

How important are the Hazards of Hazard Ratios?

Michal Abrahamowicz

**Distinguished James McGill Professor of Biostatistics
McGill University, Department of Epidemiology & Biostatistics
Montreal, Canada**

with:

Marie-Eve Beauchamp (CAN), Maja Pohar Perme (SLOVEN), Per Kragh Andersen (DK),
Hein Putter (NL), Emily K. Roberts (US), Federico Ambrogi (IT), Pierre Joly (FR),
Richard J. Cook (CAN), Terry M. Therneau (US), Malka Gorfine (ISL),
and Jeremy M. G. Taylor (US)

(on behalf of the STRATOS TG8 Survival Analysis)

Support

Canadian Institutes of Health Research (CIHR) grant PJT-180634
Natural Sciences and Engineering Research Council (NSERC) of Canada grant 228203

Main Reference

(covers most except the Literature review)

Abrahamowicz M, Beauchamp ME, Roberts EK, Taylor JMG.

Revisiting the hazards of hazard ratios through simulations and case studies.

European Journal of Epidemiology 2025; 40:611-625.

(doi.org/10.1007/s10654-025-01245-6)

Outline

- **Background:**
 - M. Hernán's "*Hazards of hazard ratios (HR)*" commentary [*Epidemiology* 2010]: Arguments & Evidence
 - Recent Critiques of HRs & Proportional Hazards (PH) model
 - Outstanding Questions
- **"Generic" Simulations:** Impact of Unmeasured Frailty on HR estimates
- **"Targeted" Simulations:** Plausibility of Hernán's Interpretations ?
- **Literature search** to assess if PH model is compatible with medical data ?
- **Conclusions**

Objectives

- OVERALL Goal:

To assess to what extent the recent Criticisms of HRs & PH model are Justifiable, using Verifiable Empirical Evidence ?

- Specific Aims:

1. Use Simulations to quantify the Frailty-induced **Bias** in HRs
2. Re-assess **M. Hernán's interpretations** of results he presents in "***Hazards of Hazard Ratios***" [*Epidemiology* 2010]
3. Scan recent large medical studies to assess **how often the PH assumption is violated in real-world settings ?**

General Paradigms: inter-connected Challenges for Advanced Analyses of Data on Human Health

1. Choose the Accurate way of MODELLING Complex Processes* underlying the Occurrence, Progression, Treatment and Outcomes of the Disease of interest**
** usually Dynamic & Multi-factorial*
*** use Sensitivity Analyses if No single “optimal” model can be defined*
2. Try to propose Clinically Plausible & “Analytically Valid” INTERPRETATIONS of the RESULTS
(requires Both Statistical & Substantive Expertise)
3. **Methodologists should rely on Verifiable Evidence** (*Not just “opinions”*)

Background I

- Hazards, and Hazard Ratio (HR) estimates based on Cox Proportional Hazards (PH) model have been the mainstays of Survival Analysis for >50 years

[e.g., Andersen et al, *Stat Med* 2021 (STRATOS TG8 guidance paper)]
- Yet, in the past 15 years, influential causal inference experts have severely criticized HRs and the PH model [Hernán *Epid* 2010; Stensrud & Hernán *JAMA* 2020; *Am J Epi* 2025]
- This started in 2010, with M. Hernán's very influential commentary "*Hazards of hazard ratios*" [*Epidemiology* 2010] which warned about the "built-in selection bias" of HRs
- Hernán's commentary, cited >1,700 times, and related more recent comments by Stensrud & Hernán [*JAMA* 2020 (>600 citations); *Am J Epi* 2025], led to concerns about the validity of Cox PH estimates and hazard-based models
- Yet, Evidence presented by Hernán is limited to Selected Results of a Single RCT [Manson et al, *NEJM* 2003]

Background II

- Hernán's cornerstone issue [2010] (well supported by statistical papers since 1980s) is that **HRs suffer from “Built-in Selection Bias”** where:
an unobserved risk factor (‘*Susceptibility*’ or ‘*Frailty*’) automatically induces:
 - (i) **a systematic underestimation (Bias to the Null) of the PH model-based HR for the exposure/treatment**
(even if Randomization ensures balanced distribution of frailty...)
 - (ii) **a gradual decrease of period-specific HRs during follow-up, possibly leading to ‘crossing hazards’**

(*The Only*) Example presented in Hernán's [2010] HRs critique:

CHD risks in RCT of Hormone Therapy

- **RCT of Hormone therapy (Estrogen + Progestin vs Placebo) in primary prevention of Coronary Heart Disease (CHD)**

