Absolute Risk: Methods and Applications in Clinical Care and Public Health, Day 1 Materials

Ruth Pfeiffer

Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
National Institutes of Health
Bethesda, Maryland, USA
pfeiffer@mail.nih.gov

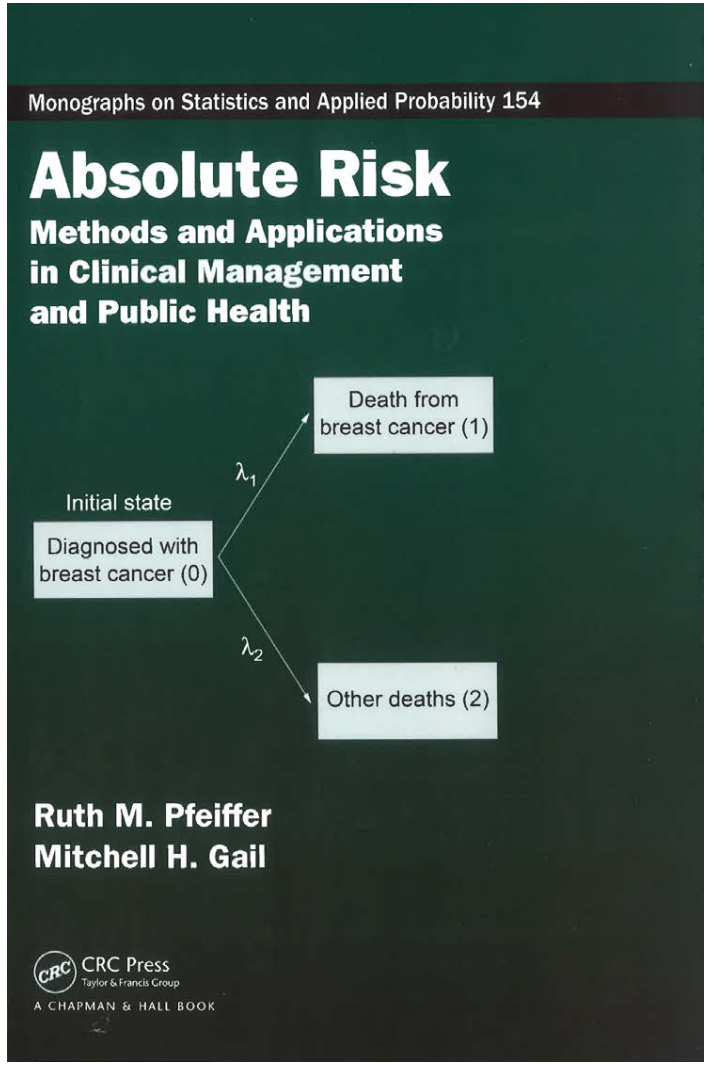
SISCER Module 12, 2024

Course Goals/Summary

- Learn what absolute risk (or crude risk or cumulative incidence) is
- Learn how to estimate it from data from various designs
- Learn how to assess the validity and usefulness of an absolute risk model
- Learn what it can be used for

Some Books on Absolute Risk/Competing Risks

- Kalbfleisch & Prentice. The Statistical Analysis of Failure Time Data, 2002
- Andersen, Borgan, Gill & Keiding. Statistical Models Based on Counting Processes, 1993
- Van Houwelingen & Putter. Dynamic prediction in clinical survival analysis. 2012
- Beyersmann, Schumacher & Allignol. Competing risks and multistate models with R. 2012
- Geskus. Data analysis with competing risks and intermediate states. 2016
- Pfeiffer & Gail. Absolute risk: methods and applications in clinical management and public health. 2018



https://www.routledge.com/Absolute-Risk-Methods-and-Applications-in-Clinical-Management-and-Public/Pfeiffer-Gail/p/book/9780367657819

Outline Day 1

1. Definitions of "risk"
2. Brief review of survival analysis
3. Competing risks
4. Estimation of absolute risk
   - Without covariates from cohorts
   - With covariates from cohort data
   - By combining observational studies with registry data
5. Assessment of risk model performance

Outline Day 2

1. Assessment of risk model performance
   - Assessing calibration with missing predictors
   - Comparing two models
2. Applications of absolute risk
3. Miscellaneous topics:
   - Updating risk models

Definitions of "Risk": Relative Risk

Ratio of probability of event in exposed versus a non-exposed group; epidemiologic/etiologic research
35 year-old woman with

- two sisters with breast cancer
- no children
- age 12 years at first period
- other risk factors at lowest risk level

Relative risk RR =5.39 compared to woman at lowest level of risk for ALL risk factors

Definitions of "Risk": Probability of Outcome

Clinical and public health applications

- Prevalence: P(condition present in patient)
- Probability of future event of cause 1 (e.g. cancer)

Probability of outcome: prevalence models

Prevalence: P(condition present in patient)

$Y = 1, 0$ condition or disease present (1)/absent (0)

modeled by e.g. logistic regression as function of factors $\mathbf{X} = (X_1, \ldots, X_p)$

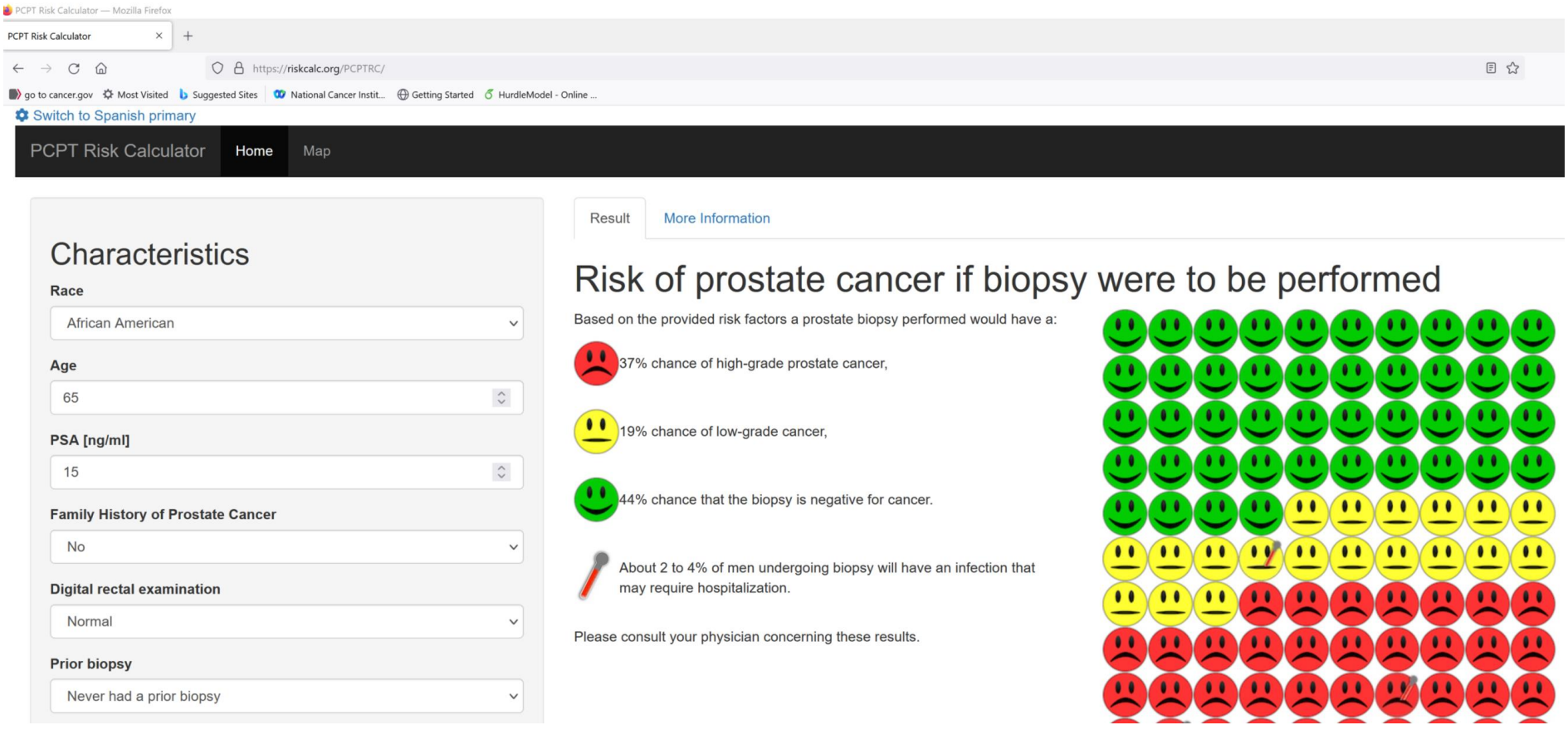
$$P(Y = 1|X_1, \ldots, X_p) = \frac{\exp(\beta_0 + \boxed{\beta_1 X_1 + \ldots + \beta_p X_p})}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

$\beta_1 X_1 + \ldots + \beta_p X_p$: " risk score"

β_0 captures disease prevalence for person with all $X = 0$

Example: Prevalence Model for Prostate Cancer

`https://riskcalc.org/PCPTRC`(Thompson et al, JNCI, 2006)



Definitions of "Risk": Probability of Outcome

Clinical and public health applications

- Prevalence: P(condition present in patient)
- Probability of **future event of cause 1** (e.g. cancer) :
  T: time to event
  - **"Pure risk":**
    $P(a < T \leq a + \tau, \text{cause=cancer} | T > a$, there are no competing risks)

Definitions of "Risk": Probability of Outcome

Clinical and public health applications

- Prevalence: P(condition present in patient)
- Probability of **future event of cause 1** (e.g. cancer) :
  T: time to event
  - **"Pure risk":**
    $P(a < T \leq a + \tau, \text{cause=cancer} | T > a$, there are no competing risks)
  - **Absolute risk:** (or "crude risk", or "cumulative incidence")
    $P(a < T \leq a + \tau, \text{cause=cancer} | T > a$, there are competing risks)

"Risk" Definitions: Probability of future event of cause 1:

Clinical and public health applications
T: time to event (e.g. breast cancer diagnosis)

- **"Pure risk":**
  $P(a < T \leq a + \tau, \text{cause}{=}1 | T > a$, there are no competing risks)

Survival Concepts and Definitions

$T \geq 0$ *survival* or *event time*

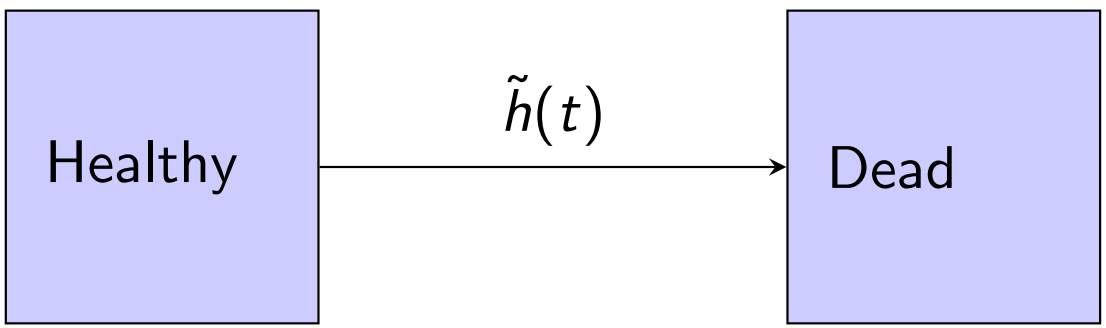
$S(t) = P(T > t)$ **survival function**

$f(t) = \lim_{\epsilon \downarrow 0} \frac{P(t \leq T < t+\epsilon)}{\epsilon}$ density function

$\tilde{h}(t) = \lim_{\epsilon \downarrow 0} \frac{P(t \leq T < t+\epsilon \mid T \geq t)}{\epsilon}$ **hazard function**

$\tilde{h}$ is infinitesimal probability that if person survives to t he/she will have event in next instant

Hazard $\tilde{h}(u)$ can be interpreted as transition rate between states



Survival Concepts and Definitions

$T \geq 0$ *survival* or *event time*

$S(t) = P(T > t)$ **survival function**

$f(t) = \lim_{\epsilon \downarrow 0} \frac{P(t \leq T < t+\epsilon)}{\epsilon}$ density function

$\tilde{h}(t) = \lim_{\epsilon \downarrow 0} \frac{P(t \leq T < t+\epsilon \mid T \geq t)}{\epsilon}$ **hazard function**

$\tilde{h}$ is infinitesimal probability that if person survives to t he/she will have event in next instant

Key relationship:

$$\tilde{h}(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log\{S(t)\}}{dt} \text{ and as } S(0) = 1$$

$$S(t) = \exp\left\{-\int_0^t \tilde{h}(u)du\right\}$$

Why Special Methods for Estimating S(t) are Needed: Incomplete Data (Censoring, Truncation)

