Gene+environment risk models: whys and hows

Peter Kraft

Professor of Epidemiology and Biostatistics

Harvard T.H. Chan School of Public Health

- Genes are important
- The environment is important
- Risk models that include both can be useful

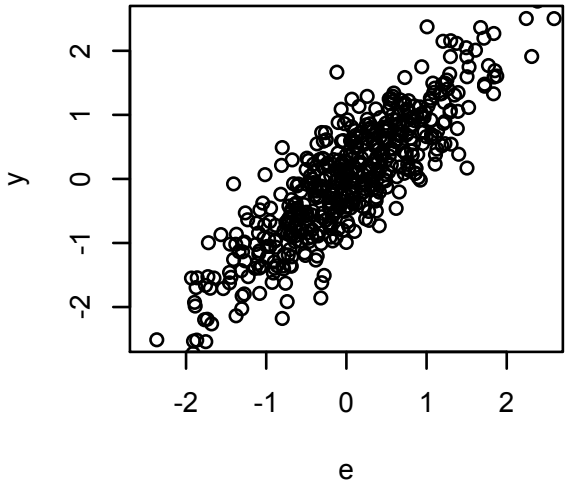
Risk models using E

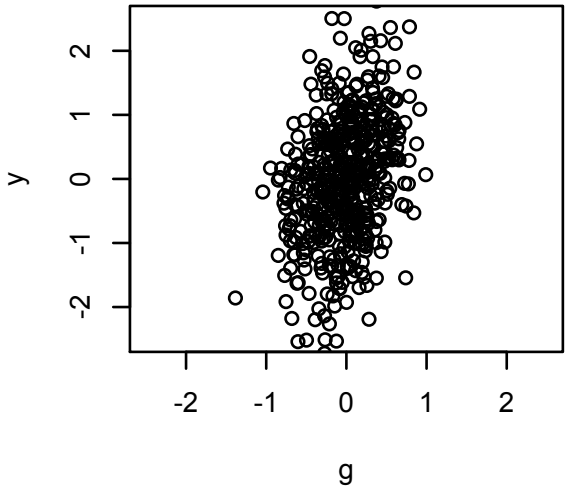
- Framingham risk model (CVD)
- Breast cancer risk assessment tool
- Chronic and acute risk of VTE

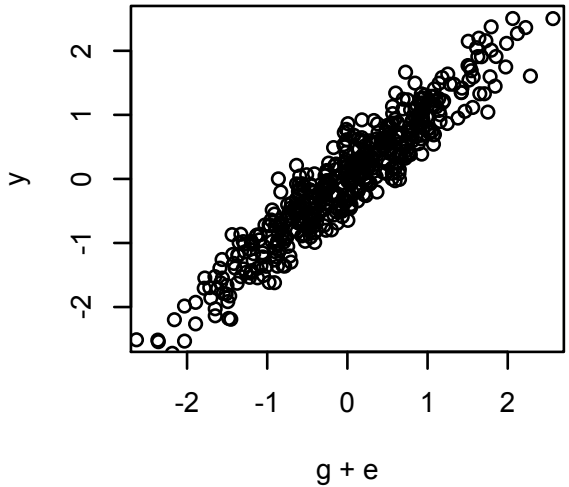
Risk models using E

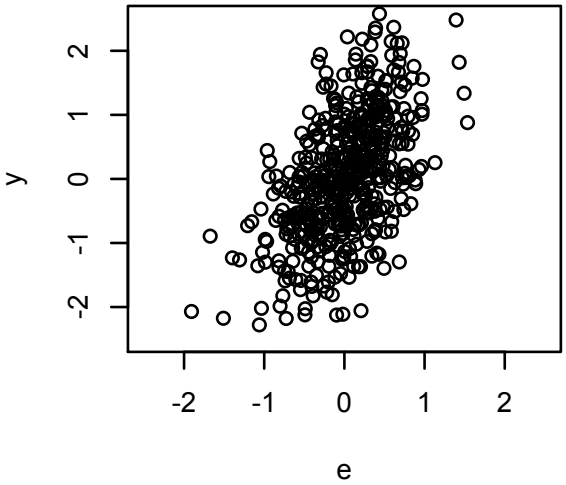
- Framingham risk model (CVD)
- Breast cancer risk assessment tool
- Chronic and acute risk of VTE
- Prostate cancer

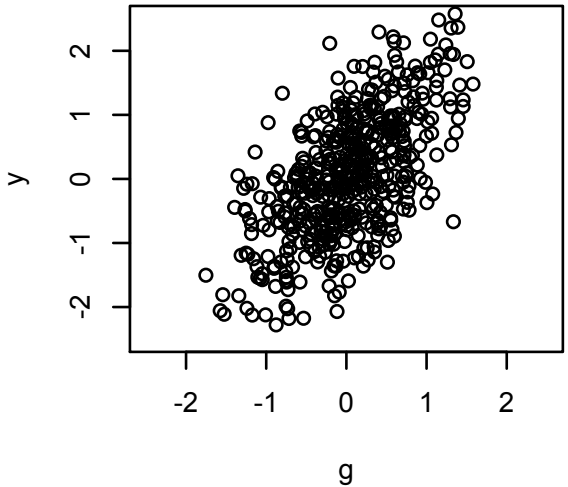
Does adding G help?

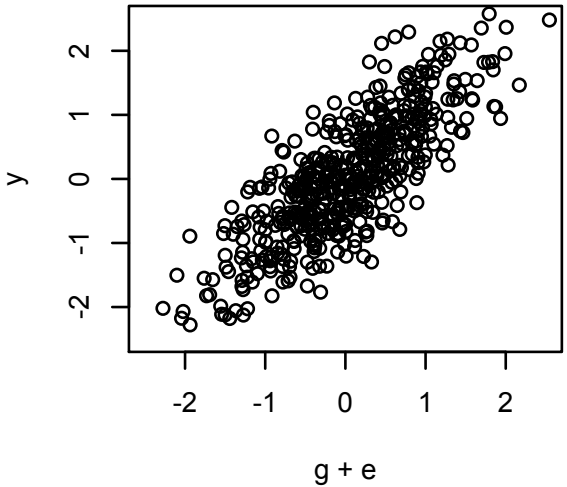

y
-2
-1
0
1
2
-2
-1
0
1
2
e


y
-2
-1
0
1
2
-2
-1
0
1
2
g


y
-2
-1
0
1
2
-2
-1
0
1
2
g + e


y
-2
-1
0
1
2
-2
-1
0
1
2
e


y
-2
-1
0
1
2
-2
-1
0
1
2
g


y
-2
-1
0
1
2
-2
-1
0
1
2
g + e

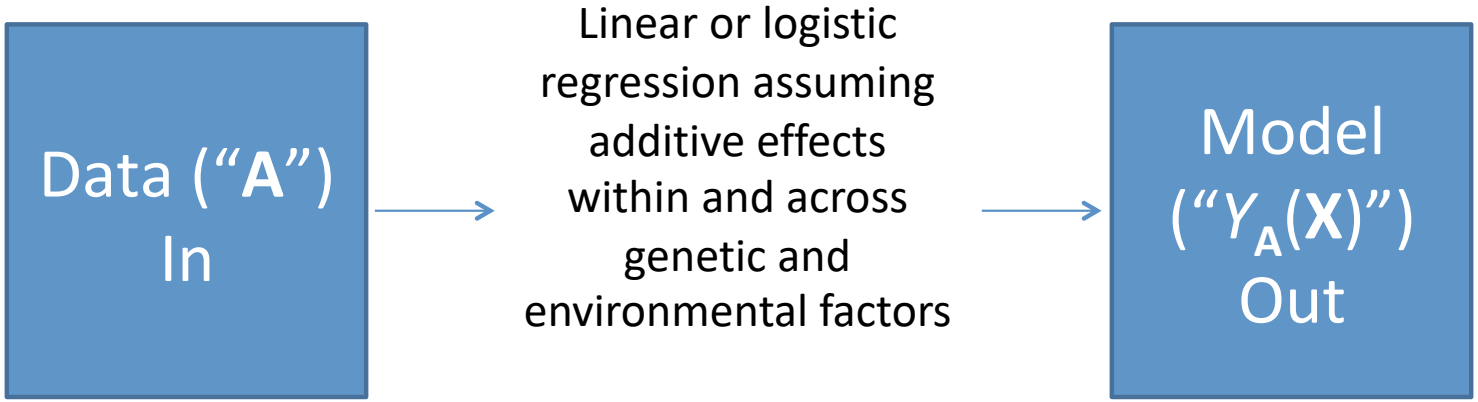
Training Models Including G & E

What functional form should the model have? What variables should be included? How should they be coded?

