**Accounting for delayed entry in analyses of overall survival in clinico-genomic databases**

Daniel Backenroth,Jeremy Snider, Ronglai Shen,Venkat Seshan, Emily Castellanos, Margaret McCusker, Dana Feuchtbaum, Mithat Gönen, Somnath Sarkar.

# Supplementary Materials and Methods

## Assumptions of risk set adjustment method

Risk set adjustment is premised on an assumption of "independent delayed entry," i.e., that the hazard of death at times after cohort entry does not depend on when a patient entered (2). Assume that to each patient there is associated a survival time X and an entry time V. Patients for whom V is less than X are observed in the follow-up study, whereas other patients are not, since the patient did not satisfy the sampling requirements before death. Such a follow-up study is said to be subject to "left truncation" given the exclusion of patients with V≥X. Assume that the density of X is given by f(x) and the density of V by g(v). Independent delayed entry requires that the conditional joint density of (V, X) given V < X is proportional to f(x)g(v) for v<x (Tsai).

If NGS testing affected survival after sequencing then one would not expect to have independent delayed entry in a clinico-genomic database, since the conditional density of (V,X) would likely also depend on the difference between V and X, i.e., the length of the time span in which physician choices relating to patient treatment were potentially affected by the results of the NGS testing. However, depending on the magnitude of the effect of NGS testing on survival, the resulting deviations from independent delayed entry may be minimal.

Risk set adjustment excludes patients from the risk set at all times prior to V and, under independent delayed entry, will properly estimate the survival function S(X) (3). Technically, risk set adjustment will estimate the distribution of X conditional on X being greater than or equal to the infimum of the support of V (i.e., we cannot say anything about the survival of patients prior to the earliest entry time), and on V being less than or equal to the supremum of the support of X (i.e., we cannot say anything about the survival of patients who have zero chance of ever being observed) (Woodroofe p. 165 and Theorem 2). In the absence of independent delayed entry, however, the survival function S(X) of X is unidentifiable, and the survival function resulting from risk set adjustment does not have a clear interpretation.

## References

1. Zhang, Q, Gossai, A, Monroe, S, Nussbaum, NC, Parrinello, CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv Res.* 2021; 1– 7.
2. Keiding N, Moeschberger M. Independent delayed entry. In: *Survival analysis: State of the art.* Springer; Dordrecht. 1992:309-326.
3. Woodroofe M. Estimating a distribution function with truncated data*. Annals Statist*. 1985;13(1):163-177.

##

##

## Supplementary Table 1

**Supplementary Table 1.** Non–Small Cell Lung Cancer (Advanced) – Comparison of Demographic and Clinical Characteristic between CGDB vs FHRD vs SEER

