Philipps Universität Marburg

Issues for analyzing competing-risks data with missing or misclassification in causes

Dr. Ronny Westerman

Institute of Medical Sociology and Social Medicine
Medical School & University Hospital

July 27, 2012

- Outline

- Introduction/Background
- Data
- Methods
- Discussion
- Perspectives

- Introduction
- Limited Failure Models (Immortal)
- Competing Risks
- Missing and Misclassification of causes (Masked causes)

- Limited Failure Model (Cure Survival Models)

- Examples: Infant Mortality
- Curability of cancer and decreasing mortality risk since diagnosis of cancer

- None defective units are not expected to fail from risk

The Competing Risk Problem:

Each subject being exposed to many competing risks, but only one will be caused the failure

Subject ist still right-censored if it do not fail within the follow-up duration

- Competing Risks

- Non-parametic, semi-parametric and full-parametric models
- Cause-spezific hazard function

- Problem: Assumption of independence through cause often violated?

- Failure Time for all risks are operatively the same, in that case, all risks being removed except the risk under consideration

- Missclassification and Missing of causes

- Cause of event for some of units or individuals not exactly identified or recored

- Partial masking: Cause is narrowed down but not exactely identified

- Reason for missclassification:
- documentation containing the information needed for attributing the cause of failure may be not collected, or the cause of diseases for some patients may be difficult to determine

- Difficulties for determination: (aetiological problems)

- Example: Cardioembolic stroke (Leary and Caplan, 2008)

- Cardioembolic stroke occurs when the heart pumps unwanted materials into the brain circulation, resulting in the occlusion of a brain blood vessel and damage to the brain tissue.

- CS diagnosed in 3-8% stroke patients, but in various current stroke registries, approximately 10-20% patients with CS have not maximal symptoms at the onset of their stroke → Exclusion

- Missclassification:

- Example: Breast Cancer

- TNM Staging vs . I-IV Staging

- Stage migration: improved detection of illness leads to movement of people from the set of healthy people to the set of unhealthy people

- Will Rogers phenomen:
  „When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states“

- Methods for treating masked cause data

- 1) Mutiple Imputations

  Should be used, when Baseline are not proportional

  Works good in case of Missing at Random (for cause)

- Problem: High-Mortality-Risks, Multiple-Specific and High-Potential-Risks often not Missing at Random

  → False classification or misinterpretation of cause-specific mortality

- Methods for treating masked cause data

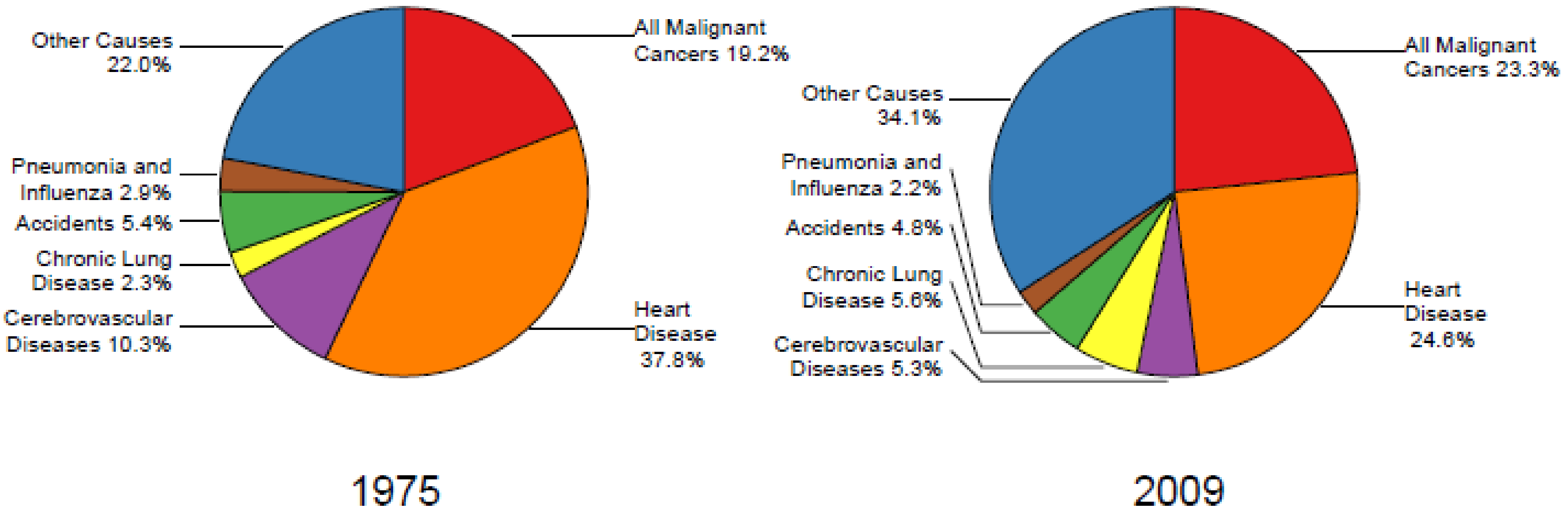
- 2) Second Stage Analysis
- Models with non-proportional cause-specific hazard