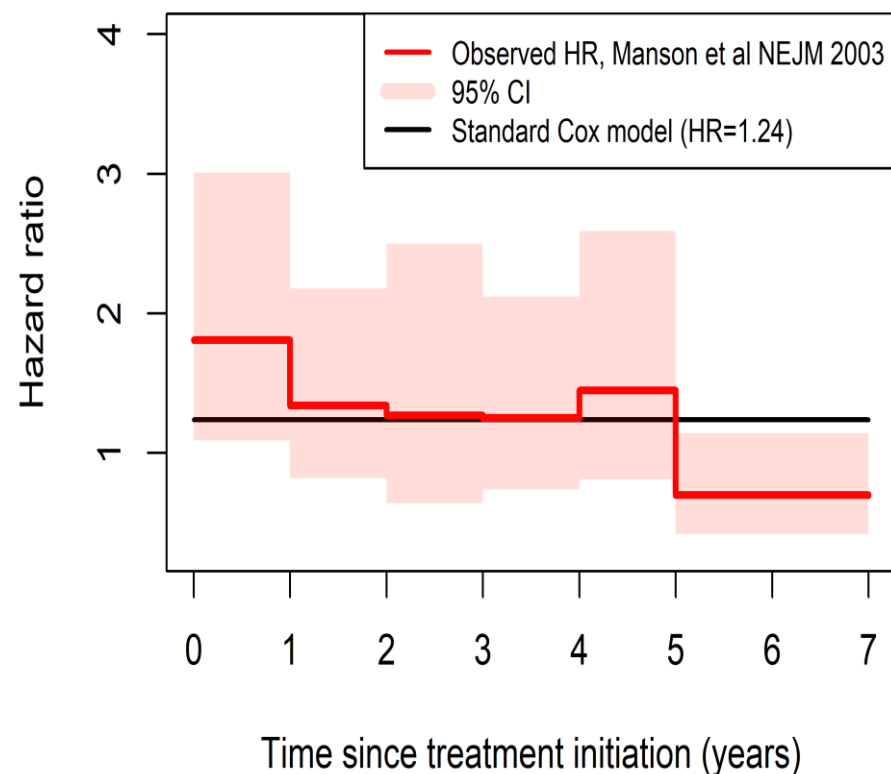
[JA Manson et al, *NEJM* 2003; 349: 523-34]

- **N = 16,608** postmenopausal Women, aged 50-79 yrs at randomization
- Design: Parallel-groups, Double Blinded RCT
- Outcome = **Time to 1st major CHD event** (non-fatal MI or CHD death)
- Follow-up: median 5.6 yrs, maximum 8.6 yrs
- CHD Events: (*Total Incidence = 335 events, only 2.0% !*)
188 in 8,506 women on **Hormone Therapy (3.9/1,000 PY)**
vs.
147 in 8,102 women on **Placebo (3.3/1,000 PY)**

Main Results of Manson et al's [2003] RCT

- ITT analyses reported by Manson et al [2003] using Cox PH model:
HR (Hormone/PB) = 1.24 (95% CI: 0.97-1.60)
- Yet, year-specific HRs decreased 'significantly' during follow-up:
from HR=1.81 in year 1 (1.09-3.01),
to HR = 1.27 in year 3, and then
to HR = 0.7 (0.42-1.14) in years 6-8
[p. 528 & Table 2 in Manson et al, 2003]

Year-specific HRs* in Manson et al's RCT



* Year-specific $HR(t)$ is estimated by: (a) using only women 'at risk' at the start of year t , and (b) censoring women with NO event during year t at the end of the year

Selected Citations (re: **Interpretation of Decreasing HRs**) from M. Hernán's Commentary [*Epidemiology* 2010]:

- “... this apparently protective effect of hormone therapy after year 5 is hardly surprising if one bears in mind that women vary in their susceptibility to heart disease.”
- “... built-in selection bias (of period-specific HRs) makes them difficult to interpret...”
- “Hazards may cross [e.g. from $HR > 1$ to $HR < 1$] at some point because of depletion of susceptibles”
- OVERALL: Hernán argues that important changes over time in year-specific HRs (& “crossing hazards”), observed in Manson et al’s RCT, reflect simply Unobserved Susceptibility (rather than true “causal” changes in the impact of treatment)

Need for more “solid” Evidence?

- However, Hernán’s interpretation is based only on (i) results of a Single Empirical study and (ii) general (qualitative) arguments
- Yet, in many trials, there may be Other, Plausible Reasons for HR(Treated/Placebo) to Change Systematically with Increasing Follow-up Time, including e.g.:
 - Gradually decreasing Adherence (with longer follow-up)
 - Cumulative Effects (*increasing* with treatment duration)
 - Developing Tolerance to treatment effects
 - Effect Modification with Disease Stage/Progression...