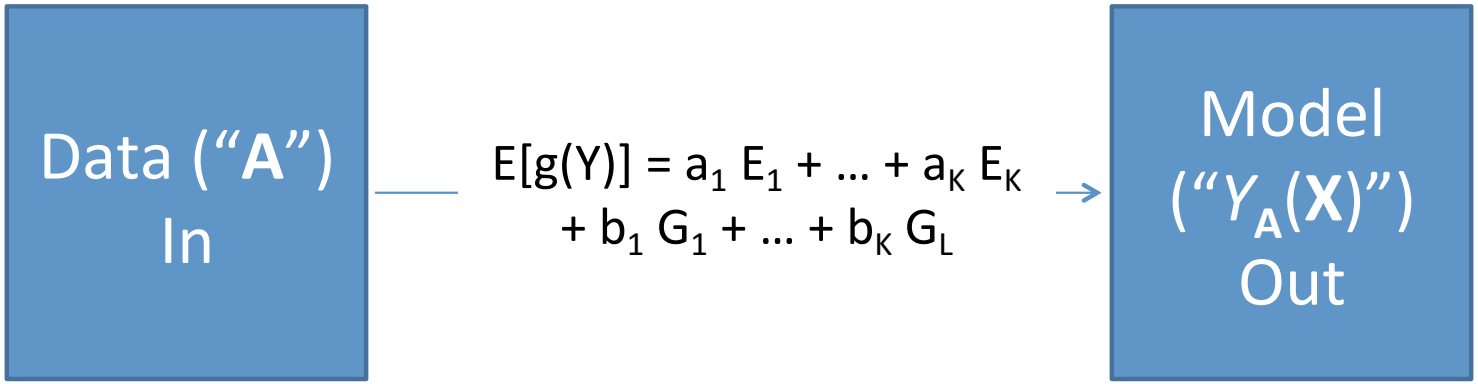
Data ("**A**") In

Model ("$Y_A(\mathbf{X})$") Out

Training Models Including G & E



Training Models Including G & E



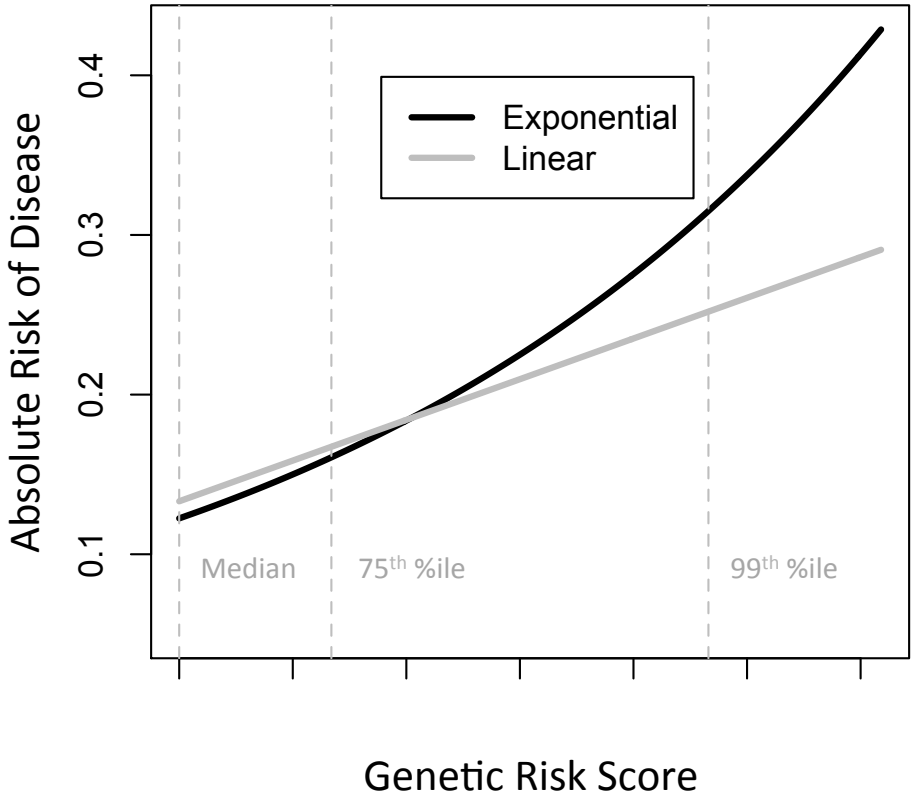
Training Models Including G & E

What about non-linear effects (aka interactions)?
School of hard knocks: linear terms pick up most of the signal, potential gains from including non-linear terms swamped by degrees of freedom.

Training Models Including G & E

What about non-linear effects (aka interactions)?
School of hard knocks: linear terms pick up most of the signal, potential gains from including non-linear terms swamped by degrees of freedom.

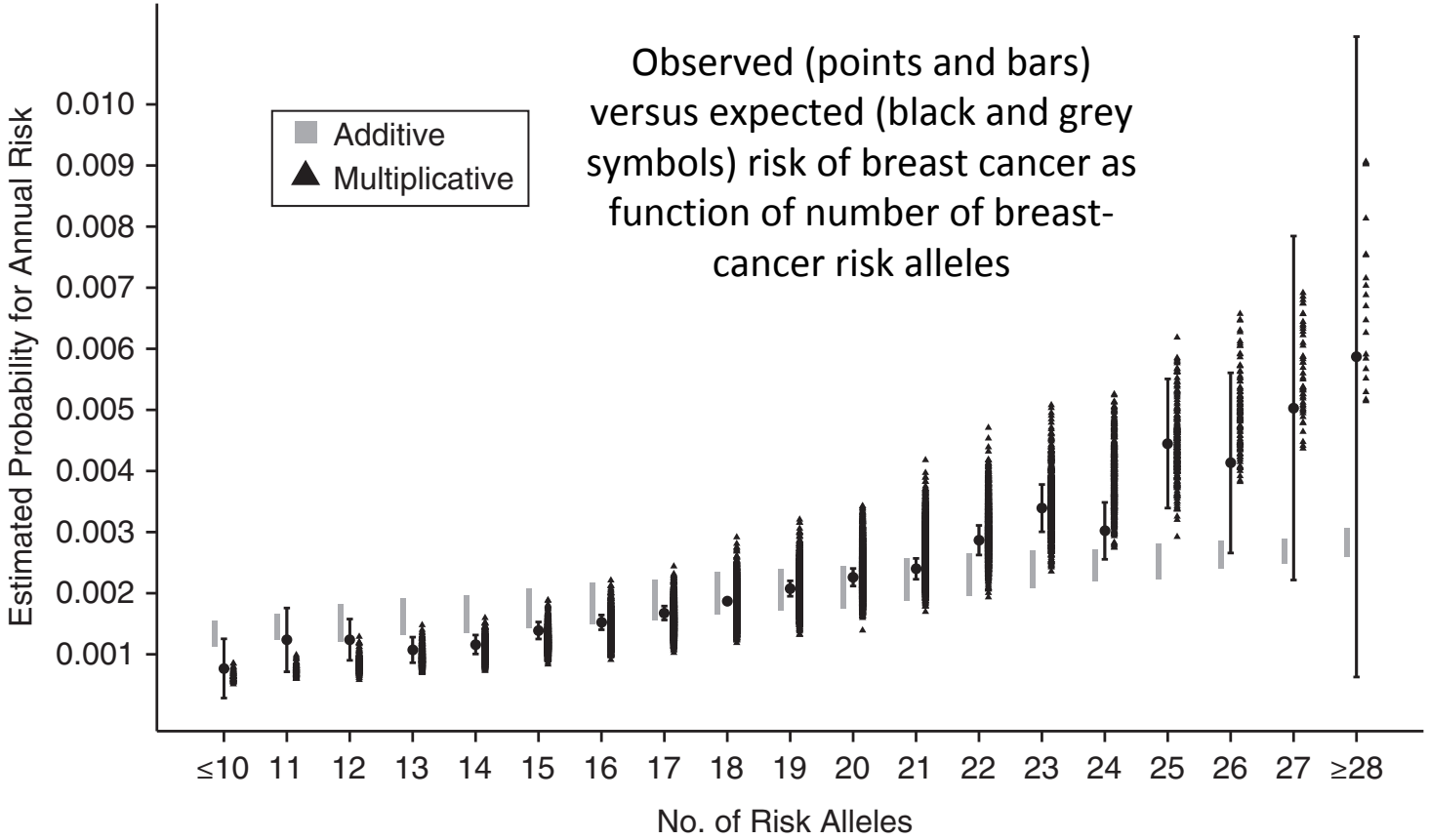
This model choice has implications for *extrapolated* risks in the tails.



