Optimal Data Analysis
Vol. 9 (July 7, 2020), 210-213

2155-0182/10/$3.00

Disparate Cancer-Incidence Rates of Caucasian *vs*. African Americans

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

Surveillance, Epidemiology and End Results (SEER) Program data were used to find cancer sites with at least moderately different rates for African *vs*. Caucasian Americans. Future research in ten cancer sites which involves subjects represented by these groups should account for associated cancer-incidence disparity in matching or via propensity score weighting methods.

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. SEER routinely includes information specific to populations defined by age, gender, geography, and race/ethnicity.[1]

This study uses SEER cancer incidence data, parsed by Caucasian *vs*. African American, to assess if this class variable identifies patient strata with different cancer incidence.[2] Cancer incidence rate is the 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 category in the SEER database was omitted because it combines heterogeneous race classes which may induce paradoxical confounding if they are combined.[3-6]

Every individual cancer category in the SEER database was evaluated to determine if the two patient groups had comparable 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 race as the binary class variable and cancer-incidence rate as the ordered attribute; weighting by prior odds was used so as 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]

Point estimates and exact discrete 95% CIs for ESS are given for cancer sites reflecting *relatively-strong* to *strong* disparity between Caucasian and African Americans in Table 1; for sites reflecting *moderate* to *relatively-strong* disparity in Table 2; and for one site reflecting *moderate* disparity between groups in Table 3.

2155-0182/10/$3.00

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

Cancer Site	Strata	MinD	ESS	D
Eye and Orbit	7	27	71.1	2.85
			64.3-77.6	2.02-3.89
			0-7.89	----
	6	47	68.1	2.81
			61.3-74.6	2.04-3.79
			0.33-7.57	73.3-1812
	5	51	62.8	2.96
			55.6-70.1	2.13-3.99
			0.33-7.57	61.1-1510
	2	202	62.5	1.20
			55.8-69.0	0.90-1.58
			0-7.24	----
Melanoma of the Skin	5	25	75.3	1.64
			69.2-81.2	1.16-2.23
			0.33-8.22	300-1510
	4	63	71.4	1.60
			64.4-77.8	1.14-2.21
			0.33-8.22	44.7-4996
	2	234	67.1	0.98
			60.4-73.7	0.71-1.31
			0-7.89	----
Skin excluding Basal and Squamous	5	14	66.1	2.56
			59.2-72.9	1.86-3.45
			0.33-7.57	61.1-1510
	4	51	63.8	16.0
			56.6-71.0	1.71-2.27
			0-7.89	----
	2	229	61.5	1.25
			54.4-68.6	0.92-1.68
			0.33-7.57	24.4-1248

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* to *Relatively-Strong* Disparity

Cancer Site	Strata	MinD	ESS	D
Acute Monocytic Leukemia	4	69	60.9	2.57
			53.4-68.3	1.86-3.49
			0.33-8.22	44.7-1208
	2	218	57.2	1.50
			49.9-64.4	1.11-2.01
			0-7.89	----
Appendix	6	25	42.1	8.25
			33.5-50.5	5.88-11.9
			0-7.89	----
	4	42	39.8	6.05
			31.4-48.2	4.30-8.74
			0.33-7.57	48.8-1028
	2	191	33.9	3.90
			25.7-42.0	2.76-5.78
			0.33-7.57	24.4-604
Cervix Uteri	3	84	40.8	4.35
			28.5-52.5	2.71-7.53
			0-10.5	----
	2	90	39.5	3.06
			28.5-50.6	1.95-5.02
			0-10.5	----
Hodgkin-Extranodal	2	251	47.7	2.19
			39.4-55.8	1.58-3.08
			0.33-7.57	24.4-604
Lip	2	289	50.3	1.98
			42.3-58.4	1.42-2.73
			0.33-8.22	22.3-604
Testes	2	118	46.1	2.34
			34.2-57.2	1.50-3.85
			0-10.5	----

See Note to Table 1.

Table 3: *Moderate* Disparity

Cancer Site	Strata	MinD	ESS	D
Peritoneum, Omentum and Mesentery	2	304	37.5	3.33
			28.8-46.3	2.32-4.94
			0-7.89	----

See Note to Table 1.

2155-0182/10/$3.00

Tables 1-3 indicate the strongest ESS point estimate obtained in this study was 75.3 by the 5-strata melanoma model that separated Caucasian *vs.* African American patients into higher, intermediate, or lower-risk ranges.[1] The Tables also show that the smallest D statistic of 0.98 occurred for the 2-strata melanoma model. The result for D is an intriguing, suggesting that identifying a single accurate attribute for the left-hand side of this model, and one for the right-hand side, will yield a parsimonious CTA model achieving close to perfect accuracy in evaluating presence or absence of melanoma in Caucasian *vs.* African American people.

Future research involving a cancer site identified herein should adjust for disparity by using matching[22-24] during subject recruitment or else by using propensity score weighting.[25-27] Findings reported presently should be replicated if more recent data which relate race and cancer-incidence exist.

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.

[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.

2155-0182/10/$3.00

[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

[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 used publically available data and was exempt from Institutional Review Board review. No conflict of interest was reported.