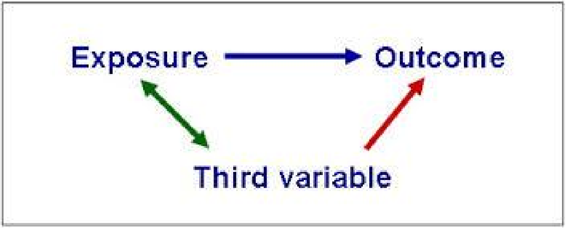
Special Topics in (Genetic) Epidemiology

Li Hsu
lih@fredhutch.org

Review: Epidemiologic Studies for Complex Diseases

- Complex diseases are contributed by both genetic and environmental risk factors
- Observational studies are powerful tools in studying these risk factors. Investigators have no control over exposure assignment. As a result, exposure of interest is often confounded by a third factor that is associated with exposure and the disease.



Genetic Factors

- Breakthroughs in high throughput genotyping and sequencing technologies have allowed researchers to assess genome-wide genetic effects on disease risk
    - High-dimension
    - Accurate measurements
    - Confounding can be effectively accounted for by principal components derived from genome-wide SNPs
- Areas that need further development: set-based association; high-dimensional risk prediction; GxE (or GxG) interaction

Environmental Risk Factors

- Besides confounding, measurement error is an issue for environmental risk factors
- Technologies such as wearable devices and metabolomics are being developed to better quantify aspects of environmental risk factors (e.g., diet and exercise)
  - Becoming more and more high-dimensional
  - Measurement error
  - Confounding remains to be a tricky issue
- Areas that need further development: functional data analysis to better characterize the effects of environmental covariates; measurement error; Mendelian randomization/instrumental variables

Study Designs

- Two commonly used study designs:
  - Case-control studies
    - Case-control studies are restrospective in nature, nevertheless the data can be analyzed as if they were prospectively collected using a logistic model and the odds ratio approximates the relative risk if the disease prevalence is low (Anderson 1972; Prentice and Pyke 1979)
    - Baseline disease probability is not identifiable
  - Cohort studies
    - Follow a group of people over a period of time to study the association of exposures with disease occurrences
    - Baseline disease probability is identifiable, even with subsampling designs (case-cohort and case-control)
- Areas that need further development: secondary phenotypes, biomarker evaluation and validation, population screening and monitoring

This week's focus:

- Population attributable fraction
- Absolute risk estimation

Relative Risk

- Measure the strength of association based on (prospective) cohort studies

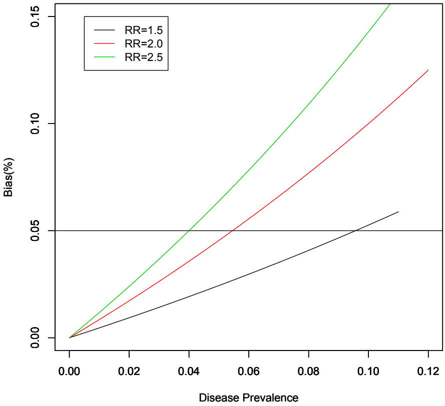
$$\text{Relative Risk(RR)} = \frac{\text{Risk in exposed}}{\text{Risk in non-exposed}} = \frac{\Pr(Y=1|Z=1)}{\Pr(Y=1|Z=0)}$$

- RR cannot be calculated directly in case-control studies. Instead, one can use odds ratio (OR)

$$\begin{aligned}\text{Odds Ratio} &= \frac{\text{Odds that an exposed subject develops disease}}{\text{Odds that a non-exposed subject develops disease}}\\ &= \frac{\text{Odds that a case was exposed}}{\text{Odds that a control was exposed}}\end{aligned}$$

When is the odds ratio a good estimate of relative risk

- When cases are representative of diseased population
- When controls are representative of non-diseased population
- When the disease is rare



Attributable Risk or Population Attributable Fraction

- Amount of disease that can be attributed to an exposure (Levin, 1953)

$$\text{PAR} = \frac{\Pr(Y = 1) - \Pr(Y = 1|Z = 0)}{\Pr(Y = 1)}$$

- It combines both the strength of association and the prevalence of exposure in the population, and therefore quantifies the population impact of risk factors

$$\text{PAR} = \frac{\Pr(Z = 1)(\text{RR} - 1)}{1 + P(Z = 1)(\text{RR} - 1)}$$

PAR

- Adjusted PAR is the reduction in incidence if a subset of risk factors is eliminated from the population while the other risk factors retain their actual levels (Whittemore 1982)

$$\mathrm{PAR}_{\mathrm{adj}} = \frac{\Pr(Y=1) - \sum_{j=1}^{m} \Pr(Y=1|Z=0, W=w_j)\Pr(W=w_j)}{\Pr(Y=1)}$$

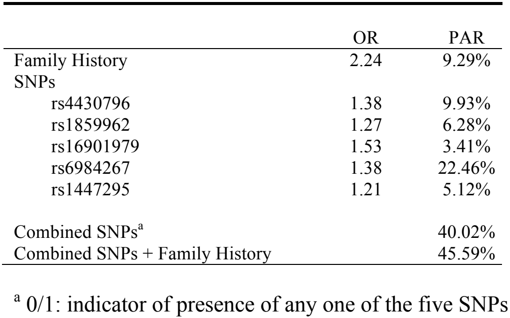
  where $w_1, \ldots, w_m$ are m levels of confounding variables

- It can also be formulated as

$$\mathrm{PAR} = \sum_{j=1}^{m} \Pr(Z=1, W=w_j|Y=1)(1 - \frac{1}{\mathrm{RR}_{\mathrm{adj}|W=w_j}})$$

- Benichou(2001); Greenland (2001); Silverberg et al. (2004); Graubard & Fears (2005)

A Case-Control Study of Prostate Cancer



	OR	PAR
Family History	2.24	9.29%
SNPs		
rs4430796	1.38	9.93%
rs1859962	1.27	6.28%
rs16901979	1.53	3.41%
rs6984267	1.38	22.46%
rs1447295	1.21	5.12%
Combined SNPs[a]		40.02%
Combined SNPs + Family History		45.59%

[a] 0/1: indicator of presence of any one of the five SNPs

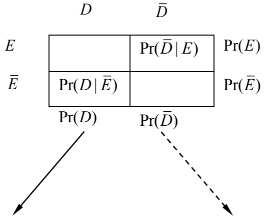
Zheng SL et al. N Engl J Med. 2008; 358(9):910-919

- A prominent property of the PAR is that PAR increases with exposure prevalence
- Q: Shall we simply increase the prevalence of exposure by including less risk alleles in genetic testing to make PAR look greater?