- **(Right) Censoring:** event times are not observed exactly, only known to fall in certain time interval
- **Reasons for censoring:**
  - individual drops out of study (lost to follow up)
  - study ends before person have event
  - individual has event that precludes having event of interest (e.g. dies from heart attack before cancer is diagnosed)
- **Ignoring censoring leads to biased estimates** of distribution of survival time and related quantities
- **(Left) Truncation:** person had to survive past certain time to get into study

Estimation of S

- Survival function can be determined directly from hazard function

$$S(t) = \exp\left\{-\int_0^t \tilde{h}(u)du\right\}$$

- Methods for estimating hazard function from sample of event times
  - Nonparametric
  - Parametric
  - Semiparametric

Non-parametric Estimate of $S(t)$: Kaplan-Meier Estimate

- $t_1 < t_2 < \ldots < t_d$ are ordered times when events occur
- n_j number of subjects still under follow-up at t_j
- d_j number of events at t_j

Probability of being alive at the t calculated iteratively from conditional survival to t_i:

$$P(T > t_j | T > t_{j-1}) = \frac{n_j - d_j}{n_j}$$

as

$$S^{KM}(t) = \prod_{j:t_j \leq t} P(T > t_j | T > t_{j-1}) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j:t_j \leq t} \left\{ 1 - \frac{d_j}{n_j} \right\}$$

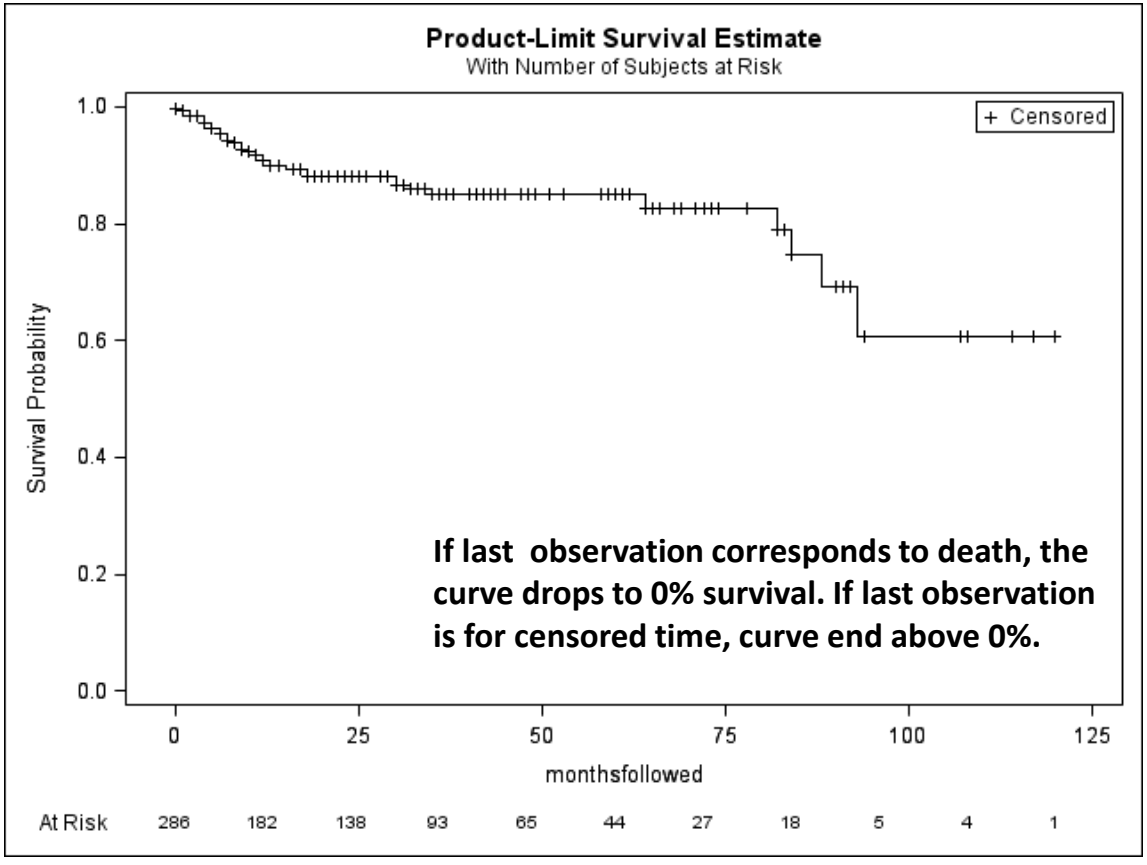
Non-parametric Estimate of $S(t)$: Kaplan-Meier Estimate

- $t_1 < t_2 < \ldots < t_d$ are ordered times when events occur
- n_j number of subjects still under follow-up at t_j
- d_j number of events at t_j
- Kaplan-Meier survival function estimate: $S^{KM}(t) = \prod_{j:t_j \leq t} \left\{1 - \frac{d_j}{n_j}\right\}$

Notice, estimate of hazard is $\hat{h}(t_j) = P(\text{event at } t_j | T > t_{j-1}) = \frac{d_j}{n_j}$
Use $\exp(-x) \approx 1 - x$ to get

$$S^{KM}(t) = \prod_{j:t_j \leq t} \left\{1 - \frac{d_j}{n_j}\right\} \approx \exp(-\sum_{j:t_j \leq t} \frac{d_j}{n_j}) = \exp\left\{-\sum_{j:t_j \leq t} \hat{h}(t_j)\right\}$$

Non-parametric Estimate of $S(t)$: Kaplan-Meier Estimate



1

Incorporating Covariates into Hazard Function

- Covariates $\mathbf{X} = (X_1, \ldots, X_p)$
- Model hazard function as

$$\tilde{h}(t|\mathbf{X}) = h_0(t) rr(\mathbf{X})$$

- $h_0(t)$ baseline hazard function (when all $X_i = 0$)
- $h_0(t)$ unspecified or assume parametric form (e.g. exponential)
- $rr(\mathbf{X})$ relative risk term, e.g. $rr(\mathbf{X}) = \exp(\beta_1 X_1 + \ldots + \beta_p X_p)$ **Cox proportional hazards model** (Cox 1972)
- $\frac{h(t|X_i)}{h(t|X_i=0)} = \frac{h_0(t)\exp(\beta_i X_i)}{h_0(t)} = \exp(\beta_i)$ is **hazard ratio for** X_i:

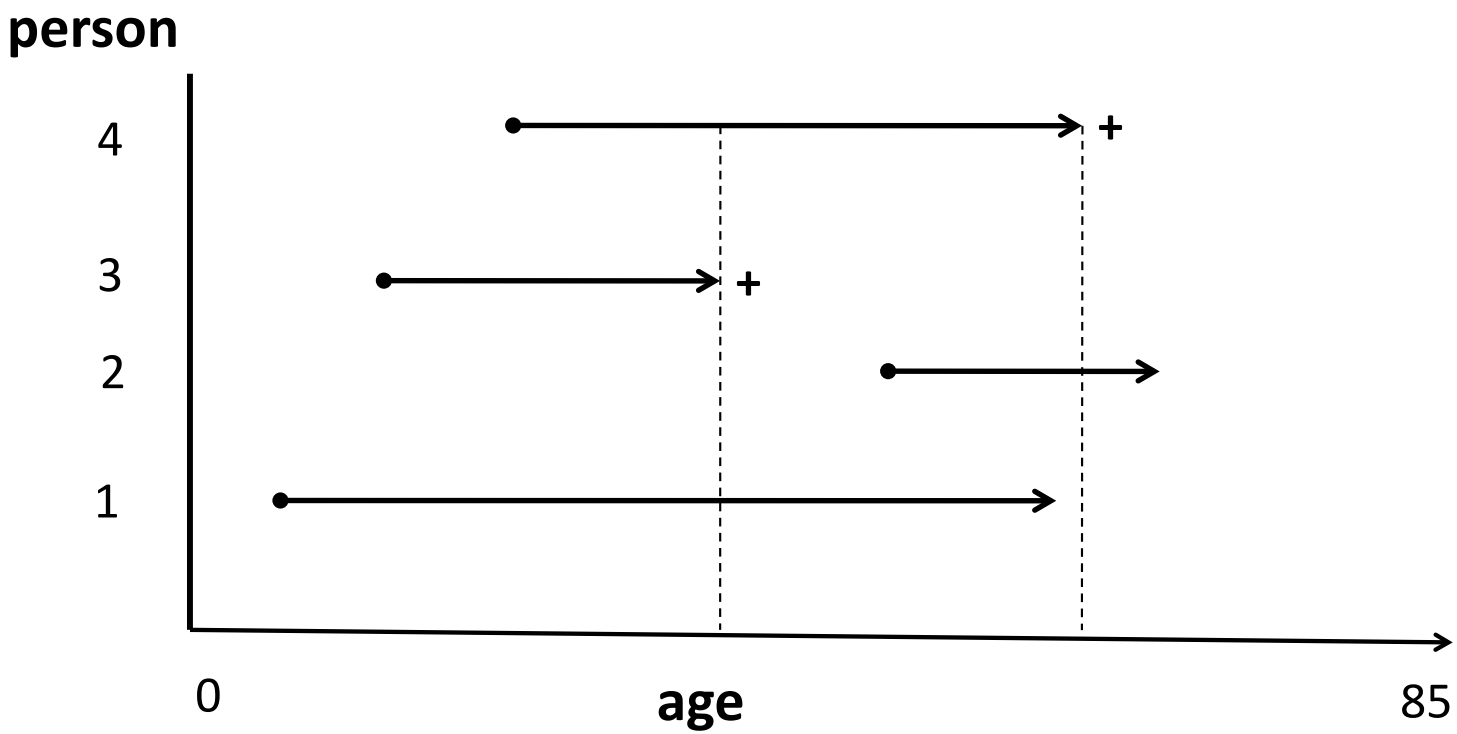
"Risk" Definitions: Pure risk for Prognosis or Predicting Incidence

$$P(a < T \leq a + \tau, \text{cause=1} | T > a, \mathbf{X}, \text{there are no competing risks})$$
$$= S(a + \tau | \mathbf{X}) / S(a | \mathbf{X}) = \exp\left\{ - \int_a^{a+\tau} \tilde{h}(u | \mathbf{X}) du \right\}$$

Choice of Time Scale

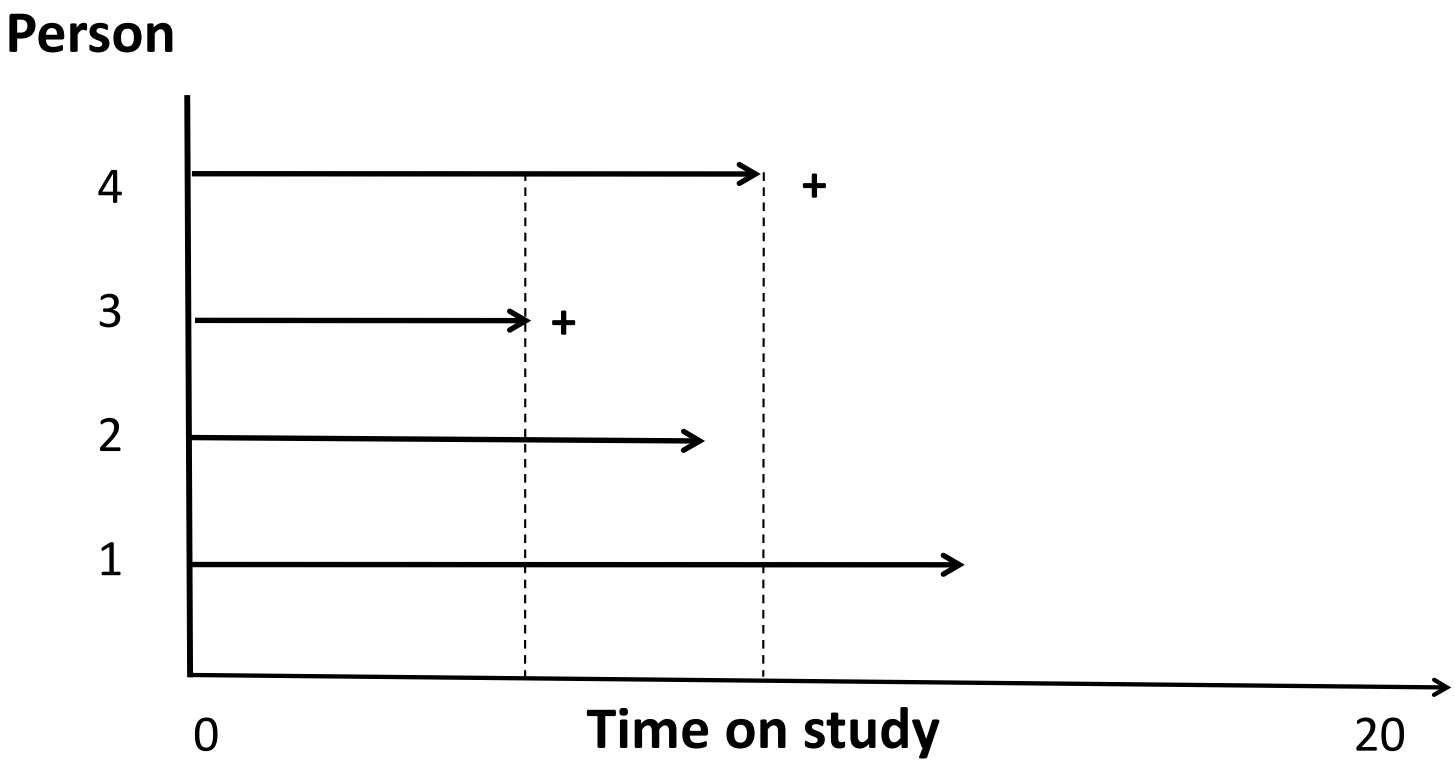
- Age
  - Important predictor of incidence
  - Need to account for left-truncation and right censoring
- Time on study
  - Requires careful modeling of age as covariate for predicting incidence
  - Better than age for predicting e.g. death from cancer following diagnosis
  - Only right censoring
- Calendar time
  - Requires careful modeling of age as covariate for predicting incidence