The *assumed* log-additive relative risk model has notable implications for individuals in the tails of the genetic risk distribution,

More work is needed to verify that this model is a good fit *in the tails*.

Kraft (2017) *J Clin Oncol*



Joshi (2014) *Am J Hum Genet*

Testing calibration of risk models at extremes of disease risk

MINSUN SONG

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA

PETER KRAFT, AMIT D. JOSHI

Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

MYRTO BARRDAHL

Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

NILANJAN CHATTERJEE*

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA

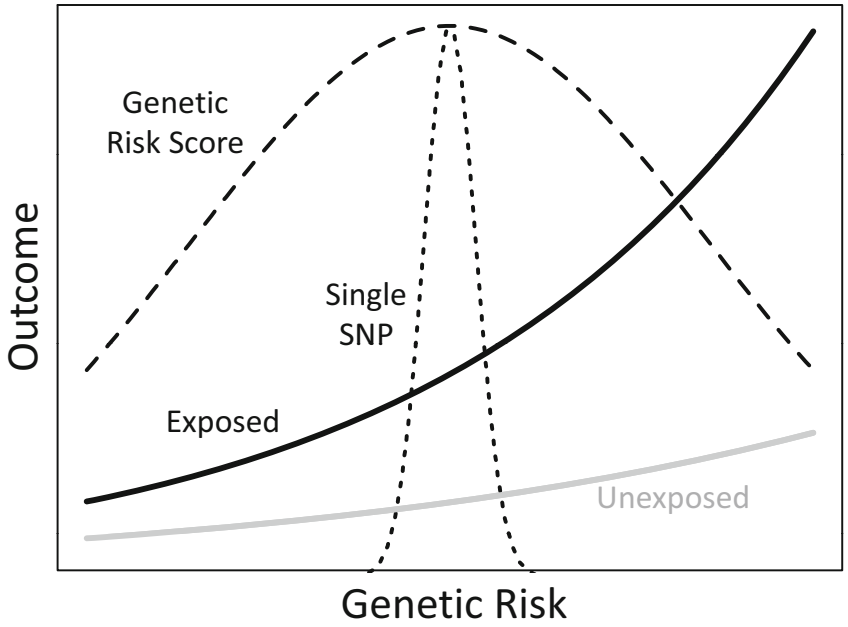
chattern@mail.nih.gov

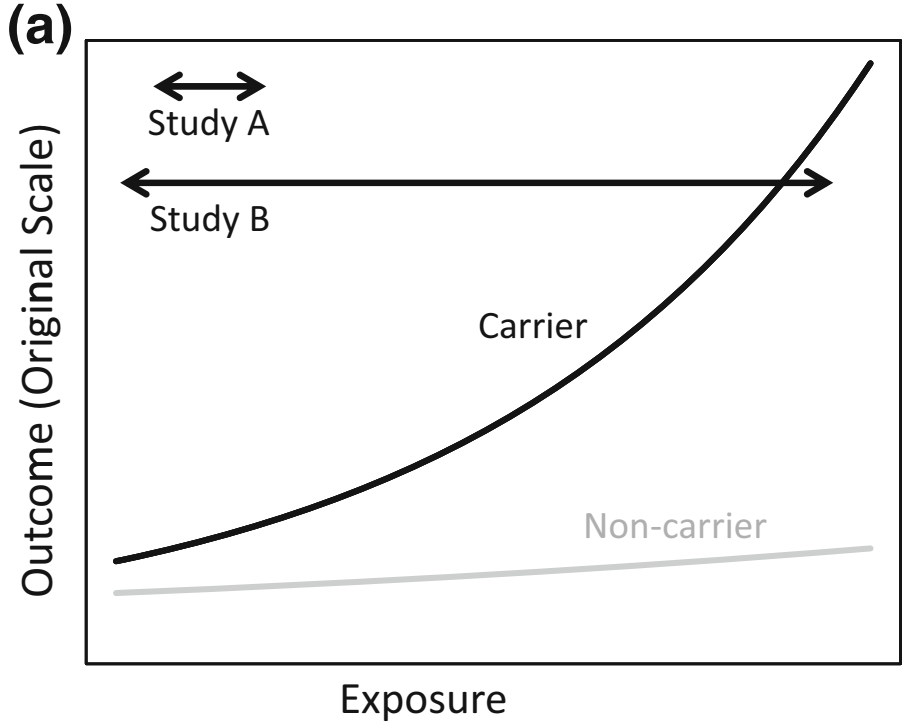
Biostatistics (2015)

Training Models Including G & E

What about non-linear effects (aka interactions)?
School of hard knocks: linear terms pick up most of the signal, potential gains from including non-linear terms swamped by degrees of freedom.

Constraining GxE effects may improve model fit
(assuming SNP-E interactions are mostly in the same direction).
Increasing exposure variance may as well.


Genetic
Risk Score
Single
SNP
Exposed
Unexposed
Outcome
Genetic Risk

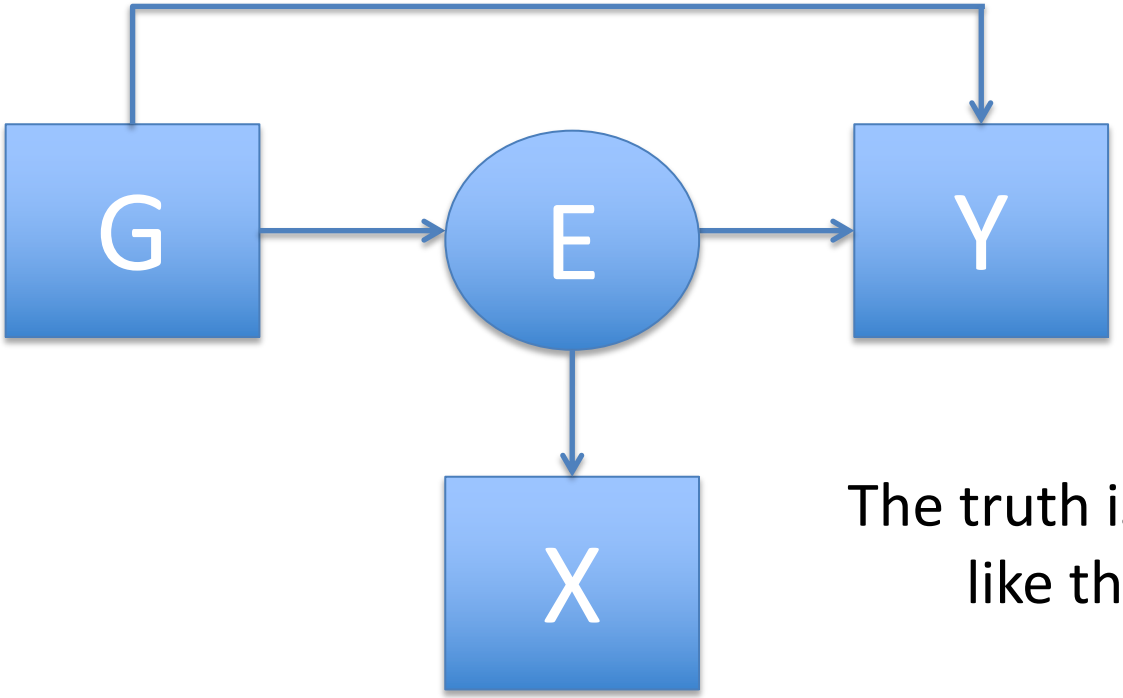

(a)
Study A
Study B
Carrier
Non-carrier
Outcome (Original Scale)
Exposure

What if G is Mediated thru E?