A Companion Measure

- Increasing PAR would lower the potential gain of disease-free, i.e., those who are not diseased, attributed to non-exposure.

Illustration of Population Attributable Risk and Population Attributable Benefit



	D	$\bar{D}$	
E		$\Pr(\bar{D} \mid E)$	$\Pr(E)$
$\bar{E}$	$\Pr(D \mid \bar{E})$		$\Pr(\bar{E})$
	$\Pr(D)$	$\Pr(\bar{D})$	

	Population Attributable Risk (PAR)	Population Attributable Benefit (PAB)
Definition	$PAR = \frac{\Pr(D) - \Pr(D \mid \bar{E})}{\Pr(D)} \times 100\%$	$PAB = \frac{\Pr(\bar{D}) - \Pr(\bar{D} \mid E)}{\Pr(\bar{D})} \times 100\%$
Interpretation	Fraction of excess risk of disease attributed to exposure	Fraction of excess gain of disease-free attributed to non-exposure
Measure	Public health impact on disease attributed to exposure	Public health impact on disease-free attributed to non-exposure

Relationship of PAR with Other Measures

- PAR is linearly associated with NPV and PAB is linearly associated with PPV.

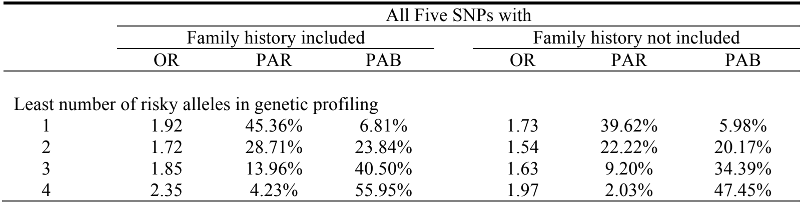
$$\text{PAR} = -\frac{\Pr(Y=0)}{\Pr(Y=1)} + \frac{1}{\Pr(Y=1)} \times NPV$$

and

$$\text{PAB} = \frac{\Pr(Y=1)}{\Pr(Y=0)} + \frac{1}{\Pr(Y=0)} \times PPV$$

Revisit the Prostate Cancer Study

OR, PAR and PAB of the Prostate Cancer Study



	All Five SNPs with					
	Family history included			Family history not included		
	OR	PAR	PAB	OR	PAR	PAB
Least number of risky alleles in genetic profiling						
1	1.92	45.36%	6.81%	1.73	39.62%	5.98%
2	1.72	28.71%	23.84%	1.54	22.22%	20.17%
3	1.85	13.96%	40.50%	1.63	9.20%	34.39%
4	2.35	4.23%	55.95%	1.97	2.03%	47.45%

Some Thoughts

- PAB addresses the question:

  How much benefit a prevention can gain to increase disease-free in the population attributed to non-exposure?

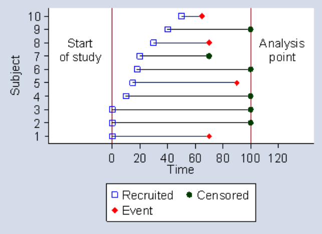
- PAR addresses the question

  How much risk a prevention can lower to reduce disease in the population attributed to exposure?

- Evan though PAB and PAR are tied in with PPV and NPV, they can be easily estimated from all major types of epidemiologic study designs (e.g., cohort and case-control); but not PPV and NPV.
- Chen, Hsu, & Peters (2010, manuscript)

Time-to-event Data

- Time-to-event outcomes have important applications in chronic disease application



- $Y(t)$: at-risk process; $N(t)$: counting process
- Hazard rate function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(t \leq T < t + \Delta t | T \geq t)$$

- Cox proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta' Z)$$

Time-Varying Attributable Risk

- Proportional reduction of probability of developing disease by time t due to an exposure over a period $[0, t]$

$$\Phi(t) = \frac{\Pr(T \leq t) - \Pr(T \leq t|Z = 0)}{\Pr(T \leq t)}$$

- Proportional reduction of hazard function due to Z

$$\phi(t) = \frac{\lambda(t) - \lambda_0(t)}{\lambda(t)}$$

Estimation of $\phi(t)$ from Cohort Data

- For time-to-event data, survival analysis techniques can be used to estimate survival function $S(t)$

$$\hat{\Phi}(t) = 1 - \frac{1 - \hat{S}_0(t)}{1 - \sum_{i=1}^{n} \hat{S}(t|Z_i)}$$

- Attributable hazard function

$$\phi(t) = 1 - \frac{\int f(T \geq t|z) f(z) dz}{\int \frac{\lambda(t|z)}{\lambda_0(t)} f(T \geq t|z) f(z) dz}$$

$$\hat{\phi}(t) = 1 - \frac{\sum_{i=1}^{n} \hat{S}(t|Z_i)}{\sum_{i=1}^{n} \exp(\widehat{\beta} Z_i) \hat{S}(t|Z_i)}$$

- Chen et al. (2006, *Biostatistics*); Chen et al. (2010 *Biometrika*); Liu et al. (2014, *JASA*)

Estimation of $\phi(t)$

- The time-varying attributable risk function can also be represented:

  $$\phi(t) = 1 - \int \frac{\lambda_0(t)}{\lambda(t|z)} f(z|T = t)dz$$

  Or

  $$\phi(t) = 1 - \left\{ \int \frac{\lambda(t|z)}{\lambda_0(t)} f(z|T \geq t)dz \right\}^{-1}$$

- $\phi(t)$ can be estimated from case or control data. E.g., a kernel estimator based on cases data is

  $$\hat{\phi}(t) = 1 - \frac{\sum_{i=1}^{n} \exp(-\hat{\beta}Z_i)Y_i K_h(t - X_i)}{\sum_{i=1}^{n} Y_i K_h(t - X_i)}$$

  where K$(\cdot)$ is a kernel function and h is the bandwidth that controls the spread of weighting window

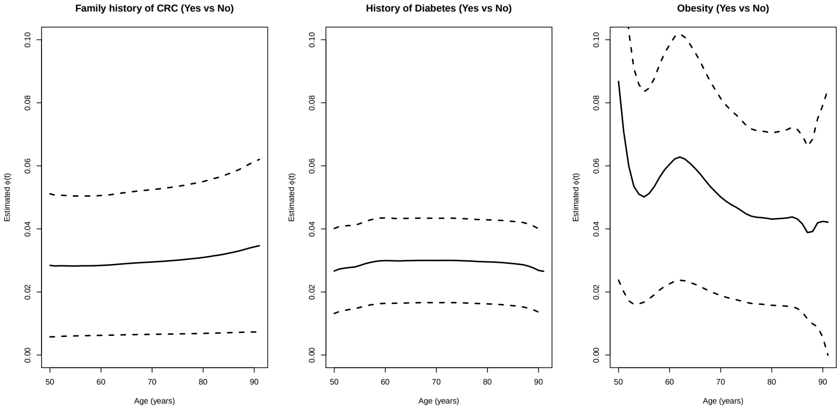
- Wei Zhao (2014, PhD dissertation, Department of Biostatistics, University of Washington)

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)

- GECCO includes 20+ population-based case-control and nested case-control studies (n $\approx$ 75,000). This example includes 2742 cases and 2756 controls from three population-based case-control studies.