Choice of Time Scale: Age



Bullet indicates start of follow up, i.e. age at entry into study. End of follow up: age at event or censoring subject to left truncation and right censoring

Choice of Time Scale: Time on Study



Multiple Types of Events

- Pure risk: one type of event, censoring
- **Absolute risk: different (competing) types of events**: e.g. breast cancer diagnosis, death from other causes

"Risk" Definitions: Probability of future event of cause 1:

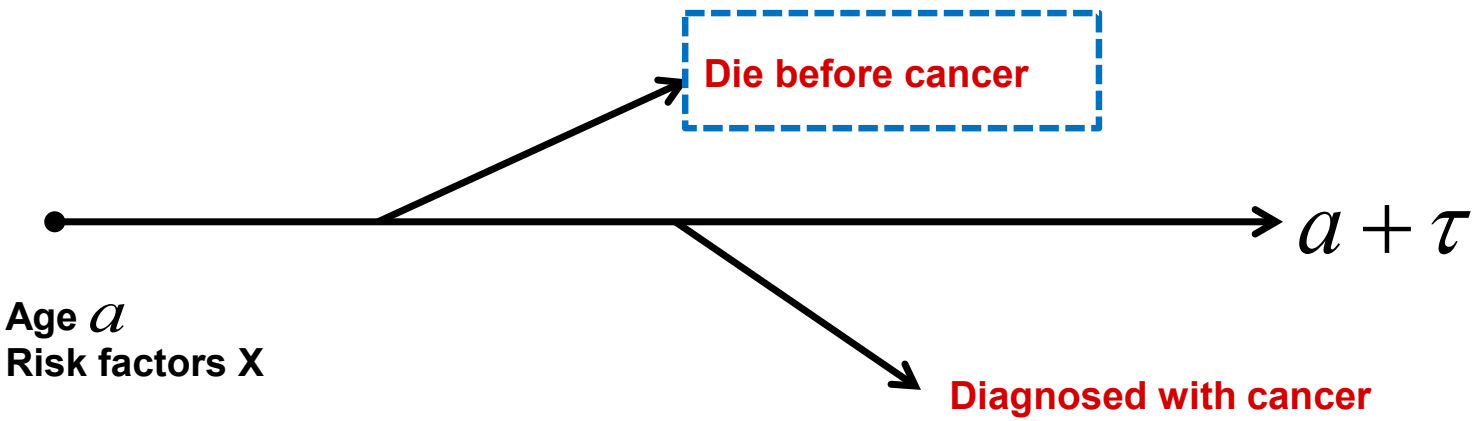
Clinical and public health applications
T: time to event

- **"Pure risk":**
  $P(a < T \leq a + \tau, \text{cause}{=}1 | T > a$, there are no competing risks)
- **Absolute risk:** (or "crude risk", or "cumulative incidence")
  $P(a < T \leq a + \tau, \text{cause}{=}1 | T > a$, there are competing risks)

Absolute Risk

T: time to event

$R(a, a+\tau, \mathbf{X}) = P(a < T \leq a+\tau, \text{cause=1} | T > a,$ there are competing risks)



Cause Specific Hazard (! Math)

Have M event types, observe $T = \min(T_1, \ldots, T_M)$
T_i have joint survival function $S(t_1, \ldots, t_M) = P(T_1 > t_1, \ldots, T_M > t_M)$
Pure (net) hazard function for cause m,

$$\tilde{h}_m(t) = \lim_{\epsilon \downarrow} \frac{P\left(t \leq T < t + \epsilon \middle| T \geq t, \text{ no other causes acting}\right)}{\epsilon}$$

$$\tilde{h}_m(t) = -\frac{\partial}{\partial t_m} \log\{S(t_1, \ldots, t_M)\}|_{t_1=0,\ldots,t_m=t,=t_M=0}$$

Cause Specific Hazard (! Math)

Have M event types, observe $T = \min(T_1, \ldots, T_M)$
T_i have joint survival function $S(t_1, \ldots, t_M) = P(T_1 > t_1, \ldots, T_M > t_M)$
Pure (net) hazard function for cause m,

$$\tilde{h}_m(t) = \lim_{\epsilon \downarrow} \frac{P\left(t \leq T < t + \epsilon \middle| T \geq t, \text{ no other causes acting}\right)}{\epsilon}$$

$$\tilde{h}_m(t) = -\frac{\partial}{\partial t_m} \log\{S(t_1, \ldots, t_M)\}|_{t_1=0,\ldots,t_m=t,=t_M=0}$$

Cause-specific hazard function

$$h_m(t) = \lim_{\epsilon \downarrow 0} \frac{P\left(t \leq T < t + \epsilon, \text{ cause} = m \middle| T \geq t\right)}{\epsilon}$$

$$h_m(t) = -\frac{\partial}{\partial t_m} \log\{S(t_1, \ldots, t_M)\}|_{t_1=\ldots=t_M=t}$$

Cause Specific Hazard, cont. (! Math)

- $S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
  $\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\} = S(t)$ estimable from observed data

Cause Specific Hazard, cont. (! Math)

- $S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
  $\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\} = S(t)$ estimable from observed data
- Cause-specific hazard $h_m(t)$ estimable because can observe instantaneous risk of event m among those at risk at t

Cause Specific Hazard, cont. (! Math)

- $S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
  $\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\} = S(t)$ estimable from observed data
- Cause-specific hazard $h_m(t)$ estimable because can observe instantaneous risk of event m among those at risk at t
- Any function that depends only on h_m is estimable

Cause Specific Hazard, cont. (! Math)

- $S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
  $\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\} = S(t)$ estimable from observed data
- Cause-specific hazard $h_m(t)$ estimable because can observe instantaneous risk of event m among those at risk at t
- Any function that depends only on h_m is estimable
- Joint survival function $S(t_1, \ldots, t_M)$ only identifiable under parametric models or under assumption that T_m independent

Cause Specific Hazard, cont. (! Math)

- $S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
  $\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\} = S(t)$ estimable from observed data
- Cause-specific hazard $h_m(t)$ estimable because can observe instantaneous risk of event m among those at risk at t
- Any function that depends only on h_m is estimable
- Joint survival function $S(t_1, \ldots, t_M)$ only identifiable under parametric models or under assumption that T_m independent
- If all T_m independent, $h_m = \tilde{h}_m$

Related Survival Functions

Joint survival function

$S(t_1, \ldots, t_M) = P(T_1 > t_1, \ldots, T_M > t_M) =$ only identifiable under parametric models or assuming that T_m independent

Overall survival function:

recall $T = \min(T_1, \ldots, T_M)$

$P(T > t) = S(t, \ldots, t) = P(T_1 > t, \ldots, T_M > t) =$
$\exp\left\{-\int_0^t \sum_{m=1}^M h_m(u)du\right\}$ estimable

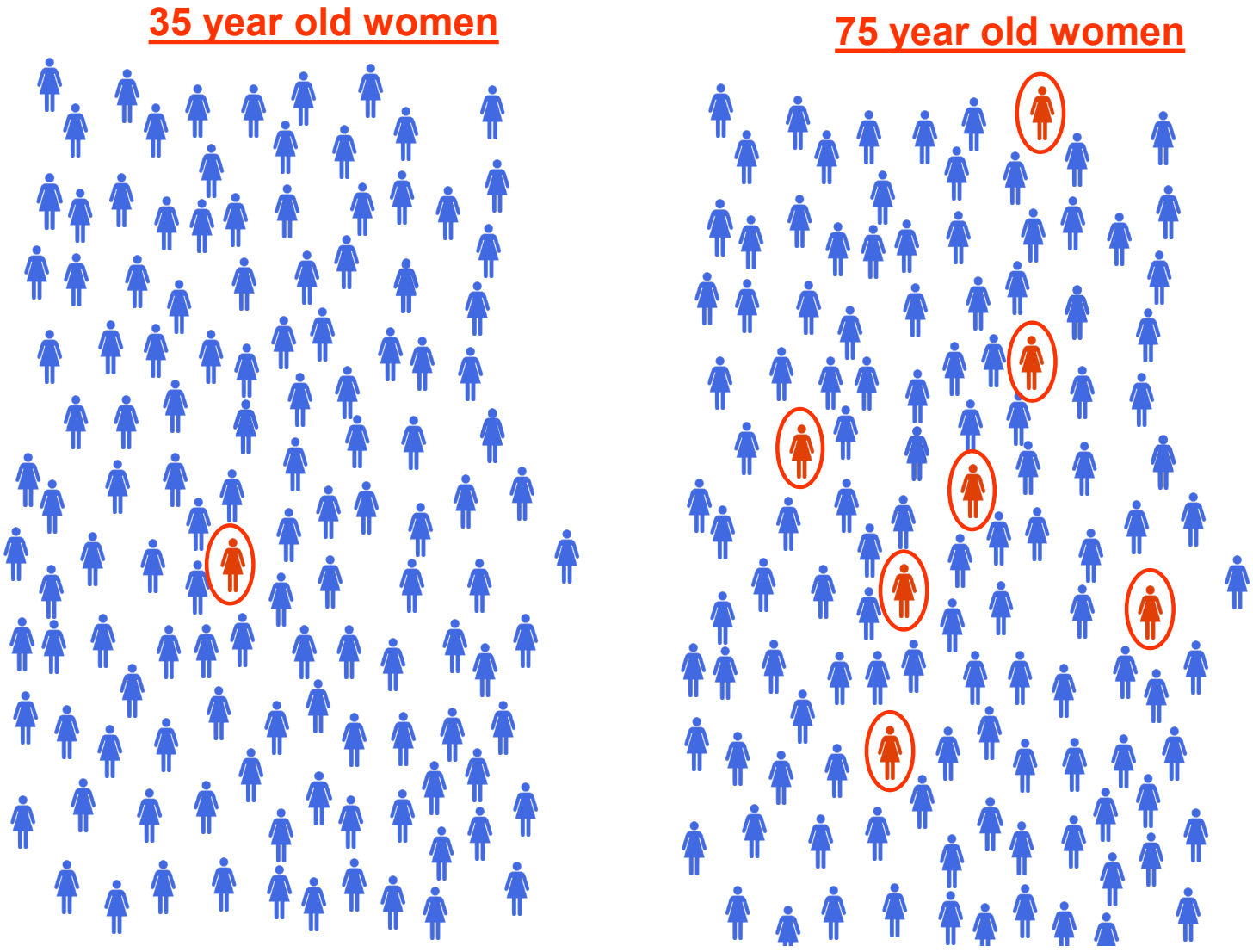
absolute risk for cause 1:
$P(T < t, \text{cause=1}) = \int_0^t h_1(s, \mathbf{X}) \exp\left[-\int_0^s \sum_{m=1}^M h_m(u)du\right] ds$

Absolute versus Relative Risk

Woman with two sisters with breast cancer, no children, age 12 years at first period: **RR=5.39**

Current age	5-Year absolute breast cancer risk
35	1.1 %
75	6.7 %

Examples of 5-year Absolute Risk of Breast Cancer



Absolute versus Pure Risk

Absolute ("Crude") and "Pure" Risk in 1000 60-Year Old Women

Age at start of interval	# at risk	# incident breast cancers	# Deaths from # other causes
60	1000	17	44
65	939	20	63
70	856	22	89
75	745	...	...