- 3) EM for grouped Survival data
  Bayesian Methods

- Assumptions for masked causes:
- Right censoring, if causes not exactly identified
- Masking probabilty is constant over time

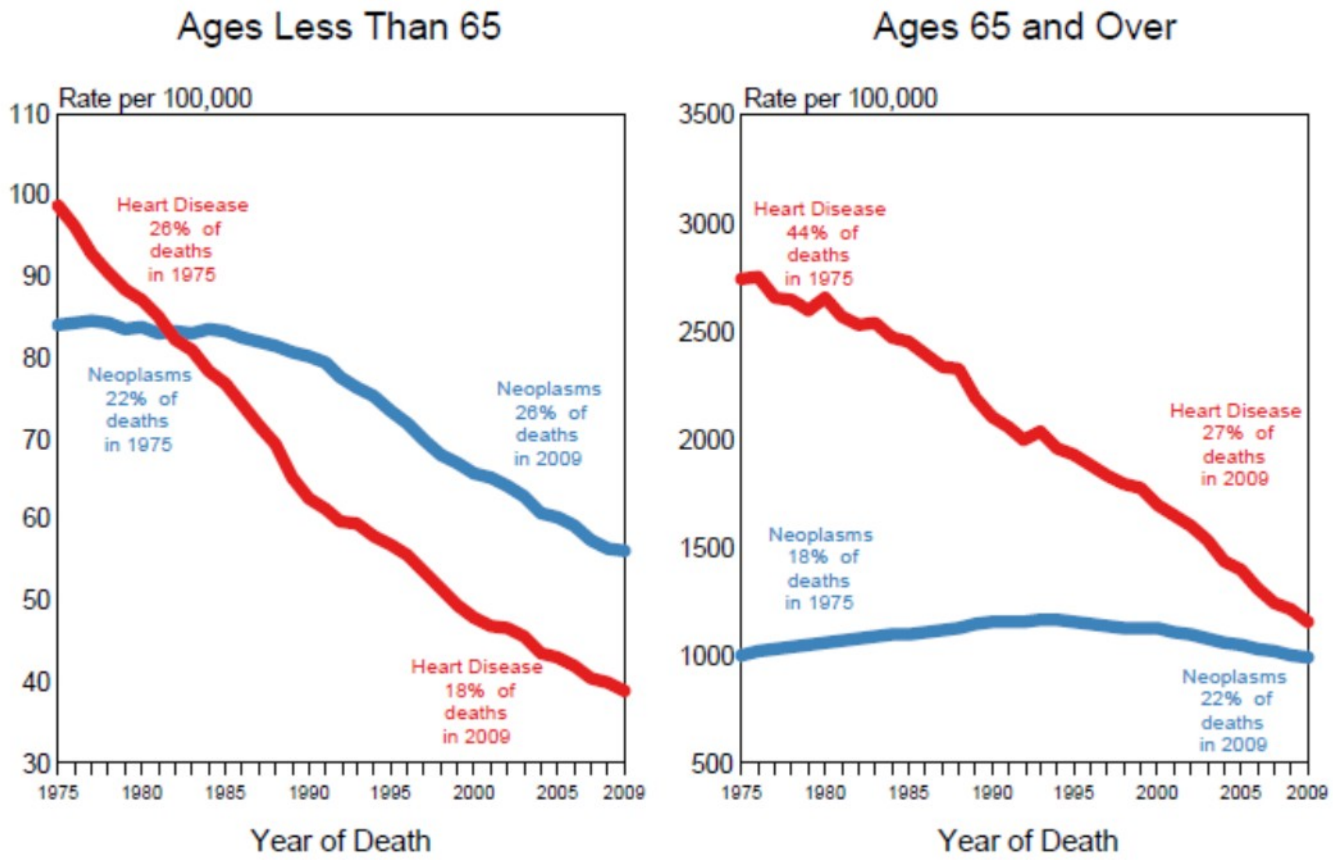
- SEER Cancer Statistic Data Base National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch (released April 2012)
- Incidence by Race, Gender and Age (different periods of time)
- Cause-Specific Mortality including all specific cancer
- SEER public use dataset on survival of breast cancer patients from 1992-2009 (n=69,990 in Situ)