Outstanding Questions

- Could the dramatic HRs changes (& “crossing hazards”) observed in Manson et al’s RCT [NEJM 2003] reflect simply Unobserved Susceptibility, i.e. Built-in Selection Bias (as implied by Hernán [2010])?

OR

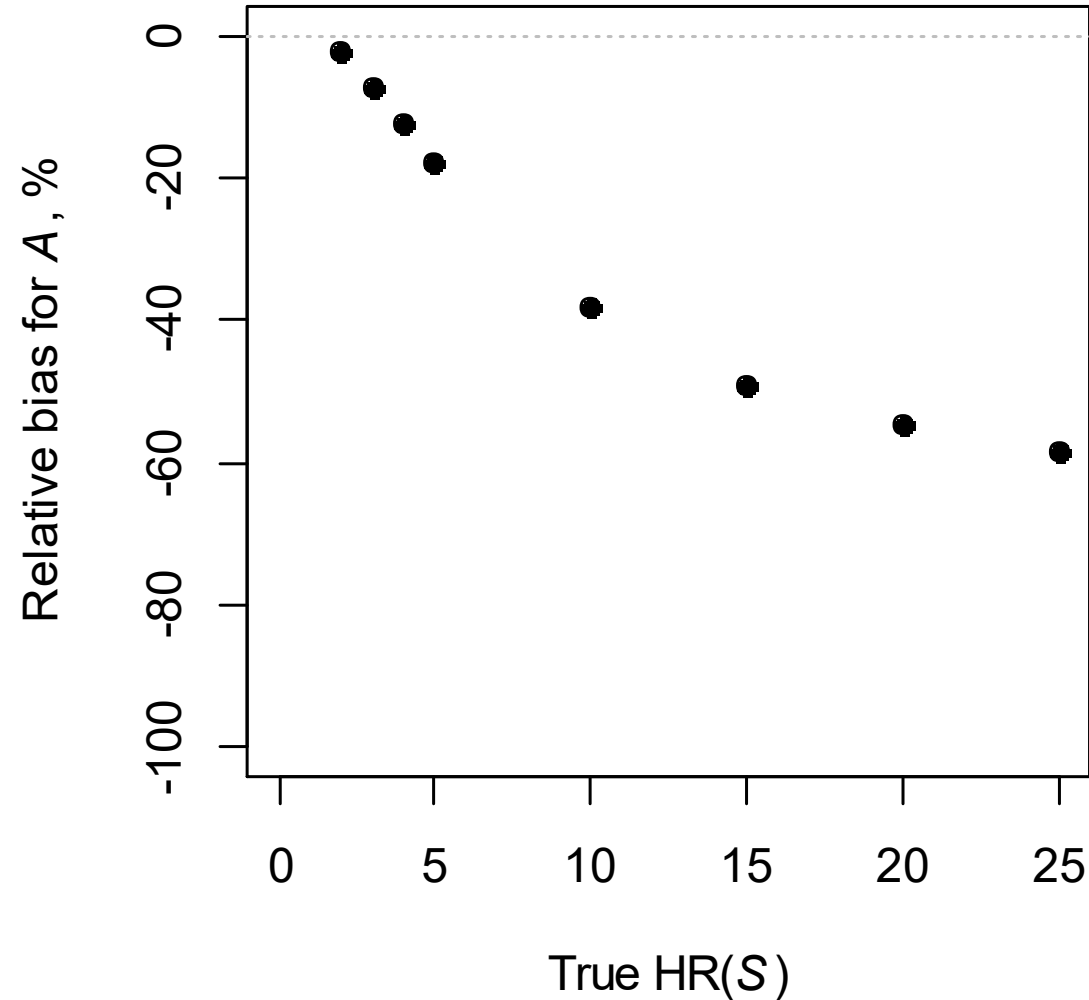
- Do these results reflect True Changes over Time in the Impact of Treatment, due to its true (Biological) Time-Dependent Effect & specific features of this RCT?
- We have addressed these questions by 2 series of Simulations:
“Generic” & “Targeted”

Goals & *Main Results of ‘Generic’ Simulations*

[Abrahamowicz et al, *Eur J Epi* 2025]

- **GOALS:** ‘Generic’ Simulations explored the impact of omitting an important independent risk factor (Frailty S) on treatment’s HR estimates from the Cox PH model
- **RESULTS** show that, for Cox-model HRs: **Unmeasured Frailty (S):**
 - induces (1a) **serious Bias-to-Null**, & (1b) **gradual decay** of year-specific HRs, **ONLY IF it has a VERY STRONG Impact on the Hazard** (e.g., $HR(S) > 5$ for Binary S)
 - (2) has a much **Weaker Impact IF:**
 - (2a) **Prevalence ($P(S)=1$) is Low And/Or**
 - (2b) **Overall Survival is High (implying Low Incidence of the events)**
- These Empirical findings motivated us to Re-Assess ** the Plausibility of M. Hernán’s Interpretation of Manson et al’s RCT results (** using ‘Targeted’ Simulations that follow)