What if G is Mediated thru E?

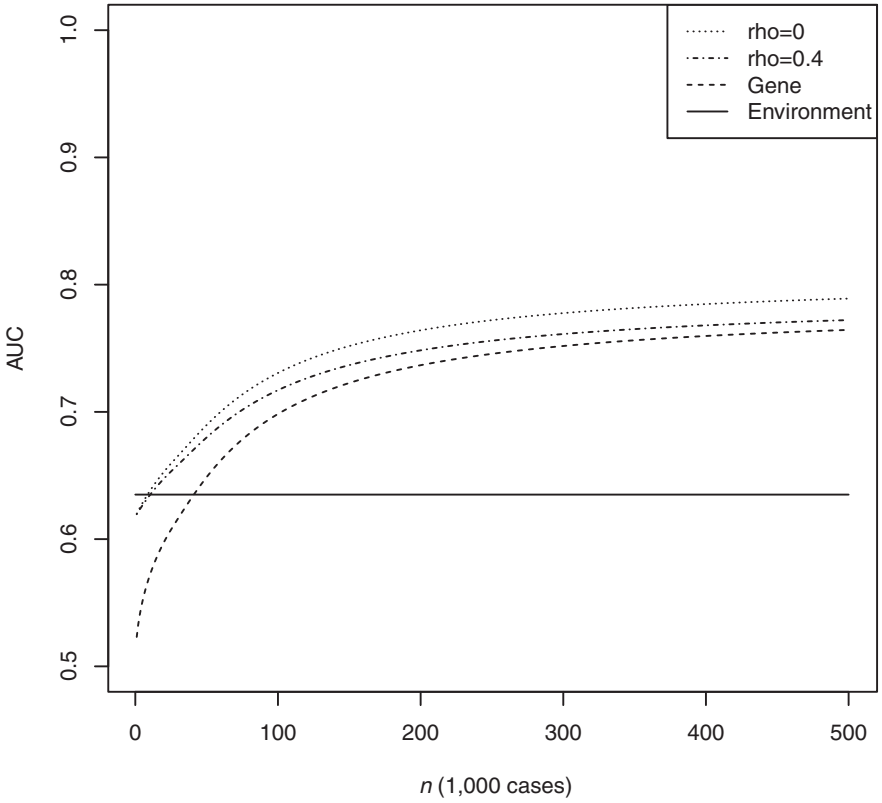
If *all* of G's effect is mediated thru E and we've measured E *without error*, then genetic terms $b_1,\ldots,b_K$ go to 0.

What if G is Mediated thru E?

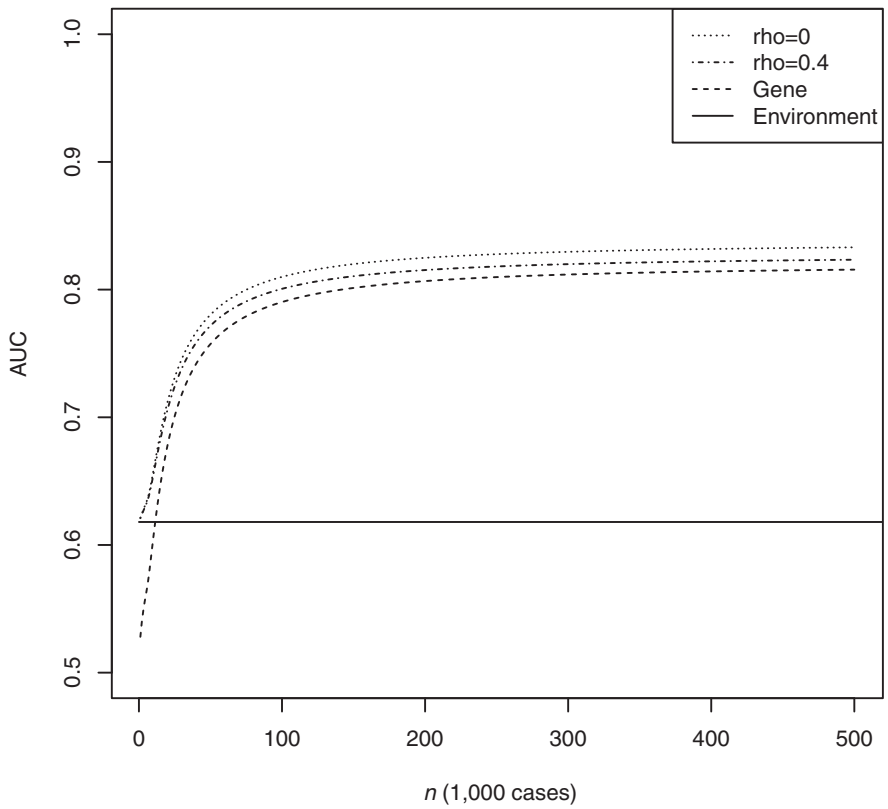


The truth is more like this.