- Leading Cause of Death in the U.S. 1975 vs. 2009

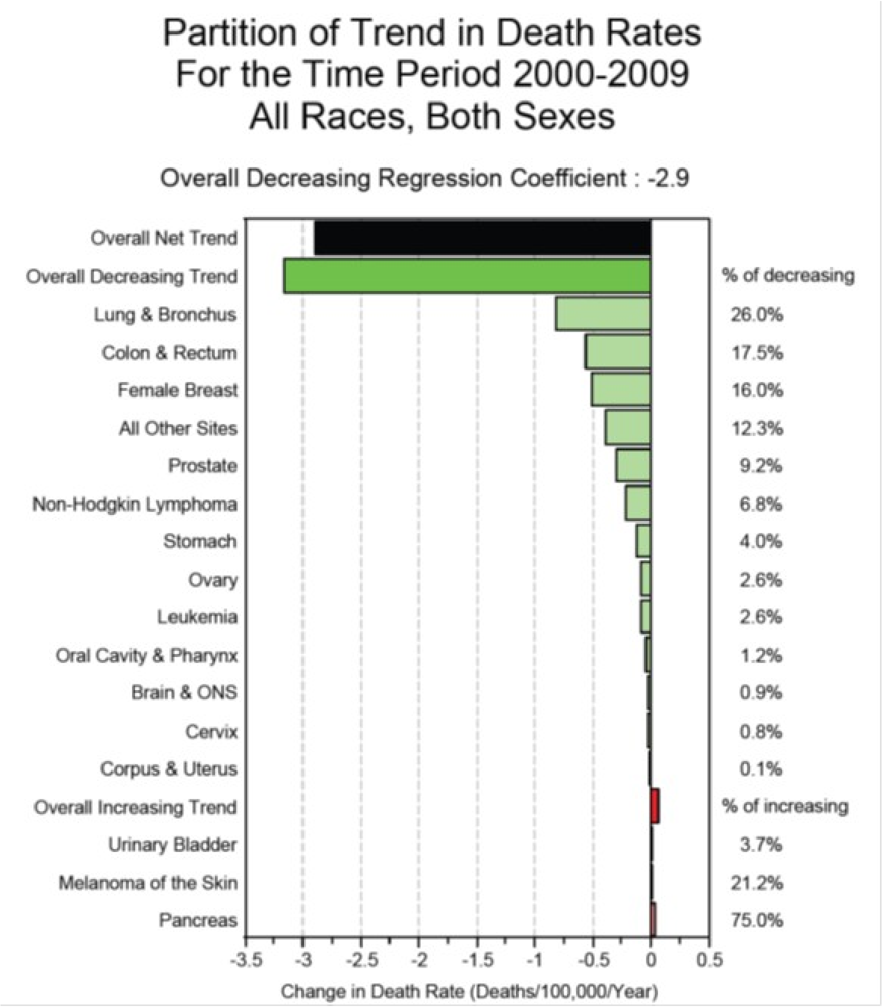


Source: US Moratlity Files, National Center of Health Statistics, Centers of Disease Control and Prevention