“Generic” Simulations: $HR(S)$ vs mean relative Bias in Tx (A) $\log(HR)$:
(1000 samples: $HR(A) = 2$; $P(S=1) = 0.5$, Cum. Inc. = 35%-60%)



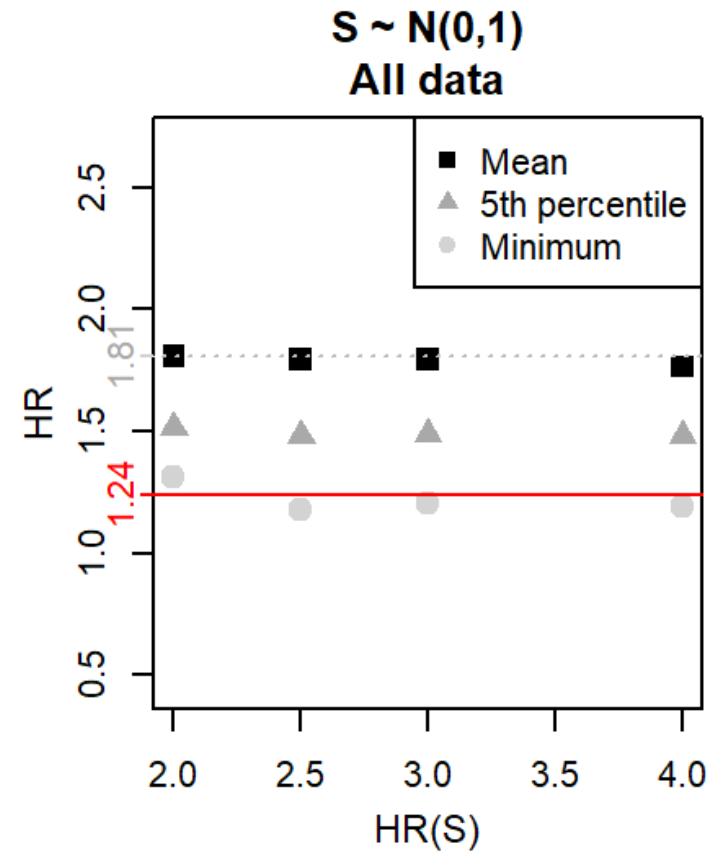
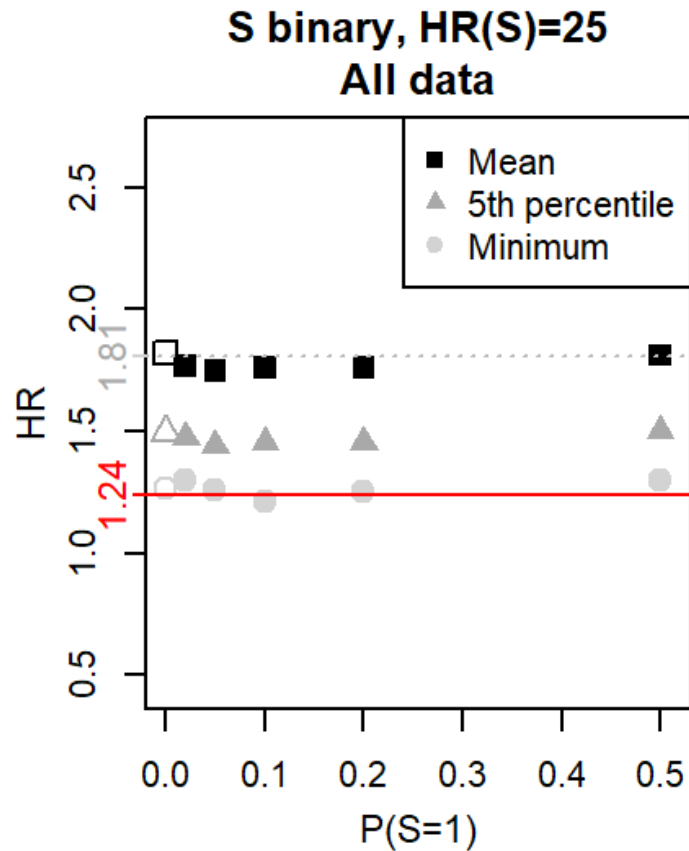
“Targeted” Simulations to Re-assess Manson’s RCT results: main Assumptions

GOAL: to assess IF results reported in Manson et al’s [NEJM 2003] RCT could be entirely due to **Unmeasured Frailty**, as postulated by Hernán [Epid 2010] ?