CVD



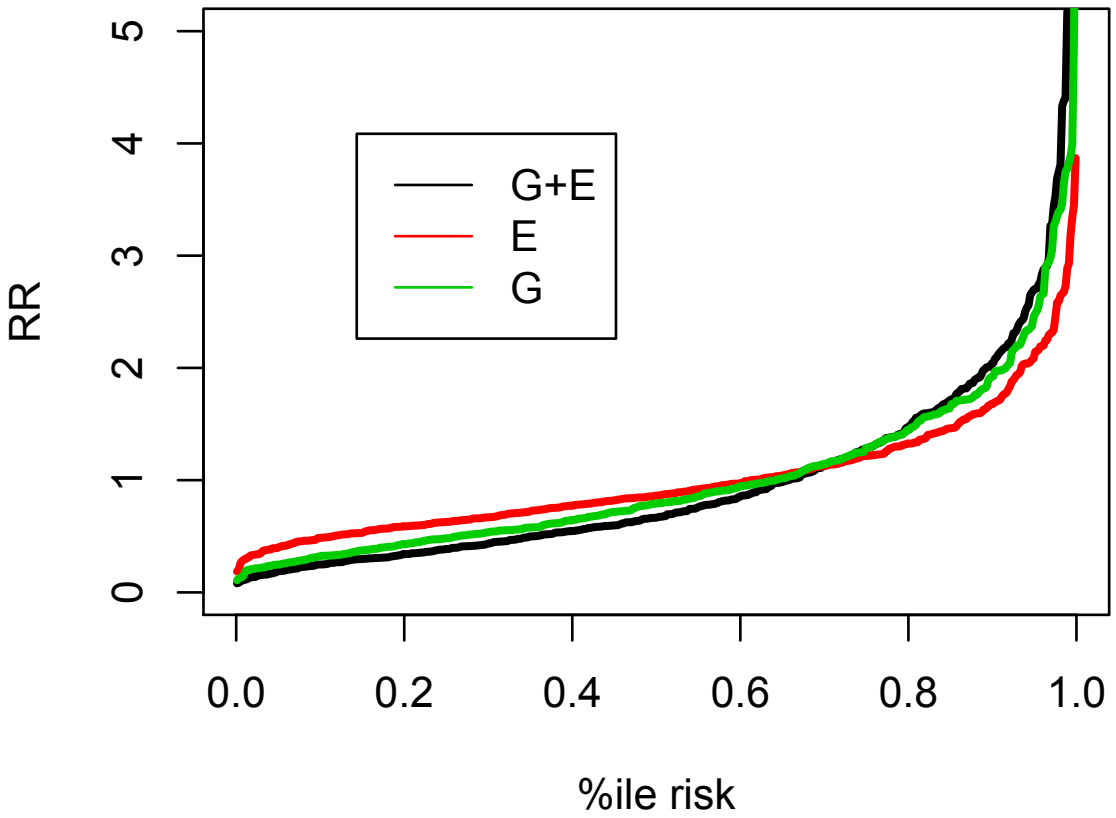
Breast Cancer



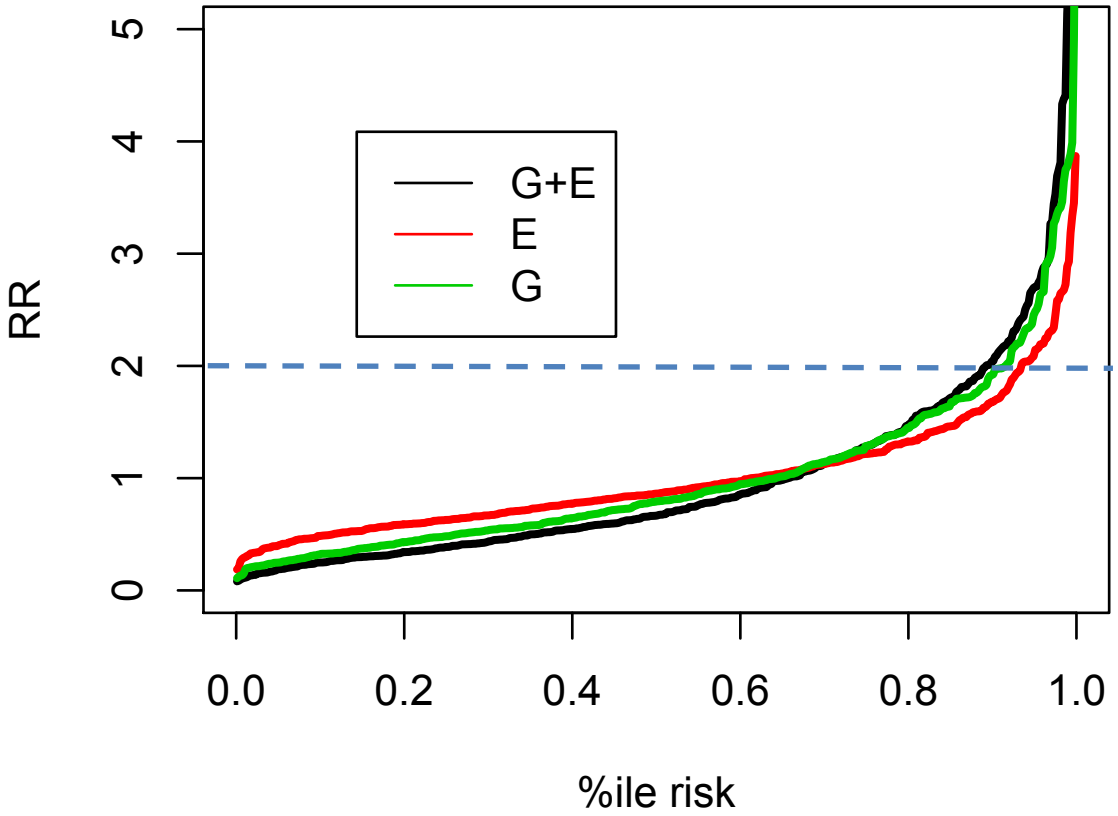
Dudbridge (2018) *Genet Epidemiol*

So, is it useful?

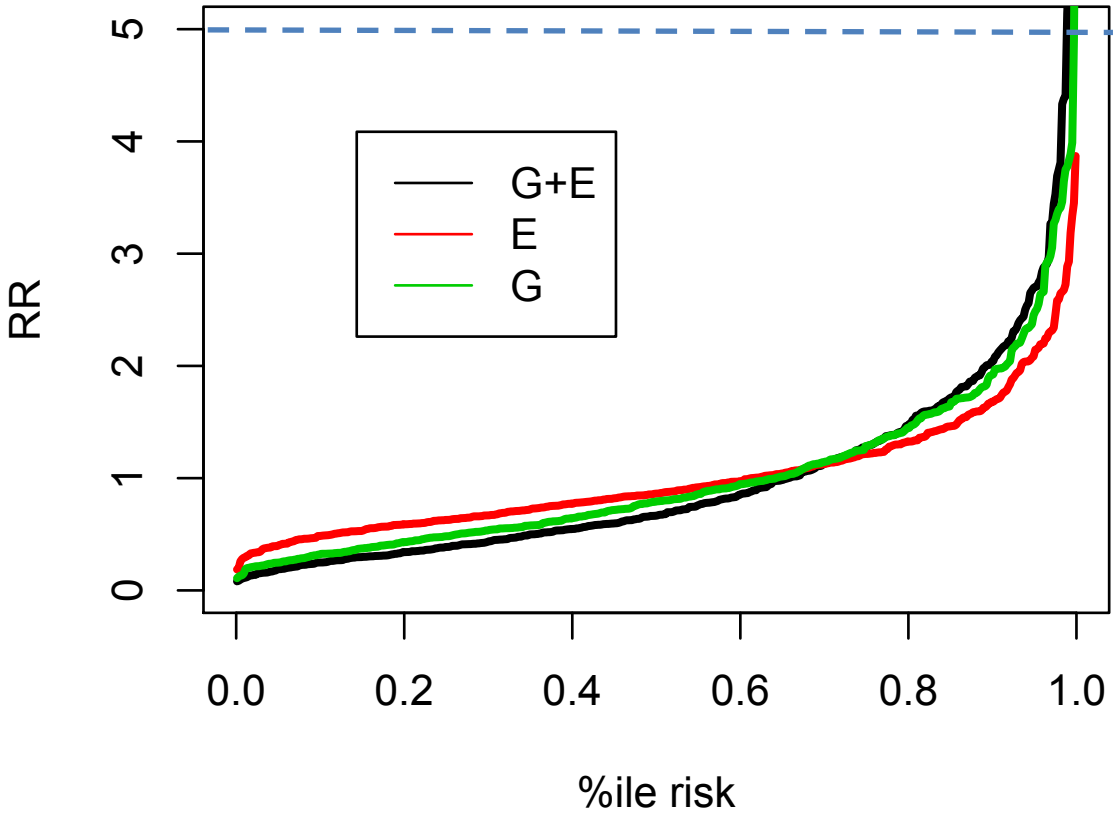
- Spreading risk distribution
- Identifying subgroups where G is actionable



Can increase the gradient of predicted risks by including G, but utility will depend on context—e.g. on the "action threshold" where expected benefits of intervention outweigh risks.