	Cases	Controls		PAF
Variable	(%)	(%)	OR (95% CI)	(%)
Family history of CRC	17.6	15.1	1.21 (1.04, 1.40)	3.0
History of diabetes	7.5	4.7	1.65 (1.31, 2.07)	2.9
BMI($>$ 30kg/m^2)	29.5	25.7	1.21 (1.07, 1.36)	5.1

Time-Varying PAF



Summary: Relative Risk vs. Population Attributable Risk

- Relative risk and odds ratio are important measures of the strength of association.
  - Important for deriving causal inference.
- Attributable risk is a measure of how much disease risk is attributable to a certain exposure
  - Useful in determining how much disease can be prevented.
- Relative risk is valuable in etiologic studies of disease
- PAR (and/or PAB) is useful for public health guidelines and planning.

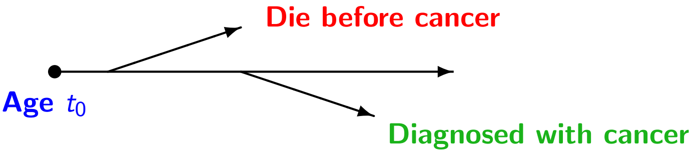
Absolute Risk

- Policy making and public health
- Counseling and personal risk management

Jolie, in her NYT article (March 24, 2015), explained:

“TWO years ago I wrote about my choice to have a preventive double mastectomy. A simple blood test had revealed that I carried a mutation in the BRCA1 gene. It gave me an estimated 87 percent risk of breast cancer and a 50 percent risk of ovarian cancer. I lost my mother, grandmother and aunt to cancer.”

Absolute Risk



- Absolute risk or crude risk is the probability that a person with a given set of risk factors, Z, and free of the disease of interest at time t_0 will develop disease before a subsequent age $t_0 + \tau$.
- Pure risk is the probability of disease if not competing causes of mortality were present.

Illustration

Life table to compare crude with pure risk.

Age at start of interval	# at risk	# incident breast cancer	# deaths from other causes
60	1000	17	44
65	939	20	63
70	856	22	87
75	745	—	—

- Crude risk by age 75 $= \frac{17+20+22}{1000} = 5.9\%$
- Pure risk by age 75 $= 1 - (1 - \frac{17}{1000})(1 - \frac{20}{939})(1 - \frac{22}{856})$ $= 6.3\%$
- Pure risk $>$ Crude risk because other causes of death are hypothetically eliminated.

Absolute Risk versus Pure Risk

- Focusing on pure risk helps understand the effects of an intervention on a particular outcome, regardless of its effect on competing causes of mortality.
- However, for clinical purposes, absolute risk is more pertinent because a patient is always subject to other causes of mortality.
- For example, it makes little sense to ask the question: "What would be your chance of developing breast cancer by age 80 if you had no risk of dying of non–breast cancer causes during the period?"

Survival Analysis and Competing Risks Framework

- Assume there are two types of events
  - $\varepsilon = 1$: the disease of interest (e.g., breast cancer)
  - $\varepsilon = 2$: competing causes (e.g., death from non–breast cancer causes)
- Let T be the time at which the first of these events occur.
- Cause-specific hazard for the disease of interest, $\varepsilon = 1$

$$\lambda(t) = \lim_{\Delta t \to 0} \Pr(t \leq T < t + \Delta t, \varepsilon = 1 | T \geq t)/\Delta t$$

- Cause-specific hazard for the competing causes, $\varepsilon = 2$

$$\lambda^{\dagger}(t) = \lim_{\Delta t \to 0} \Pr(\leq T < t + \Delta t, \varepsilon = 2 | T \geq t)/\Delta t$$

- If only one of the failure types occur, then

$$\lambda^{\text{overall}}(t) = \lambda(t) + \lambda^{\dagger}(t)$$

Absolute Risk

- Absolute risk is defined as

$$\begin{aligned} R(t|t_0; Z) &= \Pr(t_0 \leq T \leq t, \varepsilon = 1 | T \geq t_0, Z) \\ &= \int_{t_0}^{t} \lambda(u|Z) \exp(-\int_{t_0}^{u} \{\lambda(s|Z) + \lambda^{\dagger}(s|Z)\} ds) du \end{aligned}$$

- Integration of instantaneous probabilities of developing disease between t_0 and t.

Estimation

- For $i = 1, \cdots, n$ subjects
  - $X_i = \min(T_i, C_i)$, where T_i is the minimum failure time of competing causes and C_i is censoring time
  - $\delta_{ik} = I(T_i \leq C_i, \varepsilon_i = k)$: disease indicator for causes $k = 1, 2$
  - $N_{ik}(t) = I(X_i \leq t, \delta_{ik} = 1)$
  - $Y_i(t) = I(X_i \geq t)$

Estimation

- The likelihood function is

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \lambda(X_i|Z_i)^{\delta_{i1}} \lambda^{\dagger}(X_i|Z_i)^{\delta_{i2}} \exp\left(-\int_0^{X_i} \{\lambda(u|Z_i) + \lambda^{\dagger}(u|Z_i)\} du\right) \\
&= \prod_{i=1}^{n} \lambda(X_i|Z_i)^{\delta_{i1}} \exp\left(-\int_0^{X_i} \lambda(u|Z_i) du\right) \\
&X \prod_{i=1}^{n} (X_i|Z_i)^{\delta_{i2}} \exp\left(-\int_0^{X_i} \lambda^{\dagger}(u|Z_i) du\right)
\end{aligned}
$$

- Parameters can be estimated in a standard way by treating failures from other causes as censoring if there are no common parameters.