Absolute risk of breast cancer to age 75 is $\frac{17+20+22}{1000} = 5.9\%$
Pure risk of breast cancer to age 75 is
$1 - (1 - 17/1000)(1 - 20/939)(1 - 22/856) = 6.3\%$

Absolute versus Pure Risk

- Absolute risk: no competing risk assumptions like "independence"
- Absolute risk clinically relevant because eliminating other deaths is not realistic (absolute risk $<$ pure risk as competing causes reduce risk of getting cancer)
- Absolute risk nearly equals pure risk if competing causes are rare (e.g. short prediction intervals)
- Pure risk has etiologic interest as a description related to cumulative cause-specific hazard

Some Cancer Risk Models

- Cancer incidence
  - `https://bcrisktool.cancer.gov/`(breast BCRAT)
  - `https://ccge.medschl.cam.ac.uk/boadicea/` (breast BOADICEA)
  - `https://ibis.ikonopedia.com/` (breast IBIS)
  - `https://ccrisktool.cancer.gov/` (colorectal cancer)
  - `https://mrisktool.cancer.gov/` (melanoma)
  - Other models: Bladder, Endometrial cancer, Lung, Ovary, Pancreas, Prostate
- Absolute risk of death from prostate cancer after diagnosis (Albertson, Hanley, Fine, JAMA 2005)

Some Choices in Risk Modeling

- Genetic model versus empirical model
- Choice of risk factors
  - Detailed family history
  - Reproductive history (e.g. age at first live birth)
  - Medical/lab history (e.g. biopsies, mammographic density, SNPs)
- Data sources and "piecing together" the model
- Target population: e.g. general population in UK or in US; or high-risk clinic

Genetically-based Models

- Autosomal dominant
  - Use extensive family history and BRCA1/2 data
  - BRCAPRO (Berry et al, JNCI 1997)
  - Claus Model (Claus et al, Cancer,1994)
- Autosomal dominant & residual familial effects
  - BOADICEA, Antoniou et al, BJC 2008
  - IBIS, Tyrer, Duffy and Cuzick, Stat Med 2005; This model includes other factors such as LCIS, age at first live birth.

Estimating Absolute Risk

Estimating Absolute Risk from Cohort Data, no Covariates

Gaynor et al, JASA 1993

- $t_1 < t_2 < \ldots < t_d$ are ordered times when events occur
- d_{ij} number of events due to cause $i = 1, 2$ at t_j
- $d_j = d_{1j} + d_{2j}$
- n_j number of subjects still under follow-up at t_j

Maximum likelihood estimate of cause specific failure probability is

$$\hat{R}(t) = \hat{P}(T \leq t, cause = 1) = \sum_{j:t_j \leq t} \frac{d_{1j}}{n_j} \hat{S}^{KM}(t_{j-1}) = \sum_{j:t_j \leq t} \hat{h}_{1j} \hat{S}^{KM}(t_{j-1})$$

Absolute versus Pure Risk

Absolute ("Crude") and "Pure" Risk in 1000 60-Year Old Women

Age at start of interval	# at risk	# incident breast cancers	# Deaths from # other causes
60	1000	17	44
65	939	20	63
70	856	22	89
75	745	...	...

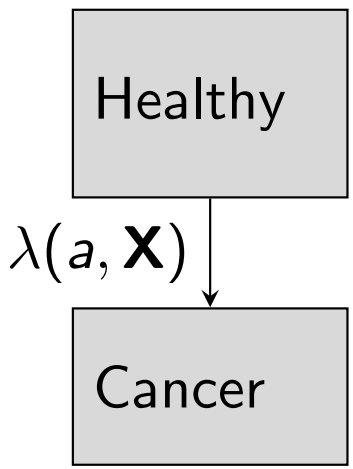
Absolute risk of breast cancer to age 75 is
$17/1000 + 20/939 * (939/1000) + 22/856 * (856/1000)\frac{17+20+22}{1000} = 5.9\%$

Modeling Absolute Risk as Function of $\mathbf{X}$

Modeling Absolute Risk as Function of $\mathbf{X}$: Two Approaches

T: time to event, $\mathbf{X}$: risk factors

$R(a, a+\tau, \mathbf{X}) = P(a < T \leq a+\tau, \text{cause}{=}1 \,|\, T > a, \mathbf{X})$



Cumulative incidence regression
(Fine & Gray, JASA 1999)

Cumulative Incidence Regression, (Fine & Gray, JASA '99)

Directly model cumulative incidence function of cause 1 as function of covariates:

$$R(a, a+\tau, \mathbf{X}) = P(a < T \leq a+\tau, \text{cause=1} \,|\, T > a, \mathbf{X})$$

R is a sub-distribution function, i.e. a non-decreasing function of time with

$$R(\infty) = P(\text{cause=1})$$

Cumulative Incidence Regression (Fine & Gray, JASA '99)

Model $R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1 | \mathbf{X})$ using link function g as

$$g\{R(t, \mathbf{X})\} = \psi_0(t) + \boldsymbol{\gamma}' \mathbf{X}$$

- ψ_0 unspecified invertible monotone increasing function (captures baseline failure probability)
- $\boldsymbol{\gamma}$ regression coefficients for covariates $\mathbf{X}$

Fine and Gray '99 used $g(t) = \log\{\log(1-t)\}$
incorporated $\mathbf{X}$ via proportional hazards model for *sub-distribution hazard*

$$\lambda(t, \mathbf{X}) = -\frac{d \log\{1 - R(t, \mathbf{X})\}}{dt} =$$
$$\lim_{\epsilon \downarrow 0} \frac{P\{t \leq T < t + \epsilon, \text{cause} = 1 |\ T \geq t \cup (T \leq t \cap \text{cause} \neq 1), \mathbf{X})\}}{\epsilon}$$

Cumulative Incidence Regression (Fine & Gray, JASA '99)

Model $R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1|\mathbf{X})$ using link function g as

$$g\{R(t, \mathbf{X})\} = \psi_0(t) + \boldsymbol{\gamma}'\mathbf{X}$$

Fine and Gray '99 used $g(t) = \log\{\log(1-t)\}$
incorporated $\mathbf{X}$ via proportional hazards model for *sub-distribution hazard*

$$\lambda(t, \mathbf{X}) = -\frac{d\log\{1 - R(t, \mathbf{X})\}}{dt} = \lambda_0(t)\exp(\boldsymbol{\gamma}'\mathbf{X})$$

where $\lambda_0(t) \geq 0$ is unspecified
Then $\psi_0(t) \equiv \log\{\int_0^t \lambda_0(t)(s)ds\}$ and

$$R(t, \mathbf{X}) = 1 - \exp\left\{-\int_0^t \lambda(s, \mathbf{X})ds)\right\}$$

If binary X has coefficient $\gamma > 0$ then $R(t, X)$ higher for those with $X = 1$

Cumulative Incidence Regression: Estimation

$$R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1 | \mathbf{X}) = 1 - \exp\left\{ - \exp(\gamma' \mathbf{X}) \int_0^t \lambda_0(s) ds \right\}$$

- partial likelihood with inverse probability of censoring weighting (IPCW; Robins & Rotznitzky, 1992)

Cumulative Incidence Regression: Estimation

$$R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1 | \mathbf{X}) = 1 - \exp\left\{ - \exp(\gamma' \mathbf{X}) \int_0^t \lambda_0(s) ds \right\}$$

- partial likelihood with inverse probability of censoring weighting (IPCW; Robins & Rotznitzky, 1992)
- **Different risk sets compared to cause specific partial likelihood:** those who failed from other causes and those who have not failed from cause 1 are at risk

Cumulative Incidence Regression: Estimation

$$R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1 | \mathbf{X}) = 1 - \exp\left\{ - \exp(\gamma' \mathbf{X}) \int_0^t \lambda_0(s) ds \right\}$$

- partial likelihood with inverse probability of censoring weighting (IPCW; Robins & Rotznitzky, 1992)
- **Different risk sets compared to cause specific partial likelihood:** those who failed from other causes and those who have not failed from cause 1 are at risk
- $\hat{\gamma}$ estimated from weighted score function

Cumulative Incidence Regression: Estimation

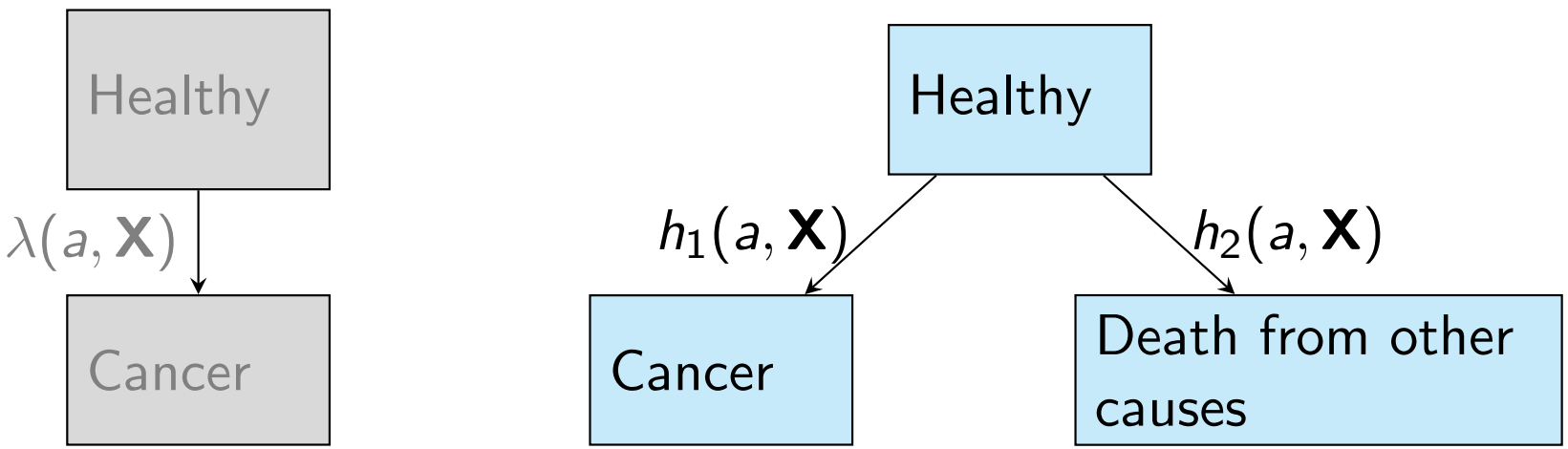
$$R(t, \mathbf{X}) = P(T \leq t, \text{cause}{=}1 | \mathbf{X}) = 1 - \exp\left\{ - \exp(\gamma' \mathbf{X}) \int_0^t \lambda_0(s) ds \right\}$$

- partial likelihood with inverse probability of censoring weighting (IPCW; Robins & Rotznitzky, 1992)
- **Different risk sets compared to cause specific partial likelihood:** those who failed from other causes and those who have not failed from cause 1 are at risk
- $\hat{\gamma}$ estimated from weighted score function
- $\hat{\Lambda}(t, \mathbf{X}) = \int_0^t \lambda(s, \mathbf{X}) ds)$ estimated using modified Breslow estimator
- $\hat{R}(t, \mathbf{X})$ asymptotically normal

Implemented in STATA and *R* package *cmprsk*; left truncation with time-dependent weights in R-package *R* package *mstate* de Wreede et al, 2001)

Modeling and Estimating Absolute Risk: Two Approaches

T: time to event, $\mathbf{X}$: risk factors
$R(a, a+\tau, \mathbf{X}) = P(a < T \leq a+\tau, \text{cause=1} | T > a, \mathbf{X})$



Cumulative incidence regression (Fine & Gray, JASA 1999)
Cause specific approach: model $h_m(a, \mathbf{X}), m = 1, 2$

Absolute Risk Model: Cause-specific Formulation

Focus on 2 types of events: 1=cancer, 2=death from other causes
Observe $T=$ time to first event

$$h_m(t) = \lim_{\epsilon \downarrow 0} \frac{P\left(t \leq T < t+\epsilon,\ \text{cause} = m | T \geq t\right)}{\epsilon}, m = 1, 2$$

Thus probability of no event in $[0, t]$ is

$$P(T > t) = S(t) = \exp\left[-\int_0^t \{h_1(u) + h_2(u)\} du\right]$$

where S is overall survival function

Absolute Risk Model: Cause-specific Formulation

$$R(a, a+\tau, \mathbf{X}) = P(T \leq a+\tau, \text{cause}=1 \,|\, T > a, \mathbf{X}) = \int_a^{a+\tau} h_1(t, \mathbf{X}) \exp\left[- \int_a^t \{h_1(v, \mathbf{X}) + h_2(v, \mathbf{X})\} \mathrm{d}v \right] \mathrm{d}t$$

- T event time
- $h_1(t, \mathbf{X})$ - hazard for cancer
- $h_2(t, \mathbf{X})$ - hazard for competing events (e.g. death)
- $\mathbf{X}$ - individual risk/protective factors
- τ - projection period

General Strategy to Estimate Cause Specific Absolute Risk

- Model cause specific hazards: $h_i(t, \mathbf{X}) = h_{0i}(t) rr(\beta_i' \mathbf{X}), i = 1, 2$

  h_{0i} - baseline hazard for cause i
  $rr(\beta_i' \mathbf{X}) = \exp(\beta_i' \mathbf{X})$ (standard Cox regression model)

- Estimate cause specific hazards $\hat{h}_1(t, \mathbf{X})$, $\hat{h}_2(t, \mathbf{X})$
- Obtain "plug in" estimate

$$\hat{R}(a, a+\tau, \mathbf{X}) = \int_a^{a+\tau} \hat{h}_1(t, \mathbf{X}) \exp\left[- \int_a^t \{\hat{h}_1(s, \mathbf{X}) + \hat{h}_2(s, \mathbf{X})\} \mathrm{d}s \right] \mathrm{d}t$$