Adding G can identify 2x as many folks with RR>2



Adding G can identify
>10x as many folks
with RR>2

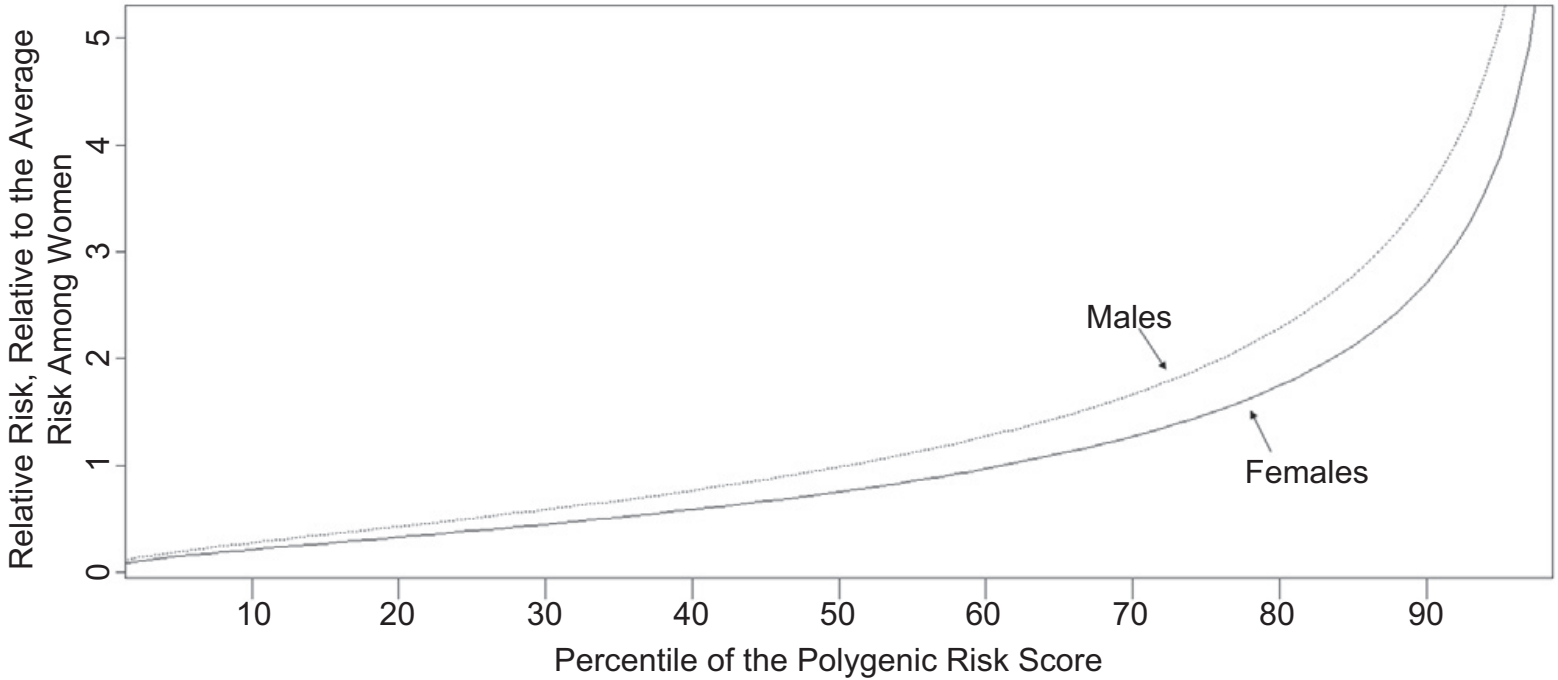
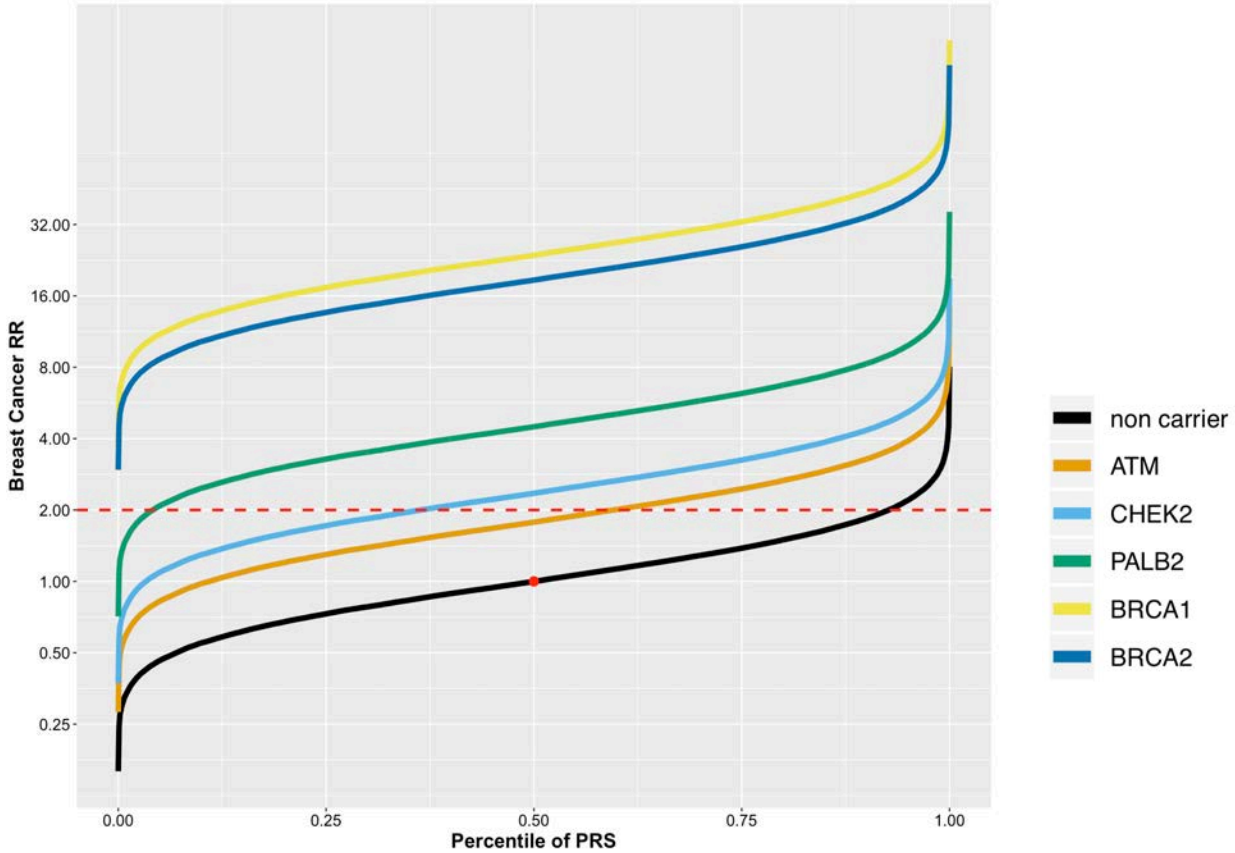


Figure 1. Cutaneous squamous cell carcinoma risk with increasing Polygenic Risk Score.

Women < 40 years old



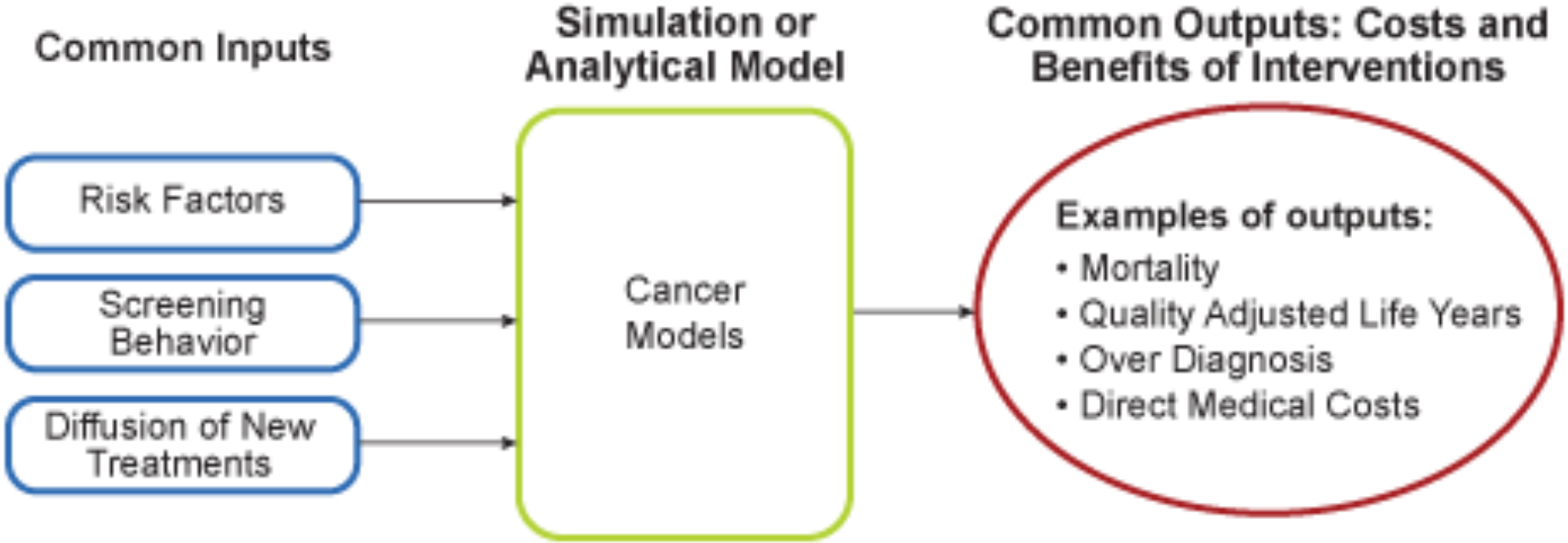
Preliminary results from CARRIERS consortium, presented at AACR 2019

Gao, Couch, Goldgar, Nathanson, Couch, Kraft et. al.

