Optimal Data Analysis
Vol. 9 (July 7, 2020), 214-217

2155-0182/10/$3.00

Differing Cancer-Incidence Rates of Male *vs*. Female Americans

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Novometric classification tree analysis was used to evaluate Surveillance, Epidemiology, and End Results (SEER) Program data to discover cancer sites moderately or relatively strongly predicted by male *vs*. female gender. Future research using any of the 13 cancer sites which met this criterion should account for gender using matching or propensity score weighting.

Data are from the Surveillance, Epidemiology, and End Results (SEER) Program, that collects and publishes cancer incidence/survival data to assemble and to report estimates of cancer incidence, survival, mortality, other measures of the cancer burden, and patterns of care in the USA.[1] SEER routinely includes information specific to populations defined by age, gender, geography, and race/ethnicity. This study parses SEER cancer incidence data by gender (male, female), to assess if this identifies patient strata which differ in cancer incidence.[2] Cancer incidence rate is number of new cancers of a specific site (type) that occur in a population in one year, expressed as number of new cancers for every 100,000 in the population who are at risk. A third ethnic category in the SEER database was omitted as it combines heterogeneous race classes which may induce paradoxical confounding if combined.[3-6]

Every individual cancer category in the SEER database was evaluated to ascertain if the male and female patients had equivalent cancer-incidence rates. Most cancer categories had 608 observations, so proportionally reducing sample size using three sequential binary parses creates 2^3=8 strata, each with 76 observations. Analysis involving a binary class variable and an ordered attribute, with endpoints each having 76 observations, requires moderate ESS=48 to achieve $\geq$90% power to detect two-tail $p<0.05$.[6-11]

Novometric CTA[12] was used to evaluate each cancer-incidence category using gender as the binary class variable and cancer-incidence rate as the ordered attribute; weighting by prior odds was used to maximize ESS; the minimum denominator selection algorithm was employed to identify the descendant family; and identical findings were required in training and in LOO validity analysis.[13-20] Findings are summarized by ESS point estimates and 95% CIs for cancer-incidence sites which reflect a moderate to relatively-strong (Table 1), or a moderate degree (Table 2) of male *vs*. female disparity.

Future research using cancer sites found moderately or more strongly related to male *vs*. female gender should account for gender using matching[22-24] during subject recruitment, or via weighting by propensity scores otherwise.[25-27]

2155-0182/10/$3.00

Findings reported presently should be replicated if more recent data relating gender and cancer-incidence exist.

Table 1: *Moderate* to *Relatively-Strong* Disparity

Cancer Site	Strata	MinD	ESS	D
Kaposi Sarcoma	2	192	47.4	2.22
			39.8-55.0	1.63-3.03
			0-7.24	----
Larynx	2	163	43.8	2.57
			36.4-51.0	1.92-3.49
			0.33-6.91	26.9-604
Thyroid	2	210	43.4	2.61
			35.4-51.3	1.90-3.65
			0-7.24	----

Note: The results for every step of MDSA analysis[12] are tabled. Cancer Site is type of cancer; Strata is number of CTA model endpoints; and MinD (minimum denominator) is the smallest sample size for any strata. ESS (effect strength for sensitivity) is a normed index: 0=the classification accuracy expected by chance; 100=perfect accuracy. The D (*distance*) statistic[17,21] adjusts ESS for model complexity, indicating the number of additional effects having equivalent mean ESS that are needed to obtain 100% correct classification (dashes indicate division by zero). For every model in the descendant family the first line gives point estimates; second line gives 95% confidence intervals (CIs) for discrete distributions obtained by using bootstrap analysis (50% with replacement)[7,18]; and the third line presents 95% CIs for chance obtained via Monte Carlo analysis using 100,000 iterations. Unless otherwise noted, N=608 for analyses reported within this paper.

Table 2: *Moderate* Disparity

Cancer Site	Strata	MinD	ESS	D
Oral Cavity and Pharynx	2	120	39.5	3.06
			33.1-46.0	2.35-4.04
			0-6.58	----
Endocrine System	2	224	38.8	3.15
			30.2-47.1	2.24-4.62
			0-7.89	----
Esophagus	2	113	32.6	4.13
			26.1-39.2	3.10-5.66
			0.33-6.25	30-604
Floor of Mouth	2	124	33.6	3.95
			26.6-40.5	2.94-5.52
			0-6.58	----
Hodgkin Lymphoma	2	289	36.5	3.48
			27.8-45.4	2.41-5.19
			0.33-7.57	0.16-3.78
Hodgkin-Nodal	2	292	36.8	3.43
			28.2-45.4	2.41-5.09
			0-7.89	----
Hypopharyanx	2	136	38.8	3.15
			31.8-46.0	2.35-4.29
			0-6.58	----
Oropharynx	2	143	34.5	3.80
			27.1-41.7	2.80-5.38
			0.33-6.91	26.9-604
Tongue	2	120	37.5	3.33
			30.8-44.1	2.27-4.49
			0-6.58	----
Tonsil	2	162	37.5	3.33
			30.0-45.1	2.43-4.67
			0-7.24	----

See Note to Table 1.

References

[1]Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, based on the November 2011 submission.

2155-0182/10/$3.00

[2]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[3]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56*, 430-442.

[4]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, *1*, 67-100.

[5]Yarnold PR (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis*, *2*, 159-171.

[6]Bryant FB, Siegel EKB (2013). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v. Illinois Bell*. *Optimal Data Analysis*, *1*, 176-198.

[6]Rhodes NJ (2020). Statistical power analysis in ODA, CTA and novometrics (Invited). *Optimal Data Analysis*, *9*, 21-25.

[7]Rhodes JN, Yarnold PR (2020). Generating novometric confidence intervals in R: Bootstrap analyses to compare model and chance ESS. *Optimal Data Analysis*, *9*, 172-177.

[8]Rhodes NJ (2020). Assessing reproducibility of novometric bootstrap confidence interval analysis using multiple seed numbers (Invited). *Optimal Data Analysis*, *9*, 190-194.

[9]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.

[10]Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, *54*, 646-653.

[11]Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, *74*, 223-238.

[12]Yarnold PR (2016). How many EO-CTA models exist in my sample and which is the best model? *Optimal Data Analysis*, *5*, 62-64.

[13]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23*, 703-712. DOI: 10.1111/jep.12744

[14]Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, *3*, 30-31.

[15]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[16]Yarnold PR (2020). What is novometric data analysis? *Optimal Data Analysis*, *9*, 195-206.

[17]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[18]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

2155-0182/10/$3.00

[19]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis*, *5*, 3-10.

[20]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis*, *5*, 11-14.

[21]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[22]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, *22*, 848-854. DOI: 10.1111/jep.12538

[23]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*, 868-874. DOI: 10.1111/jep.12592

[24]Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1309-1315. DOI: 10.1111/jep.12792

[25]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885. DOI: 10.1111/jep.12610

[26]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23*, 703-712. DOI: 10.1111/jep.12744

[27]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, *6*, 43-46.

Author Notes

This study analyzed publically available data and thus was exempt from Institutional Review Board review. No conflicts of interest were reported.