Estimation

- Breslow estimator is

$$\tilde{\Lambda}_{0k}(t) = \sum_{s \leq t} \frac{\sum_{i=1}^{n} N_{ik}(\Delta s)}{\sum_{i=1}^{n} Y_i(s) \exp(\widehat{\beta} Z_i)}$$

- Takes jumps at observed failure times, and hence it can be efficient due to sparse events.

Efficiency

- Use external disease incidence rates, denoted by $\lambda(t)$, from a national registry or other large cohort studies to improve the efficiency.
- Attributable hazard function

$$\phi(t) = \frac{\lambda(t) - \lambda_0(t)}{\lambda(t)}$$

  After rearrangement,

$$\lambda_0(t) = \lambda(t)(1 - \phi(t))$$

- Recall

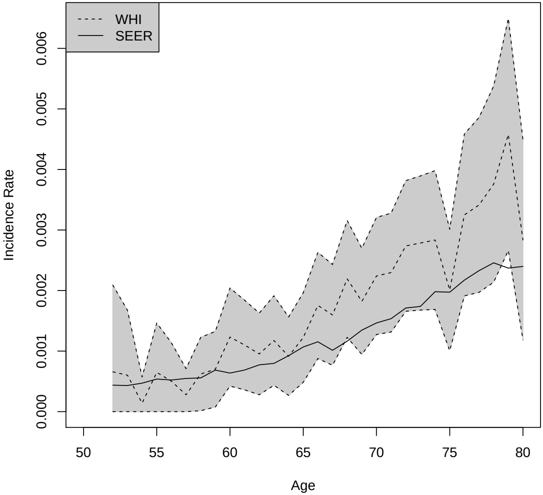
$$\begin{aligned}\phi(t) &= 1 - \left\{ \int \frac{\lambda(t|Z)}{\lambda_0(t)} f(Z|T \geq t) dZ \right\}^{-1} \\ &= 1 - \frac{\int S(t|Z) S^{\dagger}(t|Z) f(Z) dZ}{\int \exp(\beta Z) S(t|Z) S^{\dagger}(t|Z) f(Z) dZ}\end{aligned}$$

- When the competing risks is non-differential, i.e., $S^{\dagger}(t|Z)$ $= S^{\dagger}(t)$, the competing risk terms are canceled out.

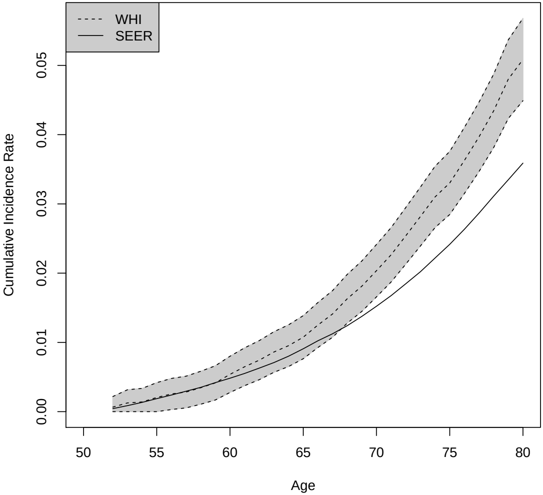
External Incidence Rate

- When a suitable external cause-specific composite incidence rate $\lambda(t)$ is available, we can plug it in for $\lambda(t)$ and estimate $\lambda_0(t) = \lambda^*(t)(1 - \phi(t))$.
- The incidence rate in the cohort may differ from the external rate because of eligibility criteria and participant characteristics such that cohort participants may not be entirely representative of the population.

WHI and SEER's Incidence Rates



WHI and SEER's Cumulative Incidence Rates



Difference between Cohort and External Incidence Rate

- The difference may be accommodated by

  $$\lambda_0(t) = \rho_0(1 - \phi(t))\lambda(t)$$

  where $\rho_0 = 1$ indicating no difference of disease incidence rates between the cohort and the external source.

- An estimator for ρ is

  $$\widehat{\rho} = \frac{\int_0^\tau \sum_{i=1}^n N_{ik}(du)}{\int_0^\tau \sum_{i=1}^n Y_i(u)\exp(\widehat{\beta} Z_i)\widehat{\phi}(u)\lambda(u)du}$$

Description of the WHI

- 1,073 (1.4%) developed CRC and 9,190(12%) died during the follow-up.
- T: age at the diagnose of CRC
- Risk factors chosen based on Freedman et al. (2009).
  - History of endoscopy (Endo) and polyps (Polyp) in last 5 years
  - Family history of CRC in the first-degree relatives (FH)
  - Current leisure-time vigorous activity (Exer, hours/week).
  - Use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs, nonuser, regular user)
  - Vegetable consumption (Veg, median portion/day).
  - BMI (kg/m^2)
  - Estrogen status

The hazad ratio (HR) estimates and the 95% CI of known risk factors

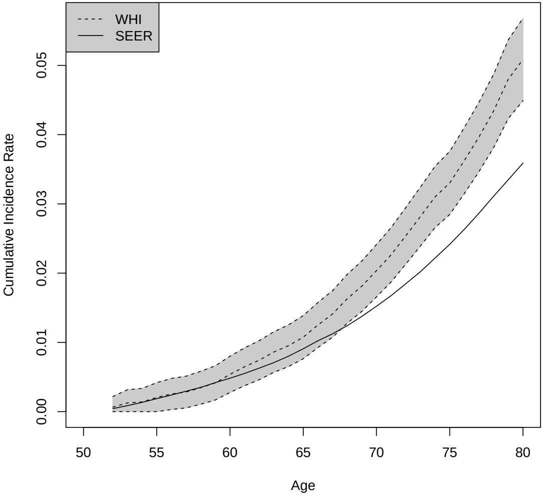
Risk Factor	*HR*	95% CI	P-value
Endoscopy and polyp history in the last 5 years			
Endoscopy and no history of polyps	1.00		
No endoscopy	1.30	(1.14, 1.48)	$< .0001$
Endoscopy and history of polyps	1.16	(0.95, 1.42)	0.0771
Endoscopy and polyps unknown	0.96	(0.66, 1.40)	0.4187
No. of relatives with CRC			
0	1.00		
≥ 1	1.23	(1.05, 1.43)	0.0051
Current vigorous leisure exercise, h/wk			
0	1.00		
$> 0, \leq 2$	0.99	(0.83, 1.18)	0.4390
> 4	0.83	(0.68, 1.03)	0.0445
Aspirin/NSAID use			
Nonuser	1.00		
Regular user	0.76	(0.65, 0.90)	0.0005
Vegetable intake, medium portion per day			
< 5	1.00		
≥ 5	0.94	(0.83, 1.06)	0.1556
BMI, kg/m^2			
< 30	1.00		
≥ 30	1.38	(1.21, 1.59)	$< .0001$
Estrogen status within the last 2 years			
Negative	1.00		
Positive	0.87	(0.76,0.99)	0.0184