- For cohort data *R* package *riskRegression* implements regression models for survival analysis with competing risks

Cox Proportional Hazards Model for Full Cohort

$h_i(t, \mathbf{X}) = h_{0i}(t) \exp(\beta_i' \mathbf{X}), i = 1, 2$

- Benichou & Gail ('90) gave variance of absolute risk estimate $\hat{R}(t, \mathbf{X})$ for
  - constant $h_{0i}(t)$
  - piecewise constant $h_{0i}(t)$
  - non-parametric $h_{0i}(t)$
- Cheng, Fine & Wei ('98): confidence bands for $\hat{R}(t, \mathbf{X})$
- **Remark:** Non-parametric estimation of absolute risk can be much less efficient (variance ratio 4) than parametric estimation

Semi-parametric Estimate of Absolute Risk

When no distributional assumption is made for $h_{0m}(t)$

$$\hat{h}_{0m}(t_j) = \frac{\sum_{i=1}^{n} \delta_{mi}(t)}{\sum_{i \in R(t_j)} \exp(\hat{\beta}'_m \mathbf{X}_i^m)}, m = 1, 2$$

where $t_1 \leq t_2 \leq \ldots$ are observed event times (possibly tied), $\delta_{kj} = 1$ if event at t_j is of type k; $R(t_j)$ risk set at t_j; $\hat{\beta}_m$ is MLE from Cox model
Nelson-Aalen estimate of cumulative cause-specific baseline hazard

$$\hat{H}_{0m}(t) = \sum_{i=1}^{n} \int_0^t \hat{h}_{0m}(s)ds = \sum_{t_i^m \leq t} \hat{h}_{0m}(t_i^m)$$

t^m observed event times for mth event type

$$\hat{H}_m(t; \mathbf{X}_m) = \sum_{t_i^m \leq t} \hat{h}_{0m}(t_i^m) \exp(\hat{\beta}'_m \mathbf{X}_m) = \hat{H}_{0m}(t) \exp(\hat{\beta}'_m \mathbf{X}_m)$$

Semi-parametric Estimate of Absolute Risk, cont.

semi-parametric estimate of absolute risk of event type 1 within interval $[t_0, t_1)$, given $\mathbf{X}$ and given no events until t_0 is

$$\hat{R}(t_0, t_1; \mathbf{X}) = \exp(\hat{\beta}_1' \mathbf{X}^1) \sum_{t_0 \leq t_j < t_1} \hat{h}_{0m}(t_j) \exp\{-\sum_{m=1}^{2} \hat{H}_m(t_j; \mathbf{X}^m)\},$$

where t_j are observed event times of type 1 occurring in $[t_0, t_1)$

Comments

- Estimate absolute risk from sub-samples of cohorts (two-phase studies)
  - **Nested case-control design** (Langholz and Borgan, 1997): at each time a case develops sample individuals from risk set
  - **Case-cohort design** (Prentice and Self, 1988): analyze data from subcohort selected at start of follow-up and all cases observed during follow up
    - Raw data (e.g. serum samples) collected on all subjects at baseline but analyzed only for cases and sub-cohort members.
  - For both designs cause-specific hazard for main cause of interest modeled by Cox model, hazard for competing causes estimated non-parametrically (no covariates)

Example: Absolute Risk Model for Breast Cancer with Modifiable Risk Factors "BCmod" Pfeiffer et al. PLoS Medicine, 2013

Combined data on white non-Hispanic women ages 50+ from two cohorts to estimate relative risks for breast cancer

Cohort	**Cohort size**	**Number of BC cases**
NIH-AARP	178,463	5,480
PLCO	62, 249	2,215
Total number	240,712	7,695

To illustrate cohort methods for absolute risk, used only AARP cohort to estimate relative risks (RRs) for invasive breast cancer and competing mortality and baseline hazards

Relative Risk Estimates for Breast Cancer Data Example from Cox Model compared to Fine and Gray sub-distribution from AARP Cohort

Risk factor	**Cause specific RR**	**subdf RR**
Family history of breast/ovarian cancer	1.39	1.39
Benign breast disease/biopsy	1.40	1.41
Age at menopause	1.18	1.14
Age at first live birth	1.17	1.16
Parity	1.32	1.29
Alcohol consumption		
$(0, < 1, 1+$ drinks/day)	1.12	1.13
Body mass index (BMI)		
$(< 25, 25-30, 30-35, 35+)$	1.09	1.10
Estrogen plus progestin HRT use		
(never, 1-9, 10+years)	1.40	1.43
Other HRT use (no/yes)	1.16	1.23

Example: Relative Risk Estimates for Breast Cancer and Competing Mortality from AARP Cohort

Risk factor	**BC RR**	**Comp RR**
Family history of breast/ovarian cancer	1.39	
Benign breast disease/biopsy	1.40	0.88
Age at menopause	1.18	0.86
Age at first live birth	1.17	0.93
Parity	1.32	1.13
Alcohol consumption		
(0,<1, 1+drinks/day)	1.12	1.13
Body mass index (BMI)		
($<25, 25-30, 30-35, 35+$)	1.09	0.80
Estrogen plus progestin HRT use		
(never, 1-9, 10+years)	1.40	0.75
Other HRT use (no/yes)	1.16	0.80
Smoking (never, former, current)		1.29

Absolute BCmod Breast Cancer Risk Estimates: Two 60-year-old Women

Model Predictor	**Woman 1**	**Woman 2**
Age at first birth	25	25
Number of life births	3	3
Menopausal	yes (age 50)	yes (age 55)
Past diagnosis of benign breast disease	no	no
Family history breast/ovarian cancer	no	no
Hormone replacement therapy use (duration)	no	yes (5yrs)
BMI, kg/m^2	24	35
Alcohol consumption (drinks/day)	0	> 1
Relative BC risk	**1.0**	**2.0**
5 year absolute risk estimate	**0.82%**	**2.9%**

Limitations of Cohort Data

- Non-representative absolute risks
- Often assumed relative risks are generalizable but not baseline hazards
- Imprecise and unrepresentative data on competing cause of death
- Lack of detailed covariate data on whole cohort

Comments

- Estimate absolute risk from sub-samples of cohorts (two-phase studies)
  - **Nested case-control design** (Langholz and Borgan, 1997): at each time a case develops sample individuals from risk set
  - **Case-cohort design** (Prentice and Self, 1988): analyze data from subcohort selected at start of follow-up and all cases observed during follow up
    - Raw data (e.g. serum samples) collected on all subjects at baseline but analyzed only for cases and sub-cohort members.
  - For both designs cause-specific hazard for main cause of interest modeled by Cox model, hazard for competing causes estimated non-parametrically (no covariates)
- Cause-specific approach allows combining data from different sources to estimate $h_m(t, \mathbf{X})$

Combine Relative Risk Estimates with Registry Data

- Estimate relative risk $rr(\mathbf{X}, t)$ and attributable risk $AR(t)$ from
  - Cohort data
  - Nested case-control data
  - Case-cohort data
  - Case-control data
- Registry (e.g. NCI's Surveillance, Epidemiology, and End Results (SEER) Program, population-based cancer registries covering 48% of US population): Obtain composite age-specific hazard $h_1^*(t)$ and competing mortality hazard $h_2^*(t)$

Attributable Risk, AR

Model hazard function: $h_1(t, \mathbf{X}) = h_{10}(t) \exp(\boldsymbol{\beta}_1' \mathbf{X})$

$h_{10}^*(t)$ composite sex- age-specific registry incidence rates (no covariates)

Attributabe risk of factors $\mathbf{X}$: $AR(\mathbf{X}) = \frac{h_{10}^*(t) - h_{10}(t)}{h_{10}^*(t)}$

Obtain

$$\hat{h}_{10}(t) = h_{10}^*(t)(1 - \widehat{AR}(\mathbf{X}))$$

Estimate $AR(\mathbf{X})$ from case-control study for type 1 outcomes (Bruzzi et al '85) ($\delta_{1i} = 1$ if person i has event of type 1, 0 otherwise, i.e. is case indicator):

$$\widehat{AR}(t) = 1 - \frac{\sum_{i=1}^n \delta_{1i} \exp\{-\hat{\boldsymbol{\beta}}_1{}' \mathbf{x}_i\}}{\sum_{i=1}^n \delta_{1i}}$$

Combine Data from Different Sources to Estimate $h_1(t, \mathbf{X})$

$$h_1(t, \mathbf{X}) = h_{10}(t) \exp(\beta_1' \mathbf{X})$$

Combine Data from Different Sources to Estimate $h_1(t, \mathbf{X})$

$h_1(t, \mathbf{X}) = h_{10}(t) \exp(\beta_1' \mathbf{X})$

Cohort, nested case-control, case cohort, case-control data

Estimate relative risk, $\exp(\beta' \mathbf{X})$ and attributable risk, $AR(\mathbf{X})$

Disease Registries: data on incidence by age, sex, race, no risk factors

$h_{10}^*(t)$, age, sex, race specific **composite hazard**

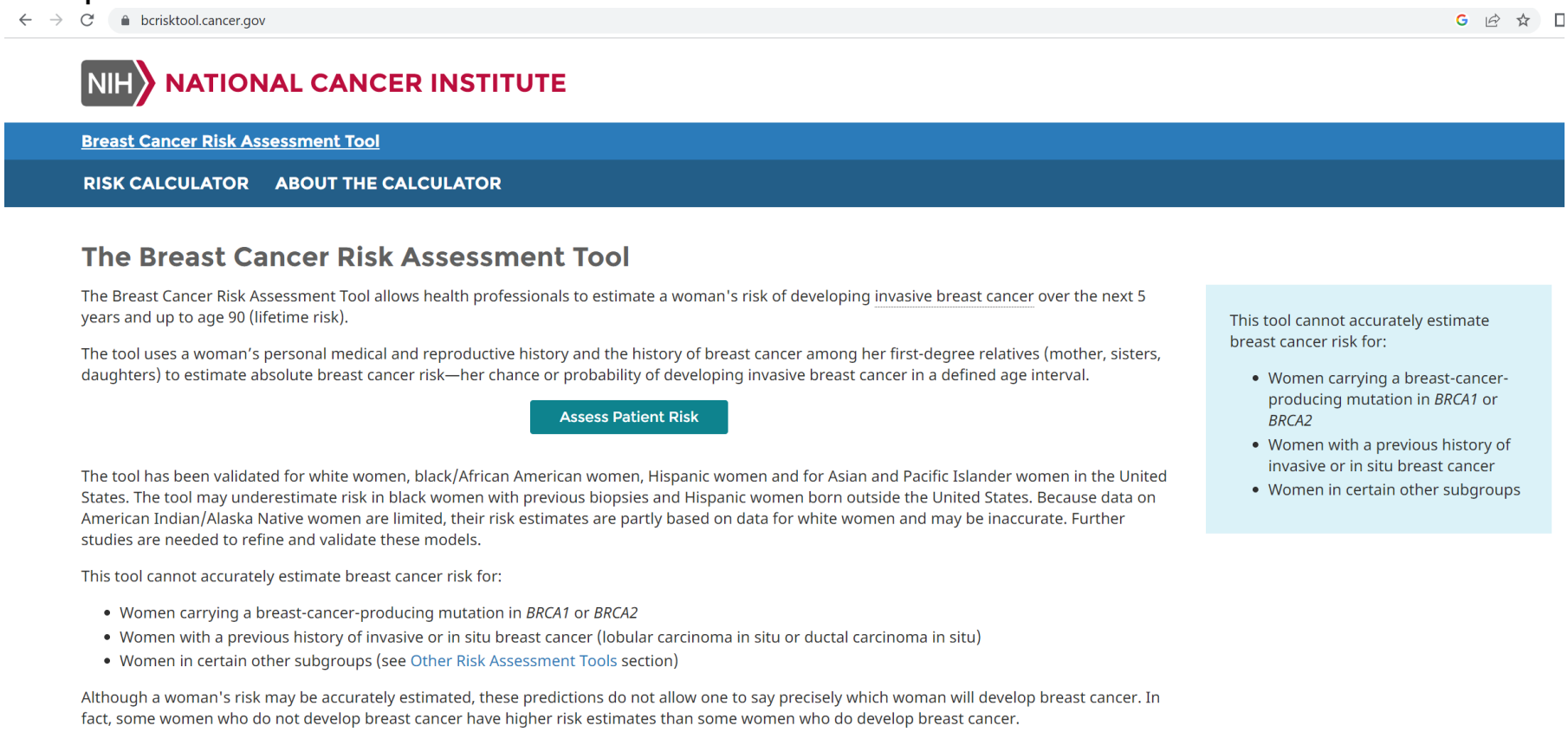
Use that $AR(\mathbf{X}) = \frac{h_{10}^*(t) - h_{10}(t)}{h_{10}^*(t)}$ to get $\hat{h}_1(t, \mathbf{X}) = h_{10}^*(t)(1 - \widehat{AR}) \exp(\hat{\beta}_1' \mathbf{X})$

$$\hat{R}(a, a+\tau, \mathbf{X}) = \int_a^{a+\tau} \hat{h}_1(t, \mathbf{X}) \exp\left[- \int_a^t \{\hat{h}_1(s, \mathbf{X}) + \hat{h}_2(s)\} ds \right] dt$$

