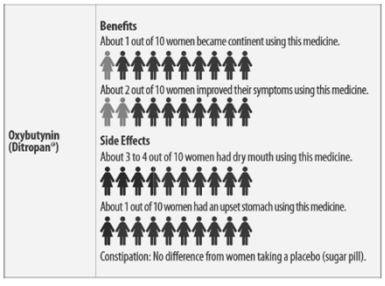
Challenging Conventional Wisdom

- Results from meta-meta-analysis don't consider differences in the power of tests
- Scenarios can be constructed in which the power of the test of homogeneity is greater for RD or for OR when both are heterogeneous
- Unclear how to measure the magnitude of heterogeneity on 3 scales (RR, OR, and RD)
- Difficult to claim that RD is inherently more heterogeneous given the lack of evidence

Poole, Shrier, VanderWeele, Epidemiology, 2015.

23

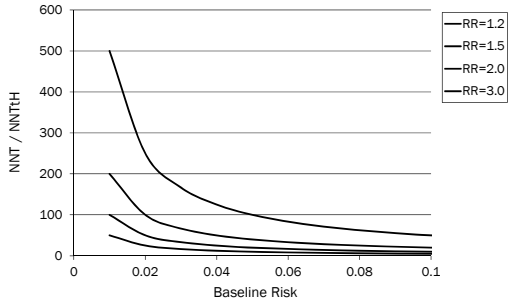
Reporting Absolute Effects to Address Benefit/Harm and Support Clinical Decision Making



Nonsurgical Treatments for Urinary Incontinence in Adult Women: Diagnosis and Comparative Effectiveness, April 2012.

24

Relative vs. Absolute Effects

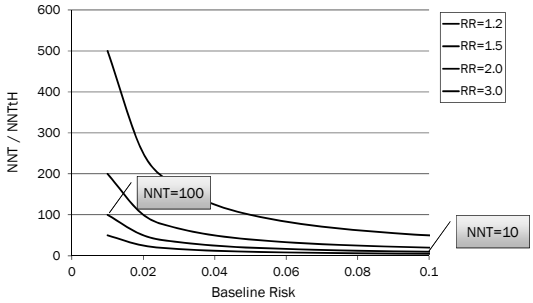


UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

25

25

Relative vs. Absolute Effects

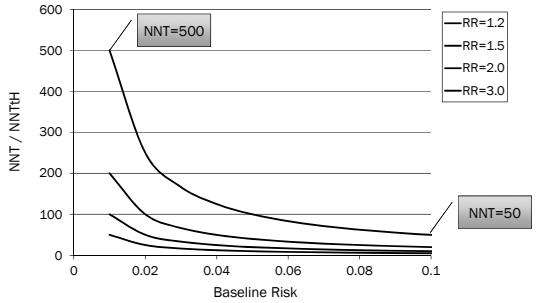


UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

26

26

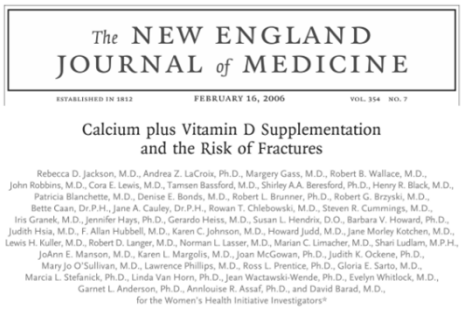
Relative vs. Absolute Effects



UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

27

27



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 16, 2006 VOL. 354 NO. 7

Calcium plus Vitamin D Supplementation and the Risk of Fractures

Rebecca D. Jackson, M.D., Andrea Z. LaCroix, Ph.D., Margery Gass, M.D., Robert B. Wallace, M.D., John Robbins, M.D., Cora E. Lewis, M.D., Tamsen Bassford, M.D., Shirley A.A. Beresford, Ph.D., Henry R. Black, M.D., Patricia Blanchette, M.D., Denise E. Bonds, M.D., Robert L. Brunner, Ph.D., Robert G. Brzyski, M.D., Bette Caan, Dr.P.H., Jane A. Cauley, Dr.P.H., Rowan T. Chlebowski, M.D., Steven R. Cummings, M.D., Iris Granek, M.D., Jennifer Hays, Ph.D., Gerardo Heiss, M.D., Susan L. Hendrix, D.O., Barbara V. Howard, Ph.D., Judith Hsia, M.D., F. Allan Hubbell, M.D., Karen C. Johnson, M.D., Howard Judd, M.D., Jane Morley Kotchen, M.D., Lewis H. Kuller, M.D., Robert D. Langer, M.D., Norman L. Lasser, M.D., Marian C. Limacher, M.D., Shari Ludlam, M.P.H., JoAnn E. Manson, M.D., Karen L. Margolis, M.D., Joan McGowan, Ph.D., Judith K. Ockene, Ph.D., Mary Jo O'Sullivan, M.D., Lawrence Phillips, M.D., Ross L. Prentice, Ph.D., Gloria E. Sarto, M.D., Marcia L. Stefanick, Ph.D., Linda Van Horn, Ph.D., Jean Wactawski-Wende, Ph.D., Evelyn Whitlock, M.D., Garnet L. Anderson, Ph.D., Annlouise R. Assaf, Ph.D., and David Barad, M.D., for the Women's Health Initiative Investigators*

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

28

28

Hip fracture, analysis of HTE by age

Group	Annualized risk (%)		HR (95% CI)
	Calcium+ Vit D	Placebo	
Overall	0.14	0.16	0.88 (0.72-1.08)
Age (yrs)			
50-59	0.06	0.03	2.17 (1.13-4.18)
60-69	0.09	0.13	0.74 (0.52-1.06)
70-79	0.44	0.54	0.82 (0.62-1.08)

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Jackson et al, NEJM 2006.

29

29

Testing HTE of the Risk Differences

Cochran's Q
Likelihood ratio (LR)
RERI (derived from RR)
Wald

See Schmidt et al, J Clin Epi 2014, Supplemental material, Appendix 1

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

30

Estimating Risk Differences

- Uniform follow-up
  - Log-binomial or Poisson (Spiegelman AJE 2005)
- Censored outcomes/time-to-event
  - Difference in the cumulative incidence at time t
  - Survival curves adjusted via weighting or matching
  - Robust variance or bootstrapping for CIs

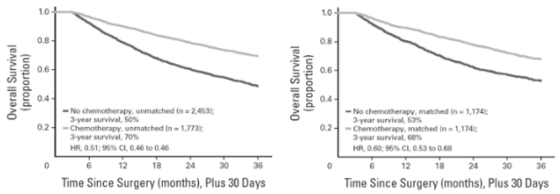
UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

31

31

Adjusting survival curves

- PS Matching or Weighting



UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Sanoff et al, J Clin Onc 2012

32

32

Conclusions

- Important subgroup differences exist
- Apply statistical criteria with care
- Estimate and report overall and subgroup effects
- Beware residual confounding
- Evaluate HTE on the absolute scale in addition to the relative scale

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

33

33