The 10-year CRC risk estimates and 95% CI

								Estrogen	10-Year Risk	
Age	Endo	Polyp	FH	Exer	NSAIDs	Veg	BMI	Status	%	95%CI
50	Yes	No	0	3	Yes	2.5	28	Pos	0.38	(0.29, 0.47)
50	Yes	Yes	1	1	Yes	2.5	29	Neg	0.62	(0.44, 0.79)
50	No		2	0	No	1.3	32	Neg	1.59	(1.21, 1.98)
									Breslow Estimator	
50	Yes	No	0	3	Yes	2.5	28	Pos	0.29	(0.15, 0.42)
50	Yes	Yes	1	1	Yes	2.5	29	Neg	0.47	(0.22, 0.72)
50	No		2	0	No	1.3	32	Neg	1.22	(0.65, 1.79)

- $\widehat{\rho} = 1.18$ (95% CI: 1.01-1.36) for age < 65, $\widehat{\rho} = 1.63$ (95% CI: 1.52, 1.73) for age ≥ 65.

WHI and SEER's Cumulative Incidence Rates



Case-Control Studies and Survival Analysis

- Assume failure time T follows the Cox model

$$\lambda(t;z) = \lambda_0(t)\exp(\beta' z).$$

- Suppose that n_1 cases and n_0 controls are sampled at time t. Then the probability of $z_1, \ldots, z_{n_1}$ corresponding to cases given $z_1, \ldots, z_{n_1+n_0}$ is

$$\frac{\prod_{i=1}^{n_1} \Pr\{z_i|d=1\} \prod_{i=(n_1+1)}^{n_1+n_0} \Pr\{z_i|d=0\}}{\sum_{l\in R(n_1,n_0)} \prod_{j\in l} \Pr\{z_j|d=1\} \prod_{j\notin l} \Pr\{z_i|d=0\}}$$

$$= \frac{\prod_{i=1}^{n_1} \exp(\beta z_i)}{\sum_{l\in R(n_1,n_0)} \prod_{j\in l} \exp(\beta z_j)}$$

- Hazard ratio β can be estimated from (conditional) logistic regression model based on case-control data (Prentice & Breslow 1978)
- However, $\lambda_0(t)$ is eliminated from the conditional likelihood function, and hence unidentifiable from case-control data.

Case-Control Studies

- Recall that PAR can be estimated from cases and controls data (Wei et al. 2014), i.e.,

$$\hat{\phi}(t) = 1 - \frac{\sum_{i=1}^{n} \exp(-\hat{\beta} Z_i) Y_i K_h(t - X_i)}{\sum_{i=1}^{n} Y_i K_h(t - X_i)}$$

- We can estimate $\lambda_0(t)$ by

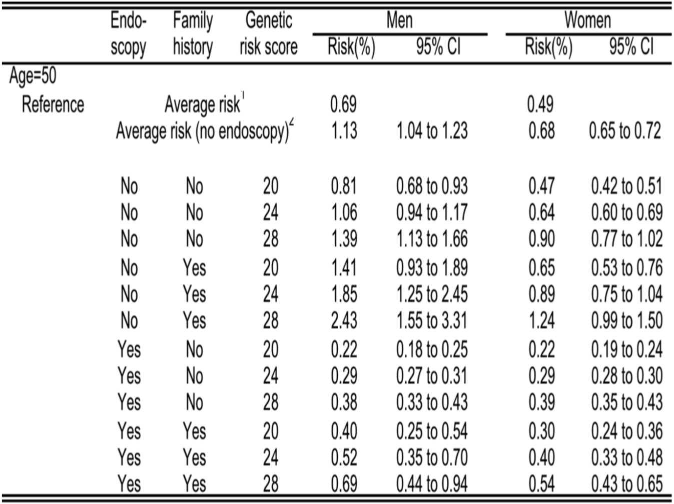
$$\widehat{\lambda}_0(t) = \{1 - \widehat{\phi}(t)\}\lambda(t)$$

where $\lambda(t)$ is external incidence rates.

Genetics and Epidemiology of Colorecal Cancer Consortium (GECCO)

- 20+ population-based case-control and nested case-control studies with n $\approx$ 75,000 GWAS, basic clinical, epidemiologic & lifestyle data
- Build a risk prediction model based on age, sex, family history of colorectal cancer (CRC), endoscopy, and genetic risk score of 27 GWAS-identified CRC loci

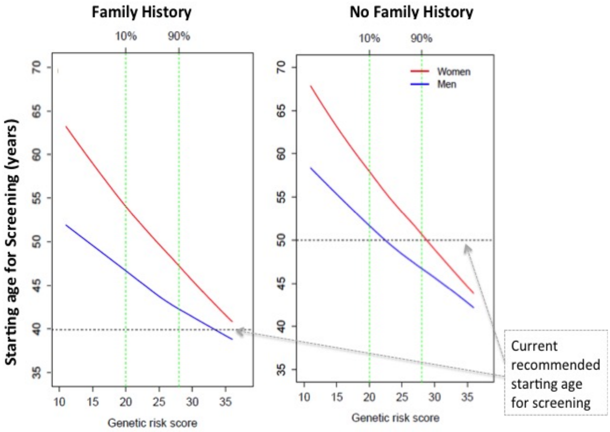
Examples of 10-year absolute risk with selected risk profiles



	Endoscopy	Family history	Genetic risk score	Men		Women	
				Risk(%)	95% CI	Risk(%)	95% CI
Age=50							
Reference	Average risk[1]			0.69		0.49	
	Average risk (no endoscopy)[2]			1.13	1.04 to 1.23	0.68	0.65 to 0.72
	No	No	20	0.81	0.68 to 0.93	0.47	0.42 to 0.51
	No	No	24	1.06	0.94 to 1.17	0.64	0.60 to 0.69
	No	No	28	1.39	1.13 to 1.66	0.90	0.77 to 1.02
	No	Yes	20	1.41	0.93 to 1.89	0.65	0.53 to 0.76
	No	Yes	24	1.85	1.25 to 2.45	0.89	0.75 to 1.04
	No	Yes	28	2.43	1.55 to 3.31	1.24	0.99 to 1.50
	Yes	No	20	0.22	0.18 to 0.25	0.22	0.19 to 0.24
	Yes	No	24	0.29	0.27 to 0.31	0.29	0.28 to 0.30
	Yes	No	28	0.38	0.33 to 0.43	0.39	0.35 to 0.43
	Yes	Yes	20	0.40	0.25 to 0.54	0.30	0.24 to 0.36
	Yes	Yes	24	0.52	0.35 to 0.70	0.40	0.33 to 0.48
	Yes	Yes	28	0.69	0.44 to 0.94	0.54	0.43 to 0.65