Absolute Risk Model for Breast Cancer in General US Population Developed at NCI (Gail MH, et al., 1989, JNCI)

https://bcrisktool.cancer.gov/ (BCRAT or "Gail model")

Predictors: History of biopsy/benign breast disease, family history of breast cancer, reproductive characteristics



Absolute Risk Model for Breast Cancer in General US Population Developed at NCI (Gail MH, et al., 1989, JNCI)

Risk factors:

- Age at menarche
- Age at first life birth
- Number of first degree relatives with breast cancer
- Number of breast biopsies
- Diagnosis of atypical hyperplasia

http://www.cancer.gov/bcrisktool/

Example: Absolute Breast Cancer Risk Model with Modifiable Risk Factors (BCmod) Pfeiffer RM, et al., 2013, PLos Medicine

Example: Absolute Breast Cancer Risk Model with Modifiable Risk Factors (BCmod) Pfeiffer RM, et al., 2013, PLos Medicine

240 712 white women ages 50+ from two cohorts, 7 695 breast cancers; predictors: reproductive & clinical factors, **BMI, alcohol consumption, hormone replacement therapy use**

Estimated $\exp(\beta'\mathbf{X})$ and $AR(\mathbf{X})$

US SEER Cancer Registries: data on race- and age-specific incidence and competing mortality

$$\hat{h}_1(t,\mathbf{X}) = h^*_{10}(t)(1-\widehat{AR})\exp(\hat{\beta}_1'\mathbf{X})$$

$$\hat{R}(a,a+\tau,\mathbf{X}) = \int_a^{a+\tau} \hat{h}_1(t,\mathbf{X})\exp\left[-\int_a^t\{\hat{h}_1(s,\mathbf{X})+\hat{h}_2(s)\}ds\right]dt$$

Relative Risk Estimates for BCmod

Risk factor	**RR**
Family history of breast/ovarian cancer	1.39
Benign breast disease/biopsy	1.40
Age at menopause	1.18
Age at first live birth	1.17
Nulliparous	1.32
Alcohol consumption (0, < 1, 1+drinks/day)	1.12
Body mass index (BMI) (< 25, 25 − 30, 30 − 35, 35+)	1.09
Estrogen plus progestin HRT use (never, 1-9,10+years)	1.40
Other HRT use (no/yes)	

Absolute BCmod Breast Cancer Risk Estimates: Two 50-year-old Women

Model Predictor	**Woman 1**	**Woman 2**
Age at first birth	25	41
Number of life births	3	1
Menopausal	no	yes (48)
Past diagnosis of benign breast disease	no	yes
Family history breast/ovarian cancer	no	yes
Hormone replacement therapy use (duration)	no	yes (5yrs)
BMI, kg/m^2	24	35
Alcohol consumption (drinks/day)	0	> 1
10 year absolute risk estimate	**1.8%**	**9.5%**
20 year absolute risk estimate	**4.1%**	**20.1%**

Advantages of Cause-Specific Modeling

- Familiar interpretation of cause-specific relative risks
- Standard survival methods & software for estimation with cohort data
- Possible to combine different data sources:
  - Relative risks from case-control or case-cohort data
  - Incorporate composite registry incidence rate $h_1^*(t)$ into baseline hazard

$$h_{10}(t) = h_1^*(t)\{1 - \text{attributable risk}(t)\}$$

  - Software
    `https://bioconductor.org/packages/release/bioc/html/iCARE.html`
- Can accommodate disease heterogeneity, e.g. hormone receptor status in breast cancer

Some Cancer Risk Models

- Cancer incidence
  - `https://bcrisktool.cancer.gov/`(breast BCRAT)
  - `https://ccge.medschl.cam.ac.uk/boadicea/` (breast BOADICEA)
  - `https://ibis.ikonopedia.com/` (breast IBIS)
  - `https://ccrisktool.cancer.gov/` (colorectal cancer)
  - `https://mrisktool.cancer.gov/` (melanoma)
  - Other models: Bladder, Endometrial cancer, Lung, Ovary, Pancreas, Prostate
- Absolute risk of death from prostate cancer after diagnosis (Albertson, Hanley, Fine, JAMA 2005)

Assessment of Risk Model Performance, Model Validation

Risk Model Validation

Before model can be recommended for practical use, need to understand its performance characteristics in independent data for rigorous assessment

- **Calibration (bias)**: Does model correctly predict number of observed events; requires cohort data
- **Classification accuracy:** How well does model categorize/classify individuals
- **Discrimination:** How different are risks in cases compared to non-cases?

Types of Validation

- **External validation:** evaluation of model performance in sample independent of that used to develop the model
- **Internal validation:** reusing same dataset on which model was developed to assess overfit and correct for resulting optimism in model performance

Independent Population for Validation (External Validation) of Absolute Risk Model

Assume population of N individuals followed prospectively over time period τ

Observe

$$Y_i = \begin{cases} 1 & \text{if ith subject develops cancer by time } \tau \\ 0 & \text{otherwise} \end{cases}$$

Notation and General Framework

$P(Y_i = 1) = \pi_i$ true (unknown) absolute risk of ith person
$R(\mathbf{X}_i)$ absolute risk estimate for ith person with baseline covariates $\mathbf{X}_i$, including age a_i
Risk model is fixed and known from previous data
In validation cohort distribution H of $\mathbf{X}$ induces distribution F of risk R

$$F(r) = P(R \leq r) = \int_{\{x:R(\mathbf{x})\leq r\}} dH_{\mathbf{X}}(\mathbf{x}),$$

If validation cohort is simple random sample estimate F from observed risks $r_1, \ldots, r_n$ by

$$\hat{F}(r) = F_n(r) = \frac{1}{n}\sum_{i=1}^{n} I(r_i \leq r)$$

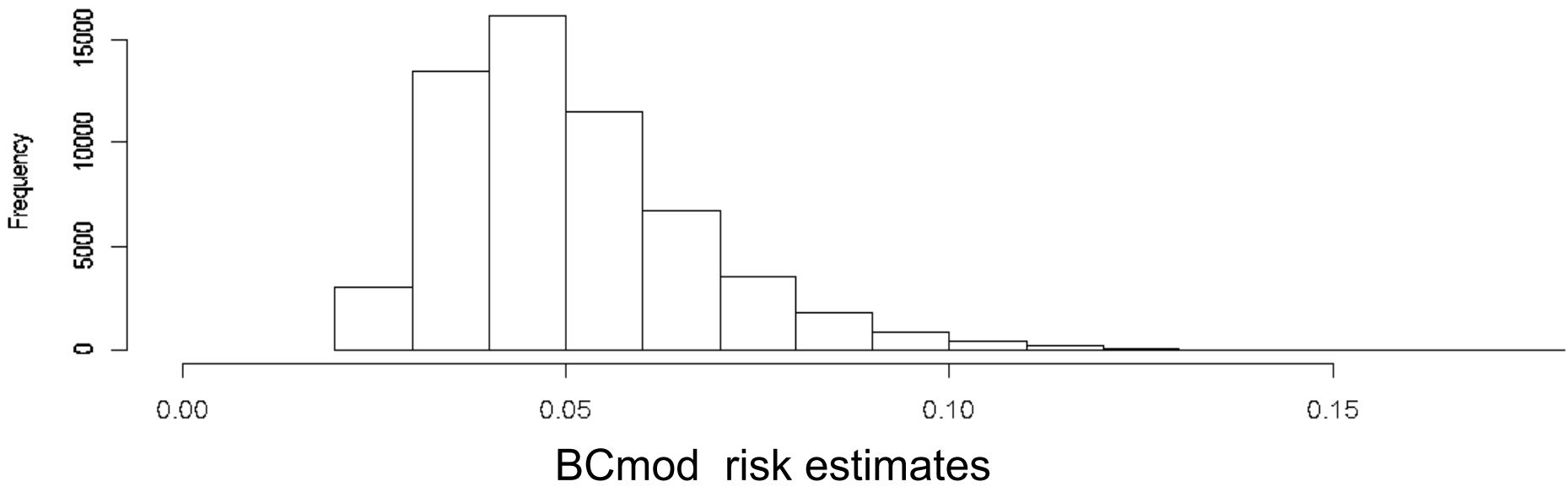
Example: Assess Model Performance of BCmod Breast Cancer Risk Model in Independent Population

Nurses Health Cohort: $N = 57,907$ women ages 51-70 at baseline
`https://www.nurseshealthstudy.org/`

Compute BCmod breast cancer risk estimate $R(a, \tau, \mathbf{X})$ for each woman, given her baseline risk factors $\mathbf{X}$, age a, and projection period τ

Observe breast cancers that develop during follow-up (median 16 years)

Distribution F of BCmod Risks in Nurses Health Validation Cohort



Assessing Calibration Using Independent Validation Cohort

Y - event indicator at end of cohort follow-up
$R(\mathbf{X}) = \hat{P}(Y = 1|\mathbf{X}, a, \tau)$ - absolute τ year cancer risk (e.g. $\tau = 5$)
Model R **well calibrated (unbiased)** in population if, for every value r

$$P(Y = 1|R(\mathbf{X}) = r) = r$$

Then $E(Y) = P(Y = 1) = E(R)$
If R well calibrated (unbiased) in cohort with $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$

$$\hat{E}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i \approx \frac{1}{n}\sum_{i=1}^{n} R(\mathbf{X}_i) = \hat{E}(R)$$

Assessing Calibration Using Independent Validation Cohort

Goodness-of fit criteria based on comparing observed ($\mathcal{O}$) and expected ($\mathcal{E}$) numbers of events overall or in subgroups of population

Here we use

$$\frac{\mathcal{O}}{\mathcal{E}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} R(\mathbf{X}_i)}$$

- If $R << 1$, then $\mathcal{O} \sim$ Poisson
- If R well calibrated $\mathcal{O}$ has mean $\mathcal{E}$ and

$$\frac{\mathcal{O}}{\mathcal{E}} \approx 1$$

- **BCmod calibration in N=57,907 women in NHS cohort:** $\mathcal{O}/\mathcal{E} = 2934/2930 = 1.00$, (95%CI: $0.96, 1.04$)

Assessing Calibration Using Independent Validation Cohort

Compute deciles $\xi_{0.1g}$ of empirical distribution of R_i and let $S_g = \{\mathbf{X}_i : \xi_{0.1(g-1)} \leq R_i < \xi_{0.1g}\}$, $g = 1, \ldots, 10$, $(\xi_0 = 0)$

$$\text{Hosmer-Lemeshow statistic} : Q = \sum_{g=1}^{G} \frac{(O^g - E^g)^2}{E^g(1 - E^g/N^g)} \tag{1}$$

where N^g is number of subjects in group g.

Assessing Calibration Using Independent Validation Cohort

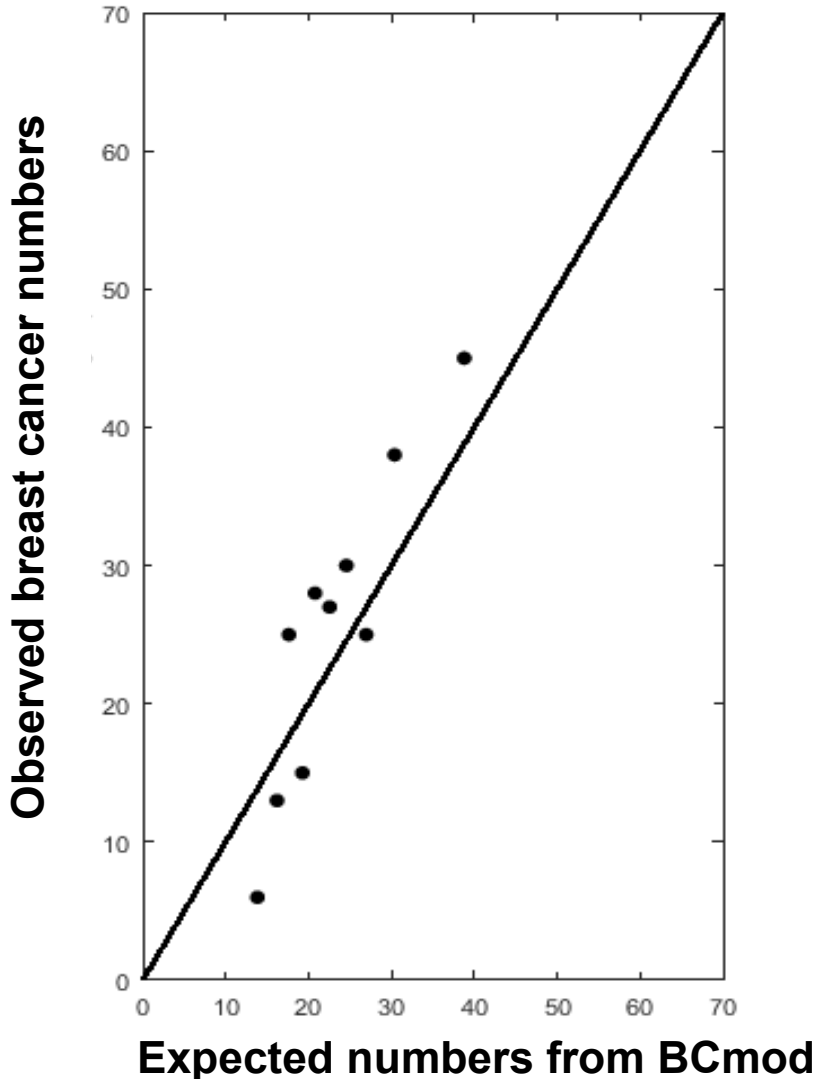
Compute deciles $\xi_{0.1g}$ of empirical distribution of R_i and let $S_g = \{\mathbf{X}_i : \xi_{0.1(g-1)} \leq R_i < \xi_{0.1g}\}$, $g = 1, \ldots, 10$, $(\xi_0 = 0)$

$$\text{Hosmer-Lemeshow statistic} : Q = \sum_{g=1}^{G} \frac{(O^g - E^g)^2}{E^g(1 - E^g/N^g)} \tag{1}$$

where N^g is number of subjects in group g.