Determining clinical utility

- RCTs
- In the absence of large, expensive, and time consuming RCTs, we can simulate effectiveness using a model of disease natural history w/ or w/o intervention

Determining clinical utility

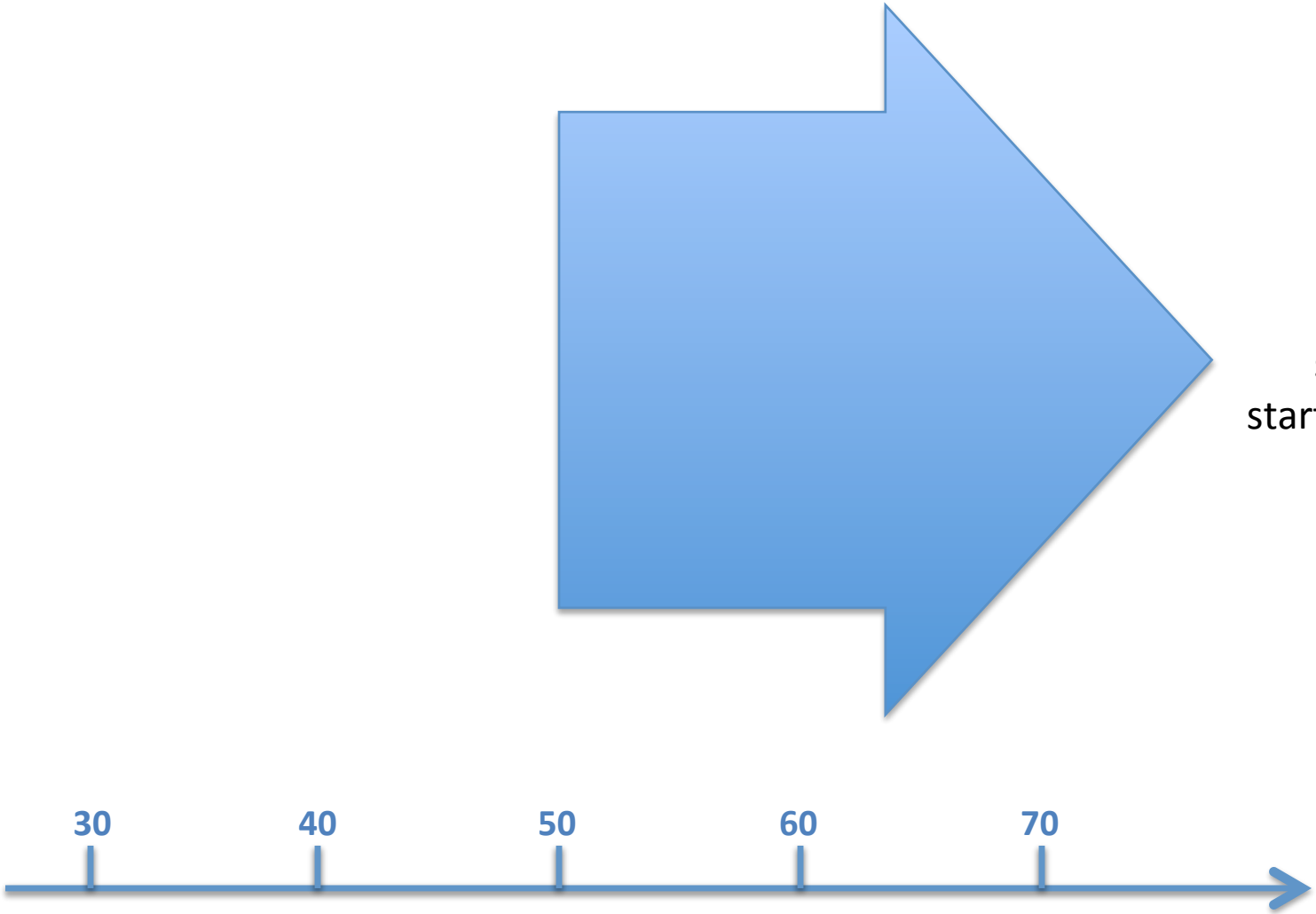


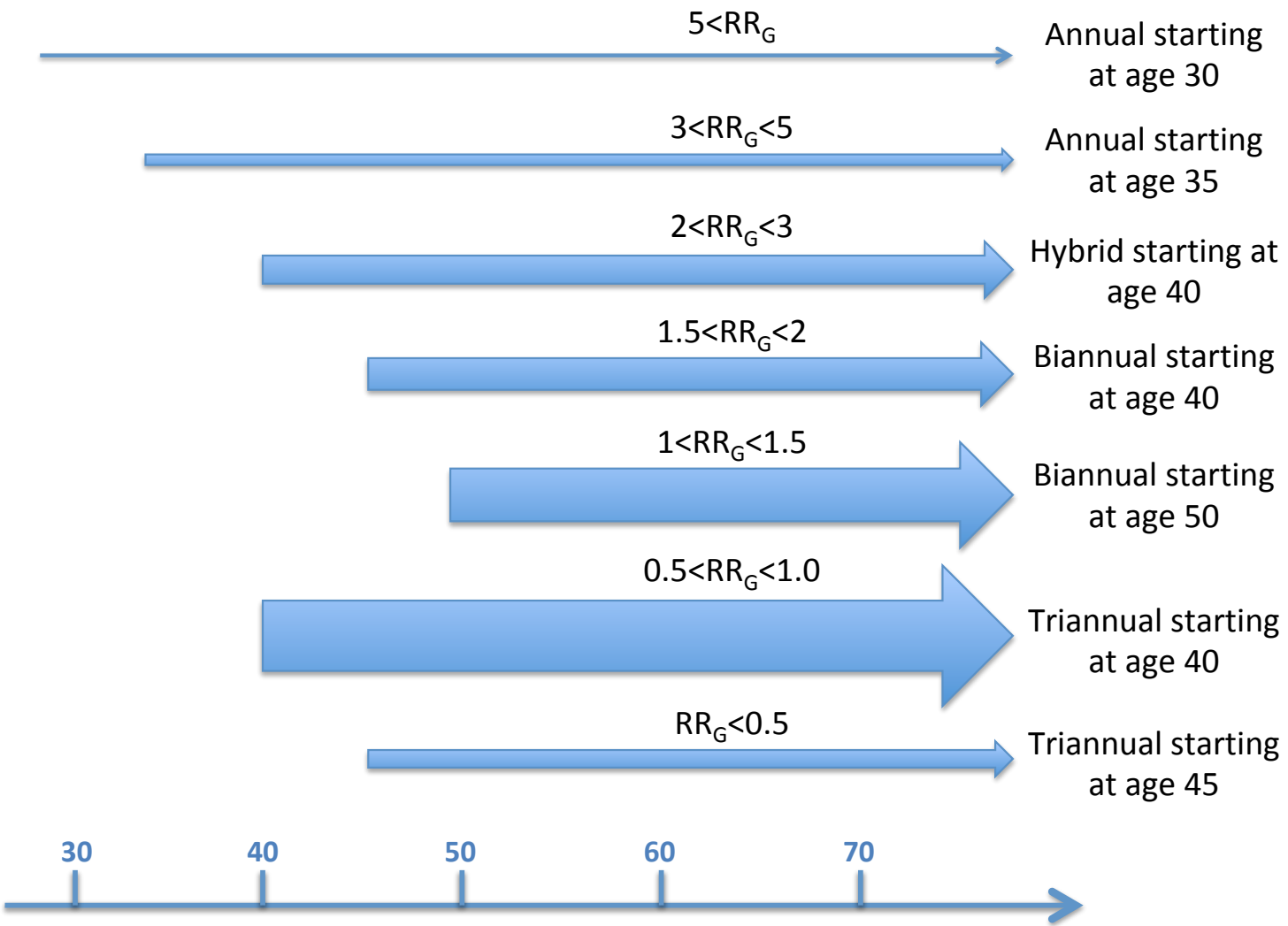
https://cisnet.cancer.gov/modeling/

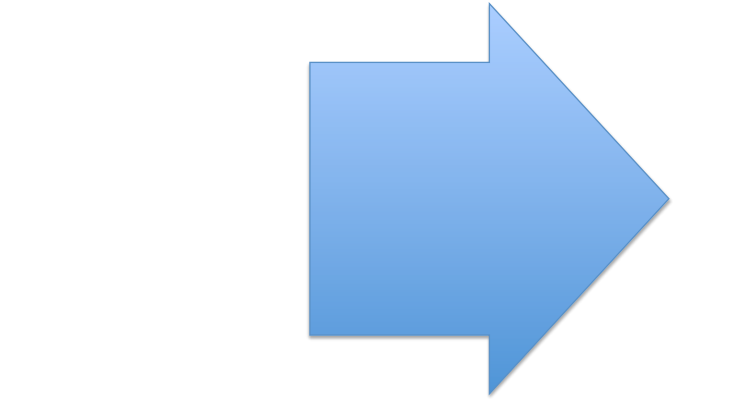
Determining clinical utility

Example: stratified breast cancer screening

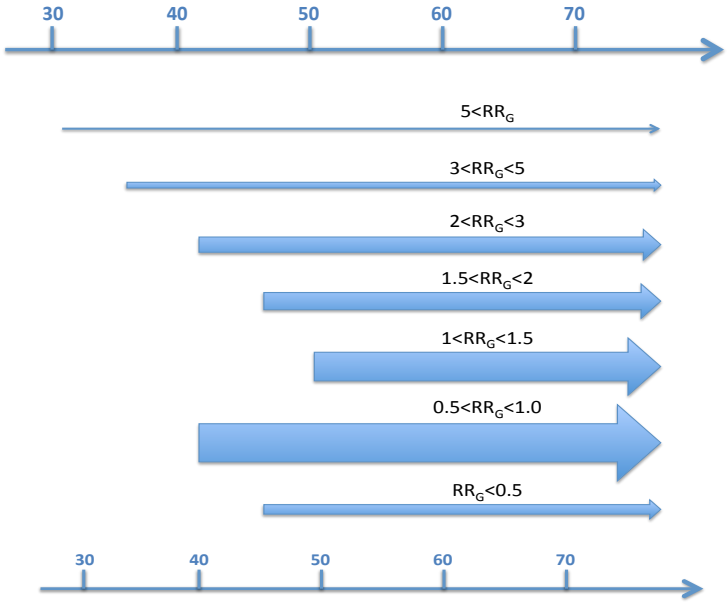
(preliminary work from CISNET breast cancer working group
—van den Broek et al.)


Bi-annual
screening
starting at age 50
30
40
50
60
70


5<RR_G
Annual starting at age 30
3<RR_G<5
Annual starting at age 35