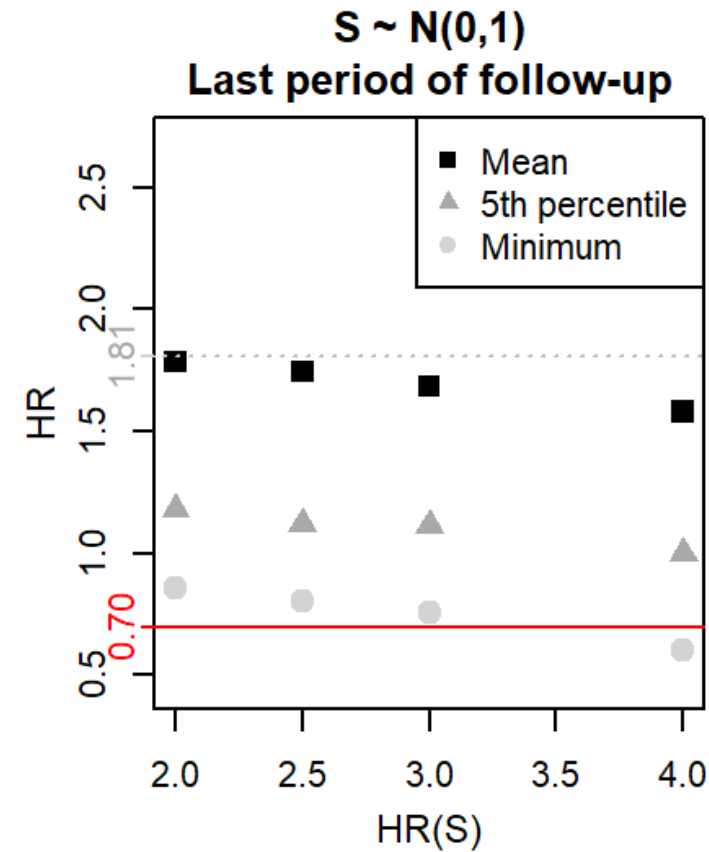
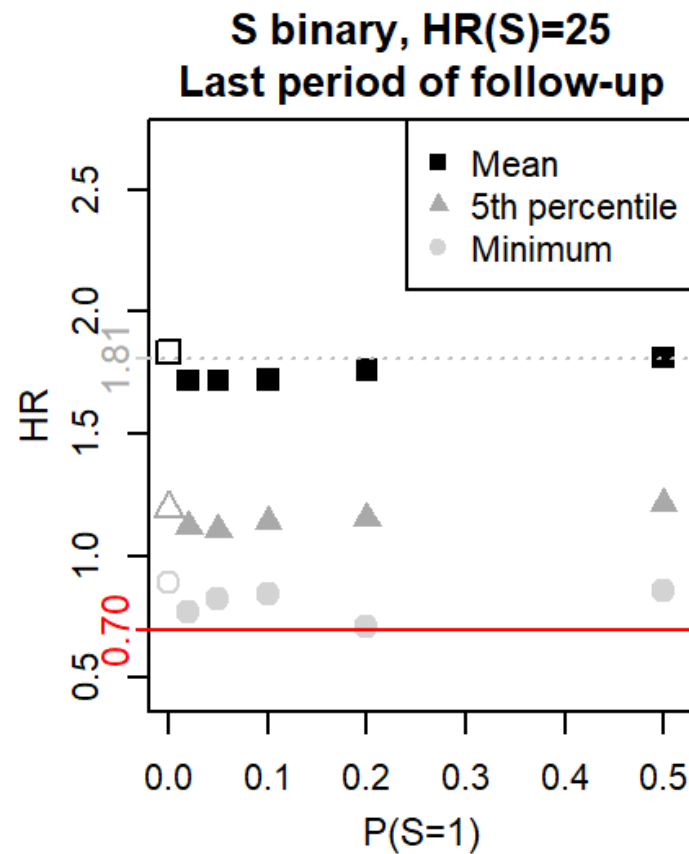
Thus, our ‘Targeted’ simulations replicated all important features of the original RCT:

- N = 16,600 ($P(A=1) = P(A=0) = 0.5$); 10 years of follow-up
- Overall Cumulative Incidence = 2.0% (335 Events), ~ 65 events in the 2 last years
- True Treatment effect: $HR(A) = 1.81$ (1st-yr estimate, assumed to be the ‘true’ impact of the hormone therapy [Hernán, *Epidemiology* 2010])
- Finally, to assess the most ‘extreme’ scenarios, we assumed **Very Strong Impact of Unmeasured S** ($HR(S)=25$ for a Binary S; or $HR=2$ to 4 for 1 SD increase in Continuous S) while varying the distribution (e.g. prevalence) of S

Simulation results: Even with the most exceptionally strong frailty (e.g. $HR(S)=25$ for Binary S in the Left panel) the overall HR is extremely unlikely to be as low as 1.24 reported by Manson et al!!