BCmod calibration in N=57,907 women in NHS cohort:
$Q = 16.7 < \chi^2_{10} = 18.31$

Calibration Plot for BCmod for Incident Breast Cancers in NHS Cohort



Why a Risk Model May be Poorly Calibrated in Validation Population

- Differences in
  - Operational definition of predictor (e.g. number biopsies)
  - Procedures used to diagnose the disease outcome
  - Screening/outcome verification
- Incomplete follow-up in validation cohort
- Missing predictor information

Assessing model calibration when predictors missing on cohort members (Shin, Gail, Pfeiffer, Biostatistics, 2020)

Assume event indicator Y always observed, $\mathcal{O}$ known
Predictors for $R(\mathbf{X}, \mathbf{Z})$ fall into two categories:

- $\mathbf{X}$ available on everybody, "*phase 1* predictors"
- $\mathbf{Z}$ only observed for subset sampled into *phase 2*

MCAR: data are *missing completely at random*
MAR: *missing at random*; Phase 2 sampling depends on observed data
Important (MAR) sub-sampling designs from cohort:

- **Nested case-control (NCC) design**: each time case develops sample controls from those at risk
- **Case-cohort (CC) design**: sub-cohort selected at start of follow-up and all cases sampled during follow-up

Estimate $\mathcal{E}$ using Multiple Imputation

Standard approach to dealing with missing data

- Create K complete copies of data by imputing missing values from model for their predictive distribution conditional on observed data, using e.g. *mice* (multivariate imputation by chained equations) in *R* (Buuren et al, '10)
- Compute

$$\hat{\mathcal{E}}^{\text{mi}} = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} R(\mathbf{x}_i, \tilde{\mathbf{z}}_i^{(k)}) = \sum_{i=1}^{n} \bar{R}(\mathbf{x}_i, \tilde{\mathbf{z}}_i)$$

  where $\bar{R}(\mathbf{x}_i, \tilde{\mathbf{z}}_i) = K^{-1} \sum_{k=1}^{K} R(\mathbf{x}_i, \tilde{\mathbf{z}}_i^{(k)})$ is mean of risks evaluated at imputed value $\tilde{\mathbf{z}}_i^{(k)}$ for kth imputation

Use **all observed data** including follow-up time in imputation model (White et al, 09)

Estimate $\hat{\mathcal{E}} = \hat{E}(R) = \sum_i R(\mathbf{X}_i, \mathbf{Z}_i)$ using weighting approaches

S_i- sampling indicator ($S_i = 1$ if person i in phase-2, 0 if not)
$\rho_i = P(S_i = 1)$ phase 2 inclusion probability known for MCAR, NCC, CC sampling
Unbiased estimate of expected number of cases in cohort is

$$\hat{\mathcal{E}}^{\circ} = \sum_{i=1}^{n} S_i w_i^{\circ} R(\mathbf{x}_i, \mathbf{z}_i)$$

where $w_i^{\circ} = 1/\rho_i$ is inverse probability of inclusion weight
Problem: (Horvitz-Thompson) estimate inefficient
Solution: improve efficiency by adjusting w° using phase 1 variables

Adjustment of sampling weights (! Math)

$\mathbf{w}^o = (w_1^o, \ldots, w_n^o)$ - inclusion probability weights
$\mathcal{X} = f(\mathbf{X}, \mathbf{U}, Y)$ - function of phase 1 data observed on all cohort members

$$\mathbf{w}^{\text{adj}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} S_i \mathrm{d}(w_i, w_i^{\text{o}}) \text{ s.t. } \sum_{i=1}^{n} \mathcal{X}_i = \sum_{i=1}^{n} S_i w_i \mathcal{X}_i$$

distance measure $\mathrm{d}(a, b) = a(\log a - \log b) + b - a$
Raking calibration (Deville & Särndal '92):

$$w_i^{\text{adj}} = w_i^{\text{o}} e^{\eta^{\mathsf{T}} \mathcal{X}_i}$$

Implemented in function *calib* in *R* package *sampling*

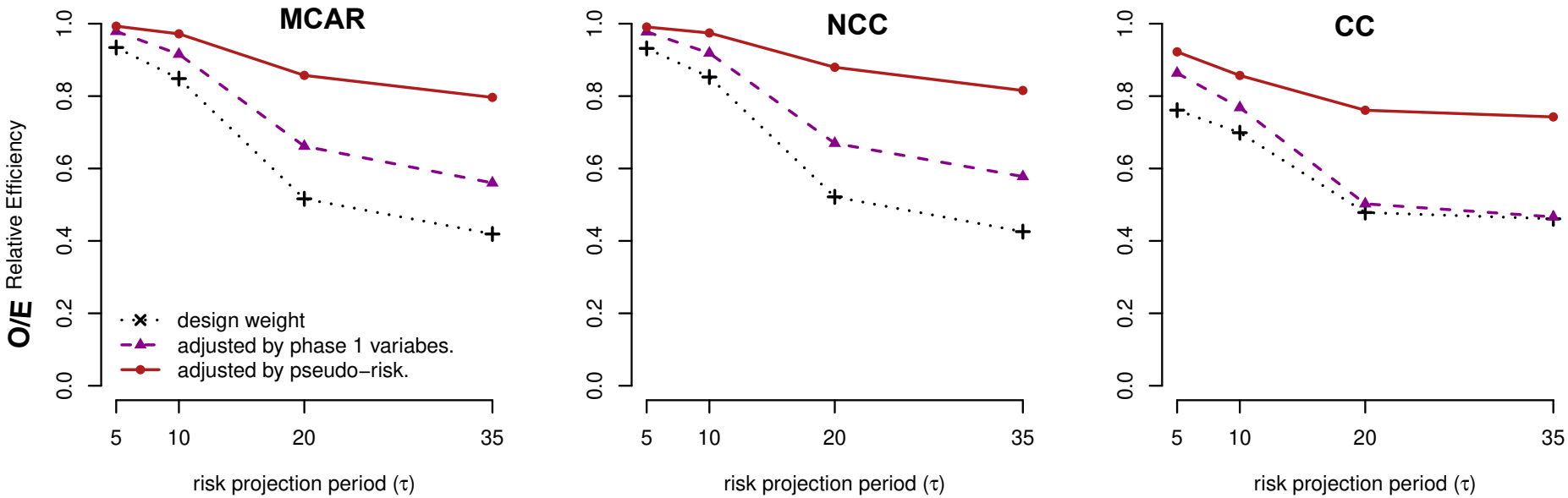
Auxiliary variables for weight adjustment (! Math)

Key: find good auxiliary variables for $\mathcal{E}$ using complete data

Our Proposal:

- Predict $\mathbf{Z}$ from phase 1 variables $(\mathbf{X}, \mathbf{U}, Y)$ using weighted ($\mathbf{w}^o$) GLMs
- Use "pseudo-risk" estimate $R(\mathbf{x}, \hat{\mathbf{z}})$ as auxiliary variable in weight adjustment

Empirical relative efficiency (ratio of empirical variance from full cohort to that from two-phase estimate) of $\mathcal{O}/\mathcal{E}$ using simulated data



Weight adjustment using pseudo-risk more efficient for all projection periods τ than using design weights or adjusting weights by phase 1 variables directly

Comment: Adjusted-weight procedure yields unbiased estimate $\widehat{\mathcal{E}}$ even if model for $\mathbf{Z}$ mis-specified

Assessing Model Accuracy

Assessing Model Accuracy

For clinical decision making a rule is needed to classify subject i as diseased based on model estimate

$$\hat{Y} = \begin{cases} 0 & \text{if } R \leq r^* \\ 1 & \text{if } R > r^* \end{cases}$$

where r^* is risk/decision threshold

Thus, rather than giving risk estimate R, one is forced to guess outcome

Classification accuracy measures how well rule identifies cases/non-cases

Accuracy Scores

Measure how well true outcome predicted

- Positive predictive value, $PPV = P(Y = 1|\hat{Y} = 1)$

Accuracy Scores

Measure how well true outcome predicted

- Positive predictive value, $PPV = P(Y = 1|\hat{Y} = 1)$
- Negative predictive value, $NPV = P(Y = 0|\hat{Y} = 0)$

Accuracy Scores

Measure how well true outcome predicted

- Positive predictive value, $PPV = P(Y = 1|\hat{Y} = 1)$
- Negative predictive value, $NPV = P(Y = 0|\hat{Y} = 0)$
- Weighted combinations of both, e.g.

$$PCC = P(\text{correct classification}) =$$
$$P(R > r^*)P(Y = 1|R > r^*) + P(R < r*)P(Y = 0|R < r^*)$$

Accuracy Scores

Measure how well true outcome predicted

- Positive predictive value, $PPV = P(Y = 1|\hat{Y} = 1)$
- Negative predictive value, $NPV = P(Y = 0|\hat{Y} = 0)$
- Weighted combinations of both, e.g.

$$PCC = P(\text{correct classification}) =$$
$$P(R > r^*)P(Y = 1|R > r^*) + P(R < r*)P(Y = 0|R < r^*)$$

Depend on sensitivity $P(\hat{Y} = 1|Y = 1) = P(R > r^*|Y = 1)$, specificity $P(\hat{Y} = 0|Y = 0) = P(R \leq r^*|Y = 0)$, **disease prevalence/incidence**, $P(Y = 1)$

Accuracy measures for BCmod in NHS validation cohort

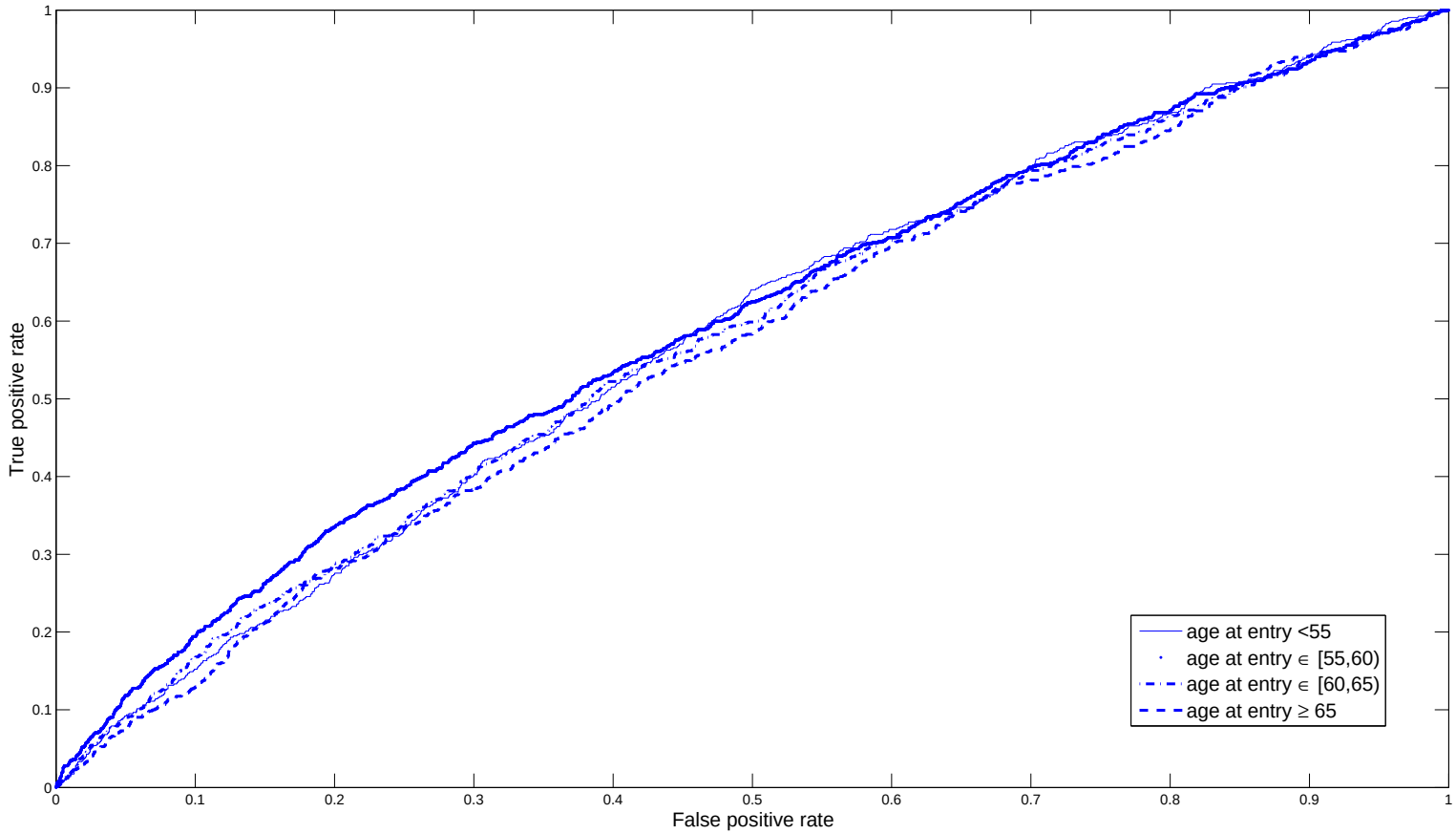
5-year absolute risk threshold $r^* = 0.0166$ for 50 to 55 year old women (252 cases, 16833 non-cases). E.g. 84 cases had $R > 0.0166$