2<RR_G<3
Hybrid starting at age 40
1.5<RR_G<2
Biannual starting at age 40
1<RR_G<1.5
Biannual starting at age 50
0.5<RR_G<1.0
Triannual starting at age 40
RR_G<0.5
Triannual starting at age 45
30
40
50
60
70

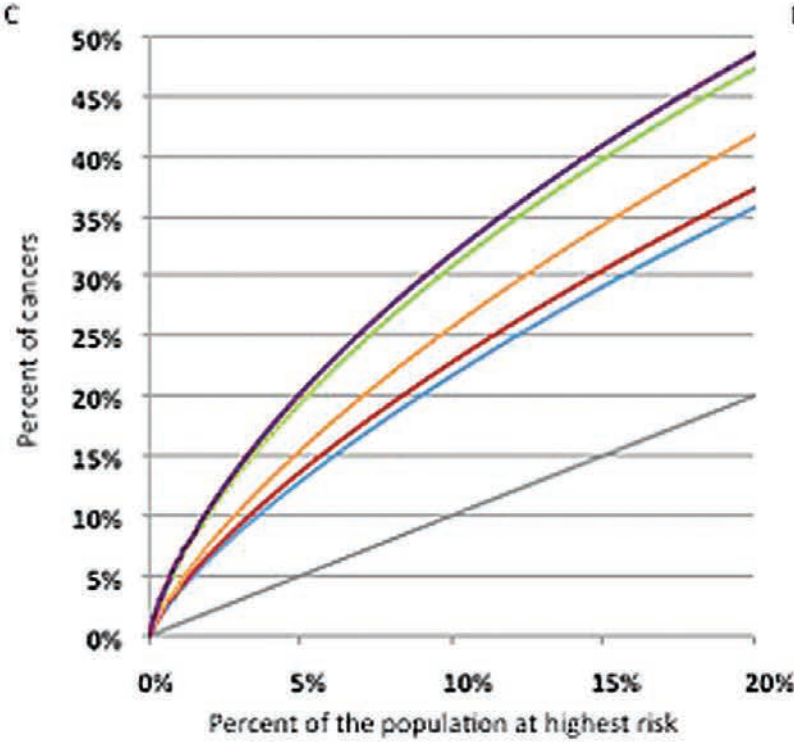


Life years gained: 118
Breast cancer deaths averted: 6.7



Life years gained: 196
Breast cancer deaths averted: 9.7

More can be better

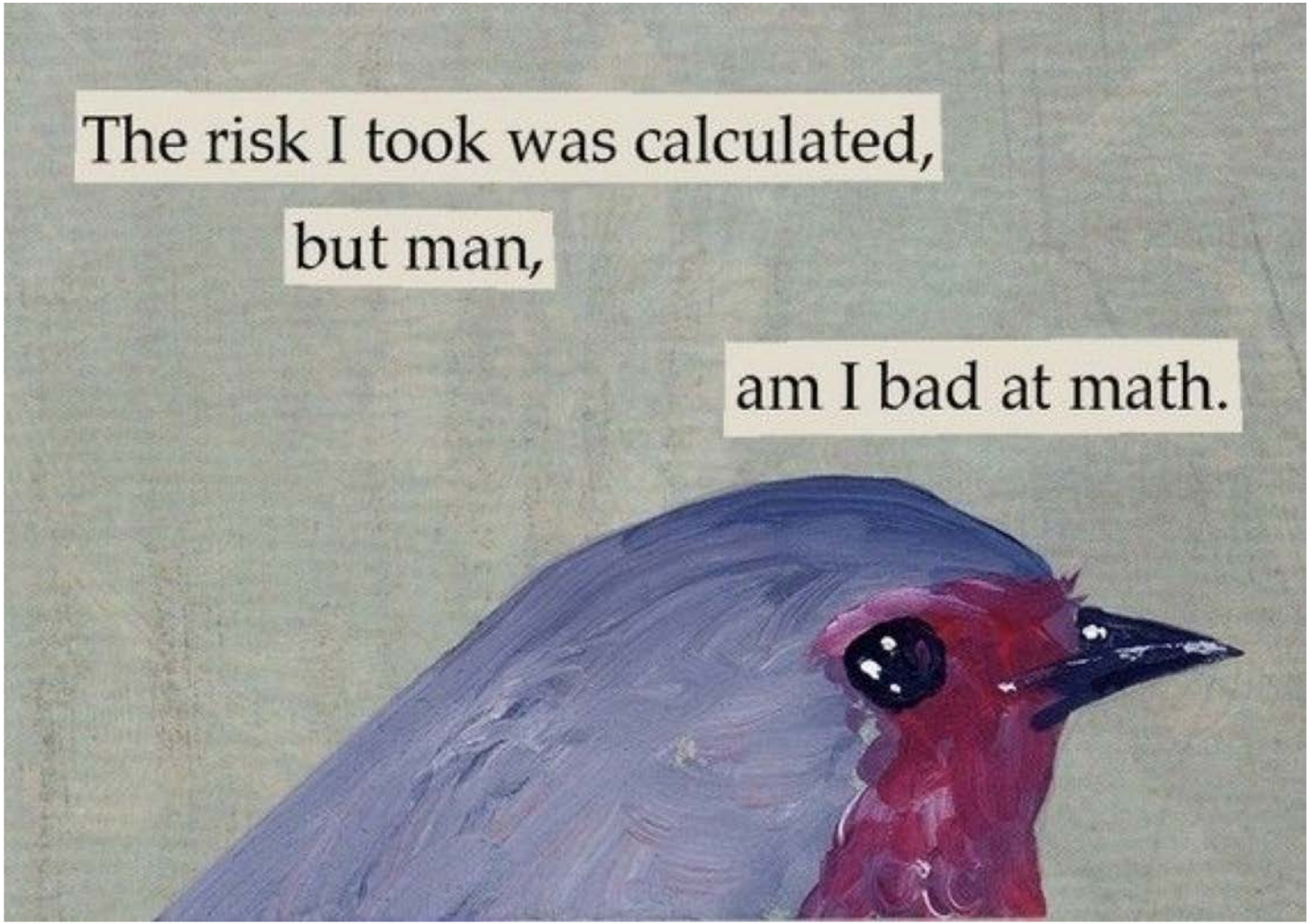


- Qx risk factors
- Mammographic density
- PRS
- Emerging biomarkers

Garcia-Closas (2014) JNCI

Misc Issues

- Implementing “complicated” E models
  —good locally, maybe not globally
- Including biology in risk models


The risk I took was calculated,
but man,
am I bad at math.