- US Death Rates, 1975-2009 Heart Disease compared to Neoplasms, by age at death

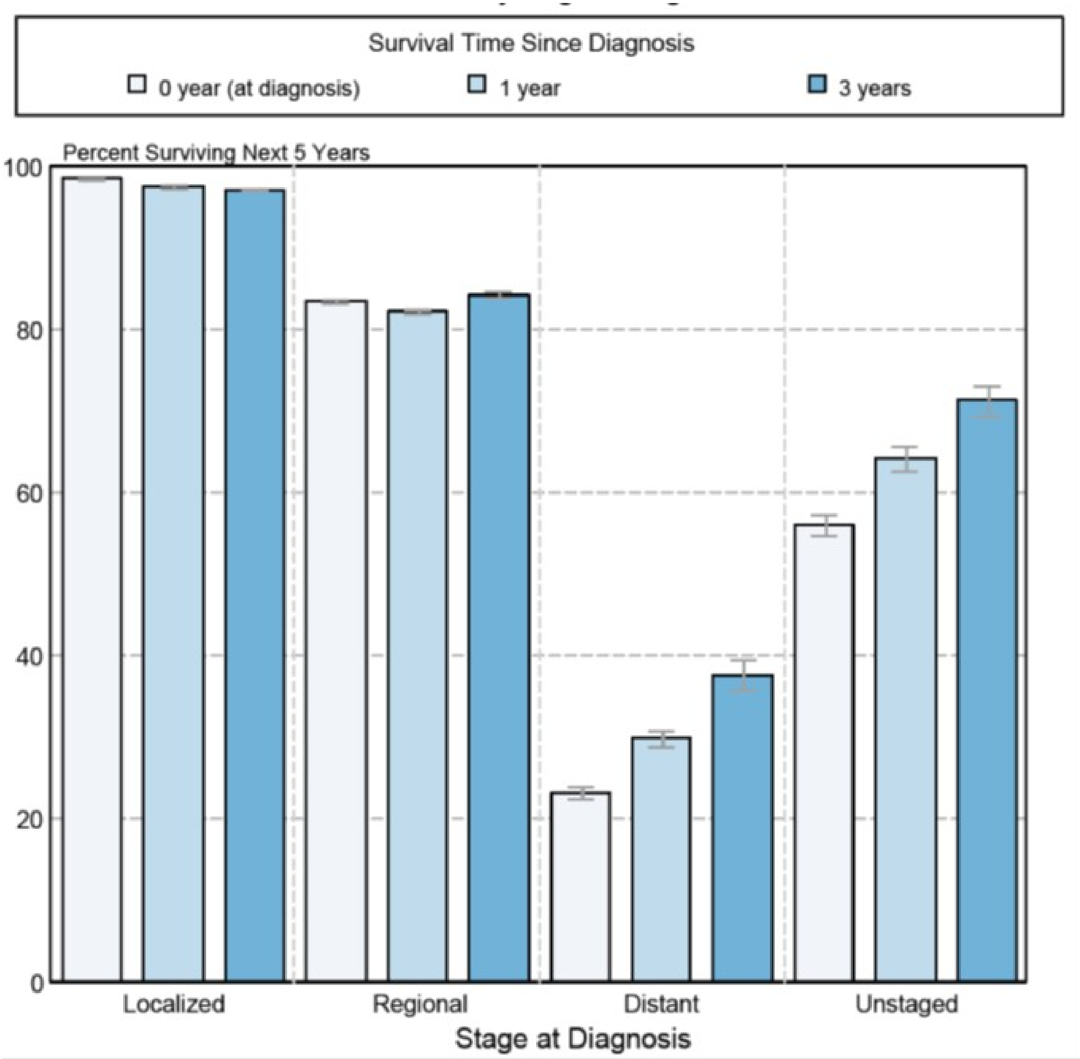


Source: US Moratlity Files, National Center of Health Statistics, Centers of Disease Control and Prevention
Rates are per 100,000 and age-adjusted to the 2000 US Std Population (19 age groups - Census P25-1103).