Similarly, year-specific HR for the last period is extremely unlikely to be as low as 0.70 reported by Manson!!



Possible Alternative Explanations of Manson's RCT results: **increasing Tx Non-Adherence ?**

- Importantly, Manson et al report [NEJM, p.526]: “relatively **High Rate of Discontinuation* of Hormone Tx**” [***42% by yr 5 !**] ^[1]
- And their secondary **analyses on Fully Adherent women yielded HR = 1.50 (1.14-1.97)** [p. 528],
i.e. **50% stronger effect than ITT-based HR=1.24 (0.97-1.54)!** ^[1]
- Thus, ~ 50% of “Bias” leading to Decaying-over-time HRs in Manson et al's RCT can be explained by Increasing-over-Time Non-Adherence with Hormone Therapy ^[1]

^[1] None of these facts is mentioned by Hernán [Epidemiology 2010]

Possible CLINICAL/Biological **Explanations** of Manson et al's RCT results ?

- **Yet, HR=1.50 (1.14-1.97) for Fully Adherent women is still considerably lower than the 1st-year HR=1.81 (1.09-3.01) [Manson, *NEJM* 2003]. Thus:**
Non-Adherence *cannot* fully explain the decrease in Tx HRs in later years!
And there may be more 'substantive' clinical/biological reasons...
- Recently, Prentice & Aragaki [*Stat Sciences* 2022] presented time-varying re-analyses of the Manson et al's RCT and suggested searching for the biological cause for the harmful hormone Tx impact decaying over time
- **Indeed, imaging studies show that hormone therapy reduces coronary artery plaque but only after several years of Tx** [Akhrass et al., *J Clin Endocrinol Metab* 2003; Barrett-Connors et al. *Menopause* 2005]
- Further, a comprehensive review reports that **decaying, and possibly 'crossing' HRs for Hormone Tx** (i) **are consistent with 2 other RCTs of same Tx** & a large cohort (Nurses Studies), and (ii) **may have a biological basis** [Hodis & Mack 2007]

Benefits of flexible modeling of time-varying HRs

- Interestingly, **Hernán does Not mention a possibility of a genuine biological change in the treatment effect [Epidemiology 2010] !**
- Yet, since 1990's statisticians working in Survival Analysis developed & validated **several flexible extensions of Cox's model that allow for Time-Dependent (TD) Hazard Ratios ** (with $\log(\text{HR}(t)) = \beta(t)$ modeled using different Splines)** e.g.:

[Zucker & Karr, *Ann Stat* 1990]; [Gray, *JASA* 1992]; [Hastie & Tibshirani, *JRSS* 1993];
[Hess, *Statistics in Medicine* 1994]; [Grambsch & Therneau, *Biometrics* 1994];
[Verweij & van Houwelingen, *Biometrics* 1995]; [Kooperberg, Stone & Truong, *JASA* 1995];
[Luo, He, Wu & Taylor, *SMMR* 2023]; [Abrahamowicz, MacKenzie & Esdaile *JASA* 1996]...
- **Hundreds of real-life applications of these models**, in many diseases, yielded strong evidence of Time-Dependent HRs for several Prognostic Factors, Exposures and Treatments**
- **Could they All be due Exclusively to Unmeasured Susceptibilities ??**

(** **Hernán & Stensrud argued AGAINST such TD analyses [JAMA 2020]...**)



Are “*hazards not expected to be proportional in almost any medical study*” ?

- M. Stensrud & M. Hernán’s *JAMA* 2020 commentary (>600 citations) includes a section “*Why are hazards usually non-proportional in medical studies*” which concludes with: “*These examples illustrate why **hazards are not expected to be proportional in almost any medical study***”
- Yet, **drawing a general conclusion based on just 3, highly selected, examples is NOT consistent with basic principles of Deductive Logic !!**
- To provide a more systematic evidence on this (contentious) point, we carried a structured Review of recent large Real-World Epidemiological/Clinical studies that used Cox PH model

Outline of Methods for our Literature Review

re: “*how often the PH assumption is violated in recent large medical studies*” ?