Examples of 10-year absolute risk with selected risk profiles

Starting Age at Screening by Genetic Risk Score

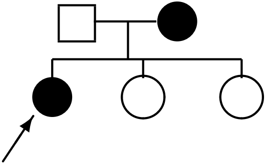


Current recommended starting age for screening

Risk threshold set to 0.91% (average 10-year risk for an individual aged 50 years person without prior endoscopy Men: 1.13%; Women:0.68%)

Alternative Approach to Obtaining $\lambda_0(t)$: Family History

- Many case-control studies of inherited diseases collect family history information including disease status (d_k) and failure time (x_k) of the relative (y_k), and in some cases, the relative' risk factors z_k.



$$\{y_k = (x_k, d_k), z_k, k = 1, \ldots, K\}$$

$$\{y_0 = (x_0, d_0), z_0\}$$

Likelihood Function

- The likelihood function is

$$\begin{aligned} L &= \prod f(y_1, z_1, z_0 | y_0) \\ &= \prod \underbrace{f(y_1, z_1 | y_0, z_0)}_{\text{relatives}} \; \underbrace{f(z_0 | y_0)}_{\text{cases/controls}} \; . \end{aligned}$$

- The first term involves the joint distribution of failure times for the family. This becomes a **bivariate** survival analysis problem.

$$\Pr(T_0 > t_0, T_1 > t_1 | z_1, z_0) = h(S(t_0|z_0), S(t_1|z_1); \theta),$$

where $S(t|z)$ is univariate survival function given z and h is a parametric function indexed by θ.

Estimation of Baseline Hazard Function

- Since the relatives' failure times are random, it is natural to formulate the hazard function for the relatives conditional on the case-control sampling.

$$\lambda(t|y_0, z_0, z_1) = \lambda_0(t) \underbrace{\exp(\beta' z_1)\mathcal{H}_\theta(t, y_0, z_0, z_1)}_{\text{time-dependent risk score}}$$

- $\lambda(t|y_0, z_0, z_1)$ has some resemblance to the Cox model, suggesting $\lambda_0(t)$ be estimated by a Breslow type estimator.
- However, $\mathcal{H}_\theta(t, y_0, z_0, z_1)$ may not be predictable at time t because $y_0 > t$.
- Use a two-stage estimator with the first stage limiting to subjects whose relatives' failure times are predictable followed by the second stage to include all subjects (Gorfine, Zucker and Hsu, 2009, Ann. Stat.)

Multiple Relatives

- When there are multiple relatives, the joint distribution of the family is

$$\Pr(T_0 > t_0, \cdots, T_K > t_K | z_0, \ldots, z_K) = h(S(t_0|z_0), \ldots, S(t_K|z_K); \theta),$$

- Take the GEE approach by breaking down a family into multiple relative-case/control pairs(Liang and Zeger 1986). The approach has the advantage of simple computation and robustness but the downside is potential efficiency loss.
- Or represent the copula model by the frailty model

$$\lambda(t|z_k, \omega) = \lambda_0(t) \exp(\beta' z_k)\omega,$$

  where ω is a common (latent) frailty shared by the relatives of the same relation.
- EM algorithm can be used to estimate the relevant parameters.

BRCA1 Data Analysis

- A population-based case-control study was conducted within the NICHD's Womens Contraceptive and Reproductive Experiences study (Marchbanks et al., 2002).
- A study of the BRCA1/2 genes was conducted to evaluate their contribution to breast cancer risk (Malone et al., 2006).
- In total, 1603 cases and controls were tested for BRCA1 mutations.
- 4568 first-degree female relatives were included, among them 634 (14%) developed breast cancer.

	# mutations
cases (n=1144)	42
controls (n=459)	1

Chen, Hsu and Malone (2009, Biometrics)

BRCA1 Data Analysis

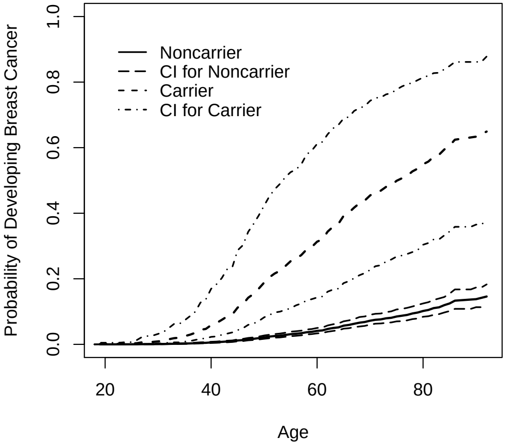
- Cumulative probabilities of developing breast cancer by age for BRCA1 mutations.

	Probability of Developing BC (95% CI)	
Age	Noncarrier	Carrier
50	0.021 (0.017, 0.027)	0.188 (0.082, 0.423)
60	0.041 (0.034, 0.050)	0.313 (0.143, 0.612)
70	0.072(0.060, 0.090)	0.454 (0.227, 0.743)
80	0.102 (0.084, 0.125)	0.549 (0.304, 0.814)

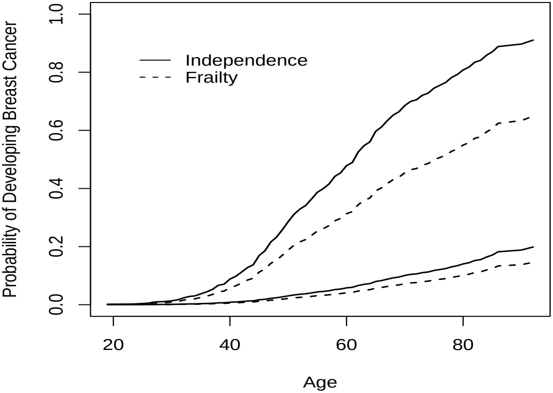
- The dependence parameter $\widehat{\theta} = 0.733$ (s.e. 0.208).