Source: US Moratlity Files, National Center of Health Statistics, Centers of Disease Control and Prevention

- 5-year Conditional Relative Survival for Cancer of female Breast



SEER, 2012

- And now, what‘s the problem?
  Preliminary Analysis with SEER- DATA (Sen et al. 2010)

- Over-sampling the masked cases
- 46 % of the women died during follow-up
- Specific mortality related to breast cancer, other cancer or non-cancer related causes

- for 56 % the exact cause of death was known
- for 35 % partial information available

- 30 % with missing cause of death: false classification (breast cancer to other or multiples cancer)
- 65 % missing causes were complety masked

- How do deal with masked causes ?
- Motivation to use Two-Component-Model for masked causes
- Risks are latent: no specific information about the cause of the component failure
- Only some individuals may susceptible to the event of interest (curability or the recessive risk for the disease)

- Useful Stata commands for cure models: lncure, spsurv, and cureregr (Lambert, 2007)

- the advances of **cureregr:** fits both mixture and nonmixture cure models

  parametric distributions: exponential, Weibull, lognormal, and gamma parametric distributions available

- Optional: **strsmix** allowing more flexible parametric distributions

Data Analysis with SEER Breast Cancer Data

- Survival of breast cancer patients
  from 1992-2009 (n=69,990 in Situ)

- cause of death: breast cancer and other causes
  other causes as competing risks

- We used a non-mixture cure fraction model with Weibull and Exponential specification

Results from Data Analysis (Estimates for the Long-Term Survival Function)

Table 1: MLEs and the standard errors for SEER Breast Cancer Data

Distribution	λ	φ	p
Weibull	0.0047 (0.00021)	0.6732 (0.1428)	0.28057 (0.1016)
Exponential	0.0041 (0.0009)		0.3032 (0.0962)