Accuracy measure	**Estimate**	**95% CI**
Sensitivity	$84/252 = 0.33$	(0.28, 0.40)
Specificity	$13456/16833 = 0.80$	(0.79, 0.81)
PPV	$84/3461 = 0.024$	(0.019, 0.030)
NPV	$13456/13624 = 0.988$	(0.986, 0.990)
PCC	$(13456 + 84)/ 17085 = 0.793$	(0.786, 0.799)

PPV=Positive predictive value, NPV=Negative predictive value
PCC=Prob. of correct classification

Measures of Discrimination for Range of Thresholds

ROC curve: plots sensitivity against 1-specificity



Measures of Discrimination for Range of Thresholds

- **ROC curve:** plots sensitivity against 1-specificity
- **Area under the ROC curve (AUC)**:

$$AUC = P(R_i > R_k | Y_i = 1, Y_k = 0)$$

  - Concordance statistic (Rockhill et al, 2001; Bach et al, 2003)
  - Mann-Whitney-Wilcoxon Rank Sum Test

Measures of Discrimination for Range of Thresholds

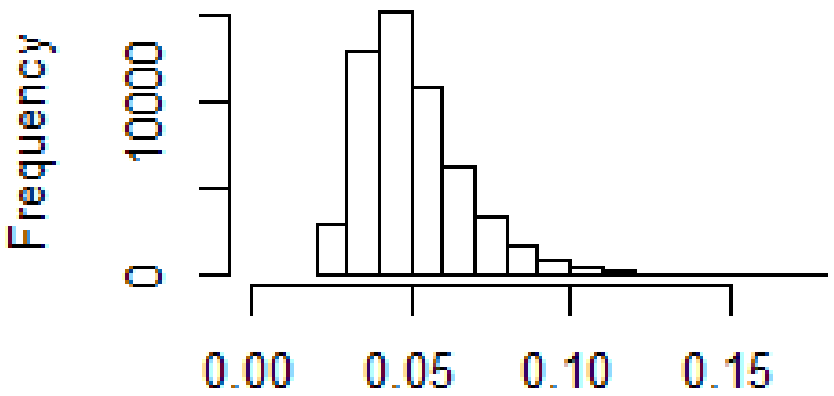
- **ROC curve:** plots sensitivity against 1-specificity
- **Area under the ROC curve (AUC)**:

$$AUC = P(R_i > R_k | Y_i = 1, Y_k = 0)$$

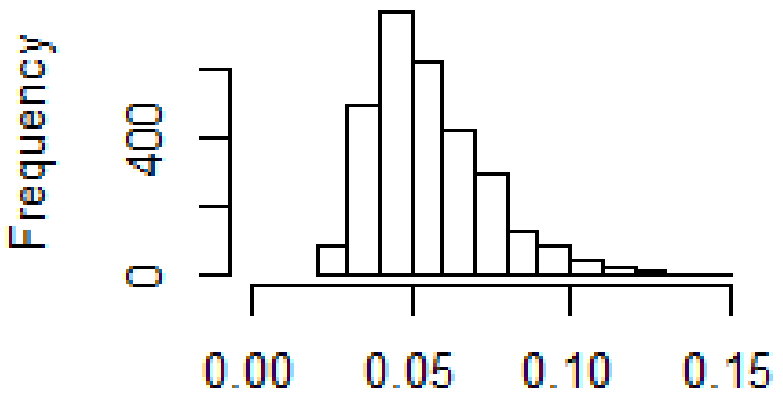
  - Concordance statistic (Rockhill et al, 2001; Bach et al, 2003)
  - Mann-Whitney-Wilcoxon Rank Sum Test

- Partial area under the ROC curve (Pepe, 2003; Dodd & Pepe, 2003)

Distribution of BCmod Risk Estimates in NHS Cohort by Breast Cancer Status at End of Follow-up



BCmod risk in women without breast cancer diagnosis



BCmod risk in women with breast cancer diagnosis

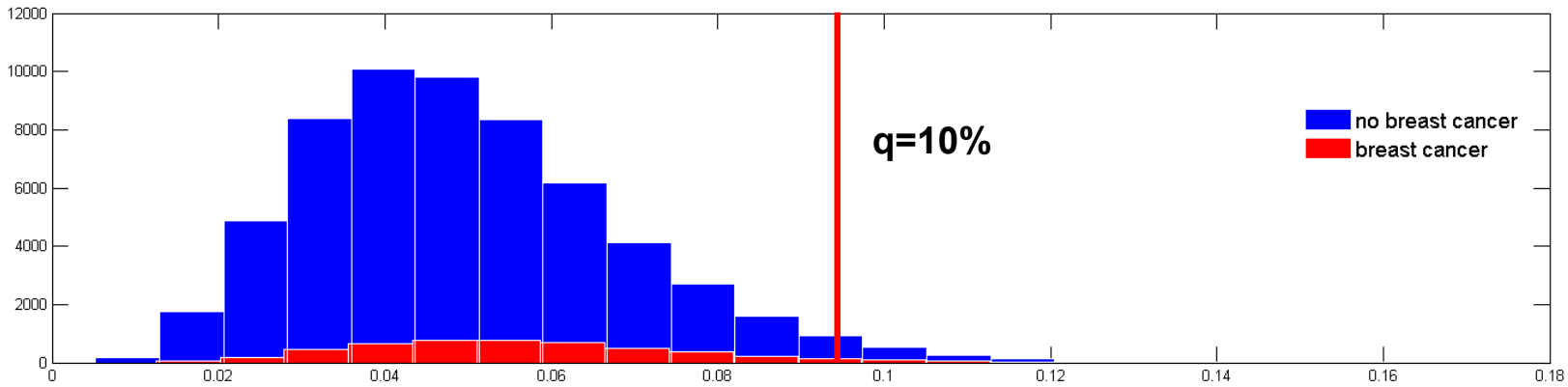
AUC=0.58

Criteria to Assess Model Performance for Screening Applications

Pfeiffer and Gail, 2011; Pfeiffer, 2013

- Proportion of cases screened: % of cases found in top 10% of population based on risk estimated from model
- Proportion needed to screen: % of population at highest risk from model needed to be screened to capture e.g. 90% of cases

Distribution of Risks in NHS Validation Cohort



Proportion of Cases Screened ("Followed"), $PCF(q)$

Proportion of cases included in proportion q of population at highest risk

Population distribution of risk: $F(r) = P(R \leq r)$

Distribution of risk in cases $(Y = 1)$: $G(r) = P(R \leq r | Y = 1)$

$$PCF(q) = P(R_{case} > (1-q)\text{th quantile of } F) = 1 - G\{F^{-1}(1-q)\}$$

Integrated PCF

$$iPCF = \int_0^1 PCF(q)dq = P(R_G > R_F)$$

Proportion of Cases Screened ("Followed"), $PCF(q)$

Proportion of cases included in proportion q of population at highest risk

Population distribution of risk: $F(r) = P(R \leq r)$

Distribution of risk in cases $(Y = 1)$: $G(r) = P(R \leq r | Y = 1)$

$$PCF(q) = P(R_{case} > (1-q)\text{th quantile of } F) = 1 - G\{F^{-1}(1-q)\}$$

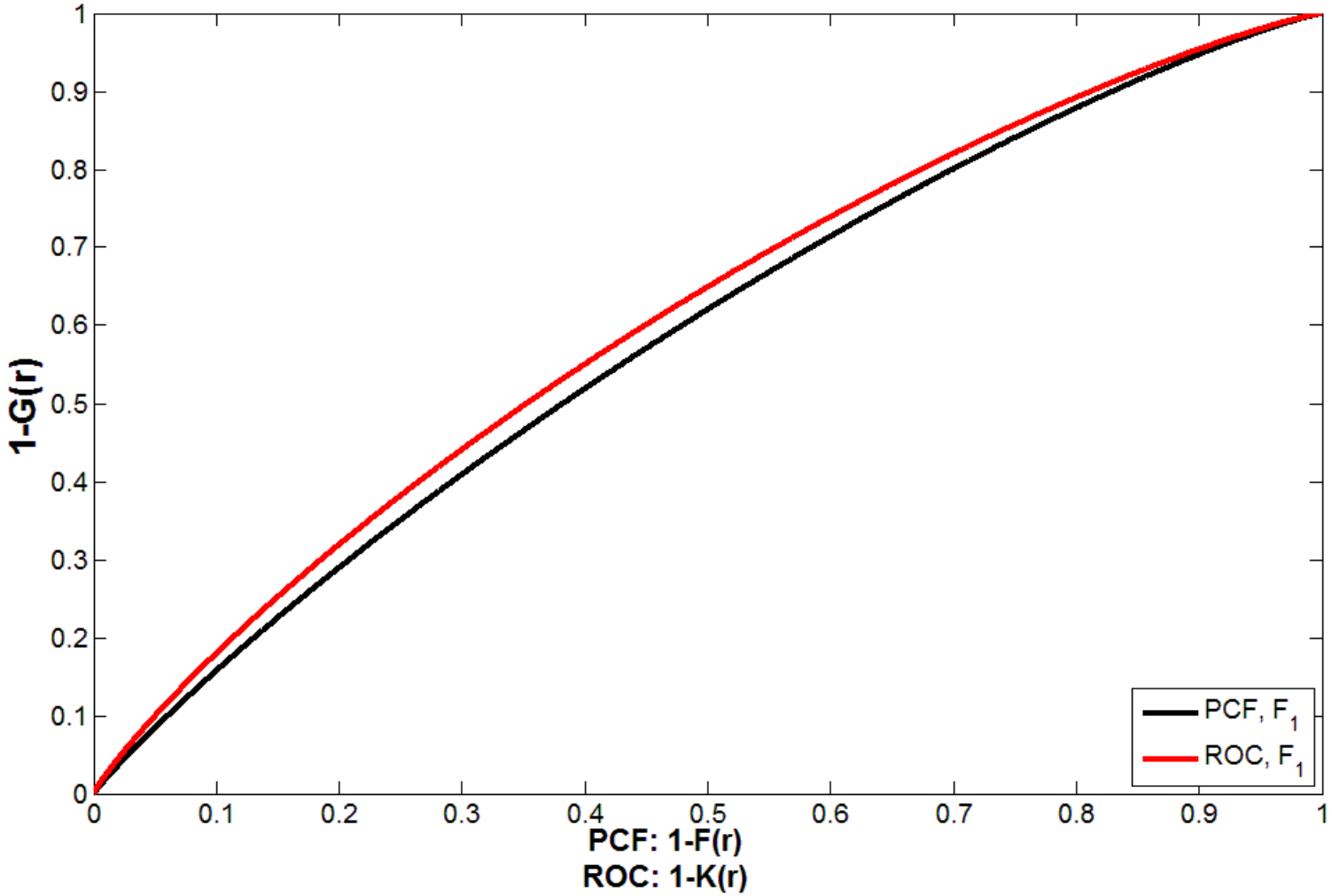
Integrated PCF

$$iPCF = \int_0^1 PCF(q)dq = P(R_G > R_F)$$

Distribution of risk in non-cases $(Y = 0)$: $K(r) = P(R \leq r | Y = 0)$

$$AUC = P(R_G > R_K)$$

ROC and PCF curves, AUC=0.58, iPCF=0.56



Accommodating population differences when validating risk prediction models (Pfeiffer, Chen, Gail, Ankerst, Stat in Med, 2023)

This topic will be discussed on day 2 of the course