Chen, Hsu and Malone (2009, Biometrics)

Probability of Developing Breast Cancer by Age



Independence vs Frailty



Missing Covariates Information on Relatives

- Instead of modeling $\Pr(T_0 > t_0, T_1 > t_1 | z_0, z_1)$, which is difficult if the relatives are missing environmental covariates, we can model $\Pr(T_0 > t_0, T_1 > t_1)$, i.e.,

  $$\Pr(T_0 > t_0, T_1 > t_1) \quad = \quad h(S^*(t_0), S^*(t_1); \theta)$$

  where $S^*(t) = \exp\{-\Lambda^*(t)\}$ is a marginal composite survival function.
- We can again use the relationship

  $$\lambda_0(t) = \lambda(t)(1 - \phi(t))$$

Some Remarks

- Family history information is more than just a risk factor. We can use it to estimate $\lambda_0(t)$ which would have not been estimable from the case-controly data alone.
- Care must be taken on the accuracy of family history information, particularly if the information is reported by the cases and controls.
- For example, the NHLBI Family Heart Study shows that sensitivity of cases and controls report on their spouse, parent, and sibling was 87%, 85%, and 81% for coronary heart disease, 83%, 87%, and 72% for diabetes, 77%, 76%, and 56% for hypertension, respectively (Bensen et al., 1999). Most specificity values are above 90% .
- These results show that the accuracy may vary by the relative type and disease, but by and large the family history information is accurate.

More Remarks

- Two assumptions are worth noting:
  - The relatives and cases/controls have the same marginal hazard function, i.e., $\Pr(T_1 > t) = \Pr(T_0 > t)$.
    - Incorporate covariates (e.g., birth cohort).
  - It requires a correct specification of the copula function $h(\cdot)$. However, extensive simulations suggest that the marginal hazard function estimator is very robust against misspecification (Chatterjee et al. 2006; Hsu et al. 2007).
    - A more flexible form of the Copula model e.g., multiple parameters, flexible piece-wise constant cross ratio (Hougaard 2000; Hsu et al. 1999).
    - Nonparametric estimator of $\Pr(T > t)$ from case-control data (ongoing work).

Nonparametric estimation of $S(t) = \Pr(T > t)$

- Consider binary outcome. Let (d_0, d_1) be the disease status of the case-control and the relative.

	$d_1 = 1$	$d_1 = 0$
$d_0 = 1$	a	b
$d_0 = 0$	c	d

  Let $P_{1|1} = \Pr(d_1 = 1 | d_0 = 1)$ and $P_{1|0} = \Pr(d_1 = 1 | d_0 = 0)$, which can be estimated empirically by $a/(a+b)$ and $c/(c+d)$, respectively. By the law of total probability, we get

$$\Pr(d_1 = 1) = P_{1|1}\Pr(d_0 = 1) + P_{1|0}\Pr(d_0 = 0)$$

  If $p \equiv \Pr(d_0 = 1) = \Pr(d_1 = 1)$, then we have

$$\widehat{p} = \widehat{P}_{1|0}/(1 + \widehat{P}_{1|0} - \widehat{P}_{1|1})$$

Nonparametric estimation of $S(t)$

- Assume $S(t) = \Pr(T_0 > t) = \Pr(T_1 > t)$
- Let conditional survival functions $S_0(t|s) = \Pr(T_1 > t | T_0 > s)$ and $S_1(t|s) = \Pr(T_1 > t | T_0 = s)$. Both can be estimated by kernel estimators.
- Let $0 < t_1 < \cdots < t_Q < \tau < \bar{t}$ be a grid of time points such that $S(\bar{t}) > 0$. For $j = 1, \ldots Q$, we have

  $$\Pr(T_1 > u, T_0 > t_j) + \Pr(T_1 > u, T_0 = t_j) = \Pr(T_1 > u, T_0 \geq t_j).$$

  We can write it in terms of the conditional and marginal survival function,

  $$S_0(u|t_j)S(t_j) + S_1(u|t_j)\{S(t_{j-1}) - S(t_j)\} = S_0(u|t_{j-1})S(t_{j-1})$$

Nonparametric estimation of $S(t)$

- This gives a recursive estimator

$$S(t_j) = S(t_{j-1})\frac{S_0(u|t_{j-1}) - S_1(u|t_j)}{S_0(u|t_j) - S_1(u|t_j)}.$$

- This estimator is not defined if there is no dependency among relatives. However, if relatives are independent, we can simply use the Kaplan-Meier estimator to estimate $S(t)$.
- The formula holds for any $u \geq 0$, which suggests we can improve the estimator by pooling over u.
- $S(t)$ does not take jumps at the observed failure times of the relatives, but at pre-fixed grid points.

Simulation results

- Gamma frailty model (Kendall's tau = 0.6)
- n=1500 cases and controls. A total of 1,000 datasets were generated.

		Naive KM estimator			Proposed estimator		
t	$S(t)$	mean	SD	95%CI	mean	SD	95% CI
45	0.967	0.978	0.004	0.275	0.961	0.083	0.980
55	0.923	0.948	0.007	0.037	0.926	0.122	0.964
65	0.858	0.904	0.009	0.003	0.863	0.136	0.963
75	0.777	0.846	0.011	0.000	0.783	0.129	0.955

- It is a first attempt to provide a nonparametric survival estimator for biased samples. The theory is still murky.

Generalizability

Can one project the risk estimates to the population? Studies can differ from the population:

- Different hazard ratios
- Different risk factor distribution of Z
- Different baseline hazard function

Improving Risk Prediction: GWAS?

- Rich literature for the prediction (e.g., various machine learning approaches)
- For GWAS data, currently a simple additive model is the most popular.
  - Using known GWAS hits ($\alpha = 5 \times 10^{-8}$)

$$S = \sum_{m} \widehat{\beta}_m G_m$$

  - Since many causal variants have very small effect sizes, they wouldnt be significant at the genome-wide significance level. It may be better to include more SNPs than just top few SNPs.