Λ-scale parameter, φ- shape parameter, p- long-term parameter

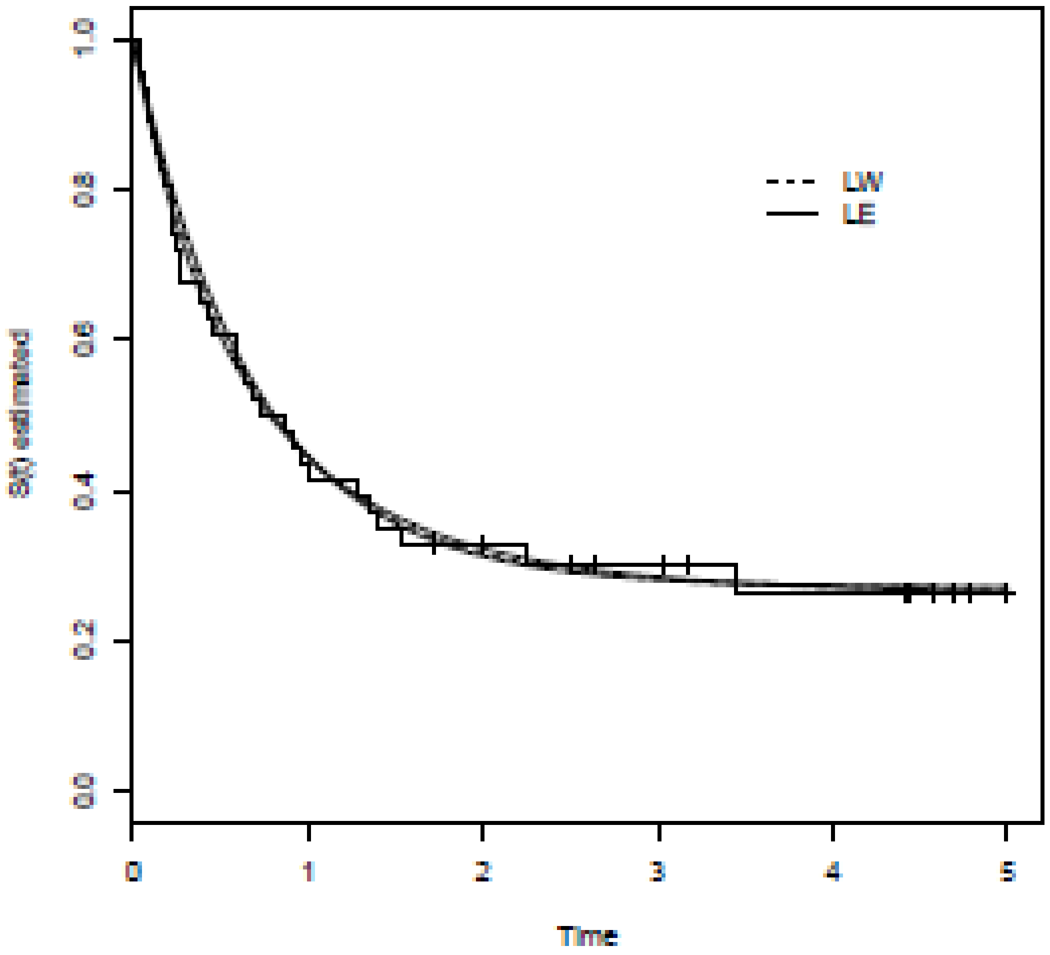
Table 2: Likelihood, AIC and BIC values

Model	$\ell(.)$	AIC	BIC
Weibull	-46.12845	98.27691	103.7626
Exponential	-46.20798	96.43586	100.1032

- Results

- no evidence that Weibull provides a better fitting than the Exponential for Seer Breast Cancer Data at 5% significance

- corrobate the empirical Kaplan-Meier Survival


LW
LE
S(t) estimated
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
Time

Thrills and Tears with Cure Survival Models

Thrills: less assumptions and minor computation problems

Tears: to overcome the naïve assumption for infinite failure time of the nonsusceptiple units

Limitations for parametric hazard functions

The complexity of the baseline hazard function (Crowther and Lambert, 2011)

- beyond standard and sometimes biologically and implausible shapes
- a turning point in the hazard function is observed
- 2-component mixture distribution e.g. Weibull-Weibull-distribution

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

other distribution families also available

- Options in STATA

- STPM2: Stata module to estimate flexible parametric survival models (Royston-Parmar models) (updated by Lampert, 2012)

- STPM2 also used with single- or multiple- record (more generalized)

- STMIX: 2-component parametric mixture survival models (Crowther and Lambert, 2011)

  distribution choices includes Weibull-Weibull or Weibull-exponential

- STMIX can be used with single- or multiple-record

- References

- Craiu and Lee (2005): Model Selection for the Competing-Risks Model with and without masking. Technometrics, Vol.25, No.4, 457-467
- Leary MC, Caplan LR (2008): Cardioembolic stroke: An updated on etiology, diagnosis and management. Annuals Indian Academic Neurology , 11, 52-63
- Lu and Liang (2008): Analysis of competing risks data with missing cause of failure under additive hazard model. Statistica Sinica 18, 219-234.
- Crowther and Lampert (2011): Simulating complex survival data. Stata Nordic and Baltic Users‘ Group Meeting
- National Cancer Institute DCCPS Surveillance Research Programme, Surveillance, Epidemiology and End Results (SEER) Programme (www.cancer.seer.gov) Research Data (1973-2009) (Released April 2012)
- Louzda et al. (2012): The Long-Term Bivariate Survival FGM Copula Model: An Application to a Brazilian HIV Data. Journal of Data Science 10, 515-535
- Roman et al. (2012): A New Long-Term Survival Distribution for Cancer Data. Journal of Data Science 10, 242-258
- Sen et al. (2010): A Bayesian approach to competing risks analysis with masked cause of death. Statistics in Medicine, 29, 1681-1695

Thank you for your attention

von behring | röntgen | stiftung

Mon, 7/30/2012, 8:30 AM - 10:20 AM CC-Room 24C

Biometrics Section

Advances in Modeling Competing Risks — Contributed Papers