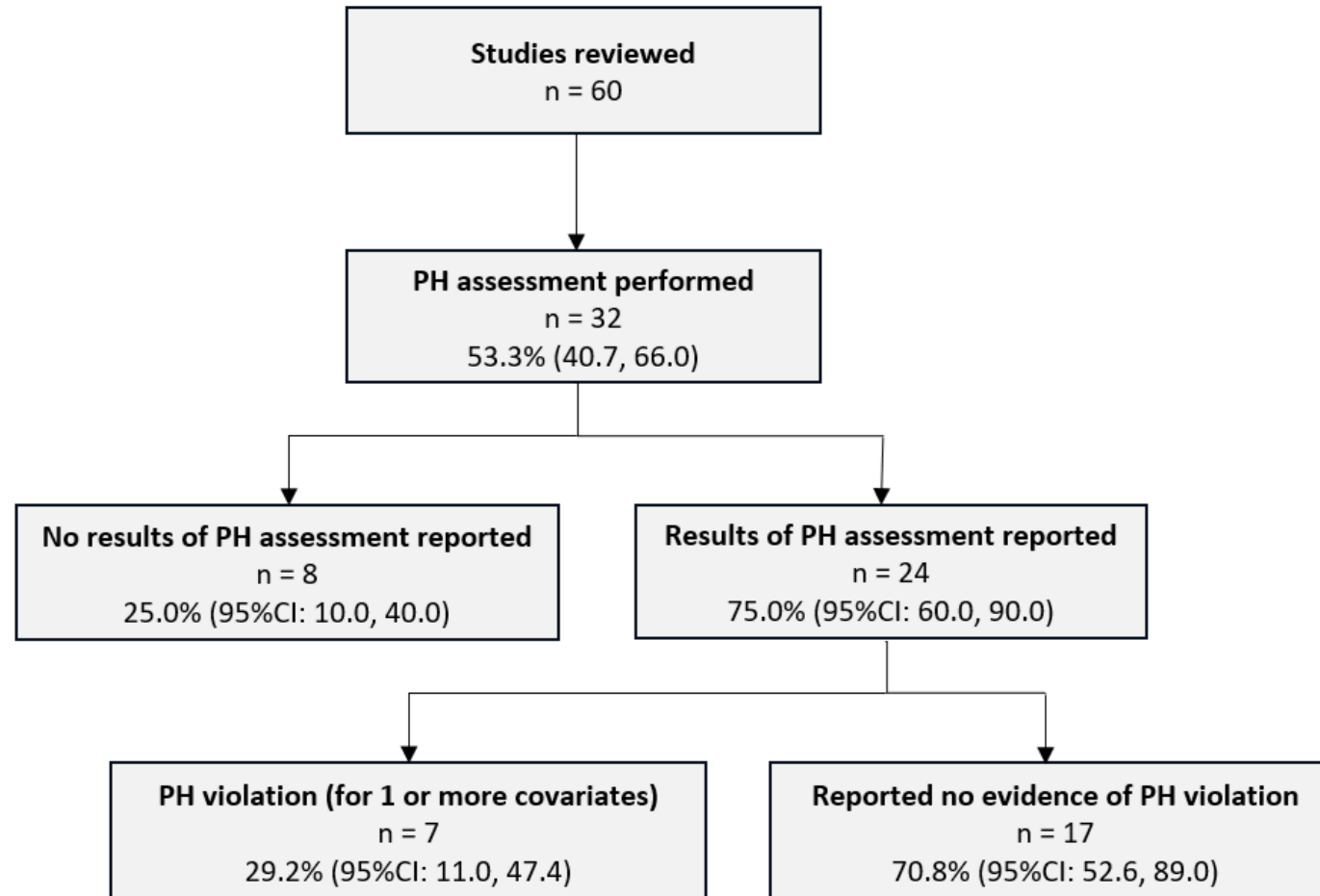
➤ Inclusion criteria & sampling:

- We used stratified sampling to get 15 papers (for a Total of 60 papers), published in 2021-26, which employed Cox PH model, from each of the 4 strata of top-ranking journals in:
(1) General medicine, (2) Epidemiology, (3) Cancer, (4) Cardiovascular research
- All 60 studies had to report >300 events (42/60 reported >1,000 events !)

➤ Review criteria focused on:

- (a) **if the PH assumption was assessed ?**
- (b) **if so, what methods were employed ?**
- (c) **whether there was evidence of violations of the PH assumption ?**

Frequency of PH assumption Violations (~29%) in recent large studies in High-ranking Clinical & Epi journals



Main findings of Literature Review

- Concern: **Even in 2020's many (28/60 = 47%) large time-to-event studies, published in top-ranking clinical & epi journals, do Not explicitly mention assessing the PH assumption** even when relying on Cox model for their main analyses!**
- ** (On the plus side: there is a *clear improvement compared to early 1990's when Altman et al [Br J Cancer 1985] found this proportion to be only ~5% !*)
- Those studies that did assess the PH hypothesis relied mostly on *graphical methods* and/or (less frequently) *time*exposure interactions*
- Most importantly: **only 29% (7/24) of the 24 large recent clinical/epi studies that reported results of assessment of the PH assumption found it was violated.**
- This solid empirical evidence is in a Sharp Contrast to (undocumented) claims that *“hazards are not expected to be proportional in almost any medical study”* [Stensrud & Hernán, JAMA 2020]