| **Characteristic** | **All eligible, n (%)** | **Stage IV at diagnosis, n (%)** |
| --- | --- | --- |
| **SEER** | **FHRD** | **CGDB** | **SEER** | **FHRD** | **CGDB** |
| **2011–18** | **2011–18** | **2011–20** | **2011–18** | **2011–20** | **2011–18** | **2011–18** | **2011–20** | **2011–18** | **2011–20** |
| **N=292,179** | **N=54,222** | **N=65,348** | **N=9,517** | **N=13,604** | **N=122,539** | **N=33,828** | **N=42,008** | **N=5,006** | **N=7,581** |
| **Age at initial diagnosis (years)** |
| 0–19 | 18 (<0.1) | 2 (<0.1) | 2 (<0.1) | 0 (0) | 0 (0) | 7 (<0.1) | 2 (<0.1) | 2 (<0.1) | 0 (0) | 0 (0) |
| 20–34 | 475 (0.2) | 106 (0.2) | 127 (0.2) | 42 (0.4) | 49 (0.4) | 322 (0.3) | 83 (0.2) | 103 (0.2) | 30 (0.6) | 37 (0.5) |
| 35–49 | 8,490 (2.9) | 1,897 (3.5) | 2,233 (3.4) | 480 (5.0) | 607 (4.5) | 4,905 (4.0) | 1,339 (4.0) | 1,615 (3.8) | 318 (6.4) | 412 (5.4) |
| 50–64 | 79,831 (27) | 16,536 (30) | 19,669 (30) | 3,204 (34) | 4,332 (32) | 37,708 (31) | 10,716 (32) | 13,025 (31) | 1,805 (36) | 2,549 (34) |
| 65+ | 203,365 (69.6) | 35,681 (65.8) | 43,317 (66.3) | 5,791 (60.8) | 8,616 (63.3) | 79,597 (65.0) | 21,688 (64.1) | 27,263 (64.9) | 2,853 (57.0) | 4,583 (60.5) |
| **Sex** |
| Female | 139,010 (48) | 25,564 (47) | 30,918 (47) | 4,859 (51) | 6,913 (51) | 56,171 (46) | 15,750 (47) | 19,751 (47) | 2,496 (50) | 3,773 (50) |
| Male | 153,169 (52) | 28,654 (53) | 34,425 (53) | 4,658 (49) | 6,691 (49) | 66,368 (54) | 18,077 (53) | 22,255 (53) | 2,510 (50) | 3,808 (50) |
| Unknown | 0 (0) | 4 (<0.1) | 5 (<0.1) | 0 (0) | 0 (0) | 0 (0) | 1 (<0.1) | 2 (<0.1) | 0 (0) | 0 (0) |
| **Race** |
| Asian | 21,553 (7.4) | 1,376 (2.5) | 1,691 (2.6) | 284 (3.0) | 407 (3.0) | 10,523 (8.6) | 972 (2.9) | 1,214 (2.9) | 184 (3.7) | 255 (3.4) |
| Black | 34,511 (12) | 4,645 (8.6) | 5,616 (8.6) | 548 (5.8) | 834 (6.1) | 15,763 (13) | 2,832 (8.4) | 3,528 (8.4) | 296 (5.9) | 479 (6.3) |
| Other | 1,611 (0.6) | 5,037 (9.3) | 6,335 (9.7) | 1,323 (14) | 1,997 (15) | 679 (0.6) | 3,203 (9.5) | 4,175 (9.9) | 662 (13) | 1,087 (14) |
| Unknown | 765 (0.3) | 5,705 (11) | 7,253 (11) | 767 (8.1) | 1,135 (8.3) | 235 (0.2) | 3,904 (12) | 5,052 (12) | 427 (8.5) | 659 (8.7) |
| White | 233,739 (80) | 37,459 (69) | 44,453 (68) | 6,595 (69) | 9,231 (68) | 95,339 (78) | 22,917 (68) | 28,039 (67) | 3,437 (69) | 5,101 (67) |
| **AJCC stage at diagnosis** |
| Stage 0 | 9 (<0.1) | 5 (<0.1) | 5 (<0.1) | 1 (<0.1) | 1 (<0.1) | NA | NA | NA | NA | NA |
| Stage I | 71,448 (24) | 4,688 (8.6) | 4,995 (7.6) | 1,155 (12) | 1,419 (10) | NA | NA | NA | NA | NA |
| Stage II | 19,762 (6.8) | 2,853 (5.3) | 3,072 (4.7) | 800 (8.4) | 1,020 (7.5) | NA | NA | NA | NA | NA |
| Stage III | 64,342 (22) | 11,309 (21) | 13,606 (21) | 2,199 (23) | 3,077 (23) | NA | NA | NA | NA | NA |
| Stage IV | 122,539 (42) | 33,828 (62) | 42,008 (64) | 5,006 (53) | 7,581 (56) | NA | NA | NA | NA | NA |
| Unknown or NA | 14,079 (4.8) | 1,539 (2.8) | 1,662 (2.5) | 356 (3.7) | 506 (3.7) | NA | NA | NA | NA | NA |
| **Year of initial diagnosis** |
| 2011 | 36,193 (12) | 4,910 (9.1) | 4,910 (7.5) | 244 (2.6) | 244 (1.8) | 14,871 (12) | 2,707 (8.0) | 2,707 (6.4) | 82 (1.6) | 82 (1.1) |
| 2012 | 36,396 (12) | 5,942 (11) | 5,942 (9.1) | 381 (4.0) | 381 (2.8) | 15,249 (12) | 3,460 (10) | 3,460 (8.2) | 112 (2.2) | 112 (1.5) |
| 2013 | 36,526 (13) | 6,749 (12) | 6,749 (10) | 620 (6.5) | 620 (4.6) | 15,303 (12) | 4,140 (12) | 4,140 (9.9) | 280 (5.6) | 280 (3.7) |
| 2014 | 36,776 (13) | 7,211 (13) | 7,211 (11) | 984 (10) | 984 (7.2) | 15,460 (13) | 4,507 (13) | 4,507 (11) | 479 (9.6) | 479 (6.3) |
| 2015 | 37,119 (13) | 7,625 (14) | 7,625 (12) | 1,440 (15) | 1,440 (11) | 15,312 (12) | 4,796 (14) | 4,796 (11) | 744 (15) | 744 (9.8) |
| 2016 | 37,214 (13) | 7,484 (14) | 7,484 (11) | 1,663 (17) | 1,663 (12) | 16,222 (13) | 4,733 (14) | 4,733 (11) | 882 (18) | 882 (12) |
| 2017 | 37,708 (13) | 7,373 (14) | 7,373 (11) | 1,939 (20) | 1,939 (14) | 16,020 (13) | 4,862 (14) | 4,862 (12) | 1,092 (22) | 1,092 (14) |
| 2018 | 34,247 (12) | 6,928 (13) | 6,928 (11) | 2,246 (24) | 2,246 (17) | 14,102 (12) | 4,623 (14) | 4,623 (11) | 1,335 (27) | 1,335 (18) |
| 2019 | NA | NA | 6,187 (9.5) | NA | 2,077 (15) | NA | NA | 4,273 (10) | NA | 1,258 (17) |
| 2020 | NA | NA | 4,558 (7.0) | NA | 1,758 (13) | NA | NA | 3,597 (8.6) | NA | 1,135 (15) |
| 2021 | NA | NA | 381 (0.6) | NA | 252 (1.9) | NA | NA | 310 (0.7) | NA | 182 (2.4) |
| Abbreviations: AJCC, American Joint Committee on Cancer; CGDB, Flatiron Health-Foundation Medicine Clinico-Genomic Database; FHRD, Flatiron Health Research Database; NA, not applicable; NSCLC, non-small cell lung cancer; SEER, The Surveillance, Epidemiology, and End Results research database.Courtesy of Snow T, Snider J, Comment L, *et al*. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health-Foundation Medicine Clinico-Genomic Database, Flatiron Health Research Database, and National Cancer Institute SEER Population-Based Cancer Registry, medRxiv *in preparation*.  |