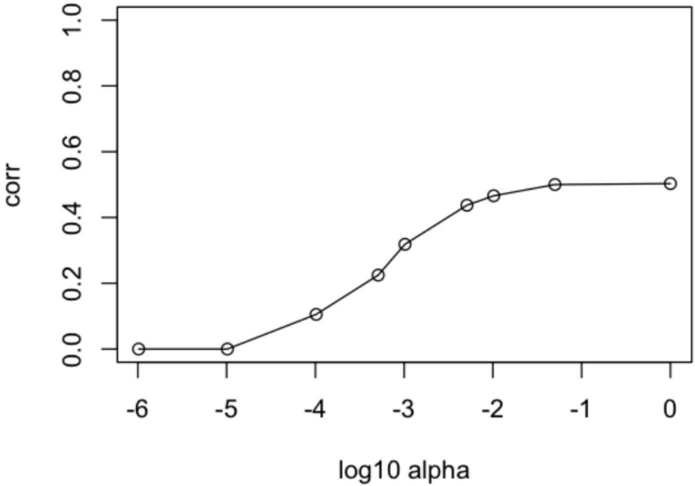
Polygenic Risk Score

- Simulation

$$Y = \sum_{m \in \text{causal}} \beta_m G_m + \epsilon$$

  - 100,000 independent SNPs (MAF=0.2)
  - Effect size follows an exp distribution
  - 1000 causal variants
  - Correlation of true score $\sum \beta_j G_{ij}$ and estimated score

$$\widehat{S} = \sum_{p_j < \alpha} \widehat{\beta}_j G_{ij}$$


corr
0.0
0.2
0.4
0.6
0.8
1.0
-6
-5
-4
-3
-2
-1
0
log10 alpha

Polygenic Risk Score (Chatterjee et al. 2013, NG)

- Model :

$$\text{logit}(\Pr(Y=1)) = \alpha + \sum_{m=1}^{M_1} + \sum_{n=M_1+1}^{M} 0 \times X_m$$

  - M: total number of variants
  - M_1: number of causal variants
  - X_m: standardized genotype value

- Estimated prediction model

$$\text{logit}(\Pr(Y=1)) = \widehat{\alpha} + \sum_{m=1}^{M} \widehat{\beta}_m \gamma_m X_m$$

where γ_m is the indicator of whether the variable is selected

Polygenic Risk Score

- Let $\widehat{U} = \sum_{m=1}^{M} \widehat{\beta}_m \gamma_m X_m$, $C_N = \sum_{m=1}^{M_1} \beta_m \widehat{\beta}_m \gamma_m = cov(\widehat{U}, U)$ and $S_N^2 = \sum_{m=1}^{M} \widehat{\beta}_m^2 \gamma_m$.

$$\widehat{U}|(Y=0) \sim N(0, S_N^2) \quad \text{and} \quad \widehat{U}|(Y=1) \sim N(C_N, S_N^2)$$

- AUC, i.e., the probability that risk-score will be greater for a randomly selected case than that of a randomly selected control, can be approximated by

$$AUC \approx \Phi(\frac{R_N}{\sqrt{2}}),$$

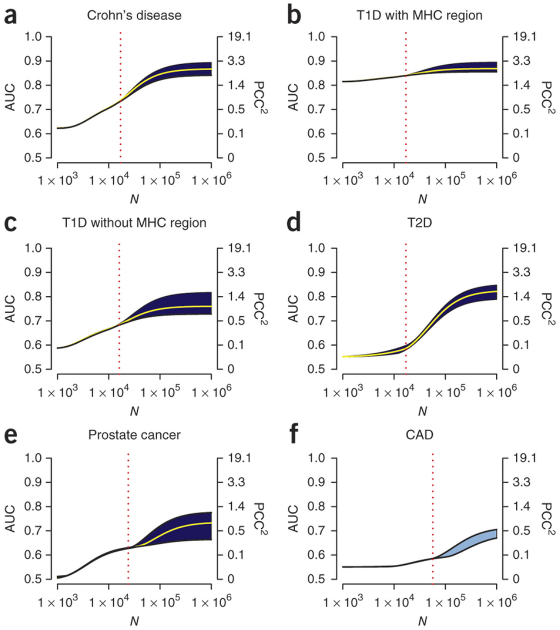
where $R_N = C_N / S_N = cor(Y, \widehat{Y})$.

Polygenic Risk Score

- Suppose the building algorithm is to include SNPs depending on whether the corresponding marginal trend-test for association achieves a specified significant level α or not.
- The expected value of R_N for such a building algorithm is approximated by

$$\mu_N(\alpha) = \frac{\sum_{m=1}^{M_1} \beta_m e_N(\beta_m) pow(N, \beta_m, \alpha)}{\sqrt{\sum_{m=1}^{M_1} v_N(\beta_m) pow(N, \beta_m, \alpha) + (M - M_1)\alpha v_N(0)}}$$

  - $pow(N, \beta_m, \alpha)$: power of the study of size N for detecting an effect size of β_m at α
  - $e_N(\beta_m) = E(\widehat{\beta}_m||Z_m| > C_{\alpha/2})$
  - $v_N(\beta_m) = E(\widehat{\beta}_m^2||Z_m| > C_{\alpha/2})$


a
Crohn's disease
b
T1D with MHC region
c
T1D without MHC region
d
T2D
e
Prostate cancer
f
CAD
AUC
1.0
0.9
0.8
0.7
0.6
0.5
PCC^2
19.1
3.3
1.4
0.5
0.1
0
1×10^3
1×10^4
1×10^5
1×10^6
N

Summary

- Models of absolute risk currently have a useful but limited role in counseling and in prevention.
- Efforts to increase discriminatory accuracy can expand that role. Will GWAS give us the boost we need for increasing the accuracy?
- Increased success in disease prevention will depend on safer and more effective interventions that may or may not need to be used in conjunction with risk models.

Recommended Reading

- Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH (2006). Case-control and case-only designs with genotype and family history data: Estimating relative risk, residual familial aggregation and cumulative risk. Biometrics 62: 36–48.

- Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature Genetics 45: 400-405.

- Chen YQ, Hu C, Wang Y (2006). Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics* 7, 515-29.

- Chen L, Lin DY, Zeng D (2010). Attributable fraction functions for censored event times. *Biometrika* 97, 713-26.

- Chen L, Hsu L, Malone K (2009). A frailty-model based approach to estimating the age-dependent function of candidate genes using population-based case-control study designs: An application to data on BRCA1 gene. Biometrics, 65: 1105-1114.

- Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C and Mulvihill JJ (1989). Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI 81: 1879–1886.

- Gail MH (2010). Personalized estimates of breast cancer risk in clinical practice and public health. Statistics in Medicine 30: 1090-1104

- Gorfine M, Zucker DM Hsu L (2009). Case-control survival analysis with a general semiparametric shared frailty model-A pseudo full likelihood approach. Annals of Statistics, 37, 1489-1517.

- Hsu L and Gorfine M (2006). Multivariate survival analysis for case-control family data. Biostatistis, 7: 387-98.

- Benichou J (2001). A review of adjusted estimators of attributable risk. Statistical Methods in Medical Research 10: 195216

- Liu D, Zheng Y, Prentice RL, Hsu L (2014). Estimating risk with time-to-event data: An application to the Women's Health Initiative. Journal of American Statistical Association 109(506):514-524.