Conclusions

- **General:**
 - **Using Empirical Results (especially of a *Single Study*...) to support Methodological Postulates requires a careful examination of All relevant aspects of the study (including limitations of data and/or analyses)**
 - **Claims about Methods' “Implausibility” should be supported by Solid Systematically collected and assessed Evidence (rather than by a few hand-picked examples) !**
- **Specific:**
 - Concerns expressed in Epidemiological literature about (i) “*Hazards of Hazard Ratios*” and (ii) Implausibility of the PH model are both exaggerated and not well supported by evidence
 - Flexible modeling of time-varying HRs, combined with Clinical/Biological expertise about the underlying exposure-outcome relationships, may yield important new insights into the etiology, progression or treatment of many diseases



Thank you = **VielenDank**

Michal.Abrahamowicz@McGill.ca

References

- Abrahamowicz M, Beauchamp ME, Roberts EK, Taylor JMG. Revisiting the hazards of hazard ratios through simulations and case studies. *Eur J Epidemiol* 2025; 40:611-25.
- Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *J Am Stat Assoc* 1996;91:1432-9.
- Akhrass F, Evans AT, Wang Y, et al. Hormone replacement therapy is associated with less coronary atherosclerosis in postmenopausal women. *J Clin Endocrinol Metab* 2003;88(12):5611-4.
- * Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer*. 1995;72(2):511-8. <https://doi.org/10.1038/bjc.1995.364-66>.
- Andersen PK, Perme MP, van Houwelingen HC, Cook RJ, Joly P, Martinussen T, Taylor JMG, Abrahamowicz M, Therneau TM. Analysis of time-to-event for observational studies: Guidance to the use of intensity models. *Stat Med* 2021; 40(1): 185–211.
- Barrett-Connor E, Laughlin GA. Hormone therapy and coronary artery calcification in asymptomatic postmenopausal women: the Rancho Bernardo Study. *Menopause* 2005;12(1):40-8.
- Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med* 2015;373(17):1627–39.
- Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrics* 1994;81:515-526.
- Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J Am Stat Assoc* 1992;87:942-951.
- Hastie T, Tibshirani R. Varying-coefficient models (with discussion). *J R Stat Soc Ser B Methodol* 1993;55:757-796.

References

- Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010;21(1):13-5.
- Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat Med* 1994;13:1045-1062.
- Hodis HN, Mack WJ. Randomized controlled trials and the effects of postmenopausal hormone therapy on cardiovascular disease: facts, hypotheses and clinical perspective. In: Lobo RA, editor. *Treatment of the Postmenopausal Woman*. 3rd edition ed. Philadelphia, PA: Elsevier Academic Press; 2007. p. 529–64.
- Kooperberg C, Stone CJ, Truong YK. Hazard regression. *J Am Stat Assoc* 1995;90:78-94.
- Luo L, He K, Wu W, Taylor JMG. Using Information Criteria to Select Smoothing Parameters when Analyzing Survival Data with Time-Varying Coefficient Hazard Models. *Stat Methods Med Res* 2023;32:1664-79.
- Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med* 2003;349(6):523-34.**
- Prentice RL, Aragaki AK. Intention-to-treat comparisons in randomized trials. *Stat Sci* 2022; 37(3)-380-393.
- Roberts EK, Luo L, Mondul AM, et al. Time-varying associations of patient and tumor characteristics with cancer survival: an analysis of SEER data across 14 cancer sites, 2004–2017. *Cancer Causes Control*. 2024.
- Stensrud MJ, Hernán MA. Why test for proportional hazards? *JAMA* 2020;323(14):1401-2.
- Stensrud MJ, Hernán MA. Why use methods that require proportional hazards? *Am J Epidemiol* 2025; 194:1504-6.
- Verweij PJ, Houwelingen HC. Time-dependent Effects of Fixed Covariates in Cox Regression. *Biometrics* 1995;51:1550-1556.
- Zucker DM, Karr AF. Nonparametric Survival Analysis with Time-Dependent Covariate Effects: A Penalized Partial Likelihood Approach. *Ann Statist* 1990;18:329-53.

In Simulations: **Non-Adherence**** consistent with Manson's RCT
(6% of women in Hormone Tx arm stop Tx each year > 42% after 7 yrs)
(**Rightmost graph) **impacted the estimates much more than even very strong S**

