Published in final edited form as:

J Acquir Immune Defic Syndr. 2017 March 01; 74(3): e60–e63. doi:10.1097/QAI.0000000000001221.

Assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines

Bryan E. Shepherd, PhD¹ and Peter F. Rebeiro, PhD² Caribbean, Central and South America network for HIV epidemiology (CCASAnet)

¹Department of Biostatistics, Vanderbilt University School of Medicine

²Department of Medicine, Vanderbilt University School of Medicine

Abstract

The HIV epidemiology literature is replete with studies that categorize continuous predictors, such as age and CD4+ cell count. It is well known that such an approach is sub-optimal, but it persists in part because results are easy to interpret after categorization. Splines may be used to incorporate continuous predictors with smoothed curves into regression models without categorization or linearity assumptions. Properly presenting and interpreting results from analyses with splines is critical for their widespread use. With data from 13,706 antiretroviral initiators in Latin America, we demonstrate how to interpret results from a Cox regression model using restricted cubic splines.

When performing a regression analysis, should we categorize continuous predictors? For studies of HIV/AIDS these may include variables such as age, CD4+ cell count (CD4), body mass index, and date of ART initiation. The statistical literature has a clear answer: No. Numerous papers have been written describing the problems with categorizing continuous variables [1–11].

Throughout, we will illustrate using data from 13,706 ART initiators with complete covariate information in Latin America and the Caribbean who are part of the Caribbean, Central and South America network for HIV epidemiology (CCASAnet) [12–13]. Figure 1A shows the predicted 5-year survival probability for adults after ART initiation as a function of age at ART initiation, with age included in the statistical model using categories (dashed lines; 18–29, 30–39, 40–49, 50 years), as a continuous variable assuming linearity (gray curve), and as a continuous variable using splines (dark curve; described in more detail later). These predicted probabilities are derived from a multivariable Cox regression model stratified by study site and adjusted for other variables (sex, clinical stage [AIDS, not AIDS], year of ART initiation [kept continuous and included with splines], CD4 [also kept

Corresponding Author: Bryan Shepherd, 2525 West End, Suite 11000, Nashville, TN 37203, Fax: 615-343-4924, Phone: 615-343-3496, bryan.shepherd@vanderbilt.edu.

This work was presented in part at the 2015 International Workshop for HIV Observational Databases; Catania, Sicily; March 27, 2015.

Conflicts of Interest: None.

continuous and included with splines], and initial ART regimen [non-nucleoside reverse transcriptase inhibitor {NNRTI}, boosted protease inhibitor {PI}, or other]).

Figure 1A illustrates some of the problems with categorization. Although only about 10% of CCASAnet patients were ≥ 50 years at ART initiation, there is considerable variation in the predicted survival probability for patients at ages ≥ 50 . This makes sense: patients initiating ART at the age of 50 years would be expected to have very different survival probabilities than patients starting at the age of 75 years. In fact, if we re-did the analysis only among those ≥ 50 years and left age in as a continuous variable assuming linearity, age is strongly associated with mortality, $p < 0.001$. The statistical model that categorizes ages ignores these differences, thereby losing important information. Similarly, if one wants to examine the association between another variable and mortality after adjusting for age, the model that categorizes age fails to completely adjust for age and therefore may result in residual confounding. Finally, the choice of categorization levels is arbitrary. Do we rush our patients to start ART right before their 50th birthday? Of course not. Everyone should understand that around 50 years, the hazard is likely somewhere between the estimated hazards of the 40–49 and ≥ 50 years age categories. However, if this is the general understanding, then why do we not fit models that explicitly permit these gradations? When continuous variables are categorized, our statistical models implicitly assume the risk of mortality drops/changes at the boundary of the categories.

Given these and other well-known problems with categorizing continuous predictors, one might ask why our HIV and epidemiology literature is replete with categorizations. We have occasionally heard researchers say that they categorize continuous variables because they do not want to assume linearity; such an assumption is not necessary with splines. Some researchers may lack technical know-how – it is easy to include categorical variables in analyses and splines may be considered statistically advanced and/or difficult to implement with standard statistical software. This need not be a hindrance – most major statistical software packages including R, Stata, and SAS have functions that allow the inclusion of splines in almost any regression model (e.g., linear, logistic, Cox, Poisson, negative binomial, Weibull, proportional odds, and median regression, to name a few). Another reason for categorization is that clinicians often have to make dichotomous decisions (treat or do not treat), and categorical results are easier to incorporate into this framework. However, we believe the primary reason our literature contains extensive categorization is that we simply like the interpretation of variables that have been categorized and that we struggle with interpreting alternative approaches. This need not be the case, and the bulk of this report will focus on the interpretation of analyses with splines.

A spline is a smoothed curve included in a regression model. There are many types of splines and many options for their fitting [14–17]. We will focus on restricted cubic splines, which are also known as natural splines [18]. The amount of smoothness in a model with restricted cubic splines is governed by the number of degrees of freedom, or parameters, used by the spline for a specific variable in the model. Increasing the degrees of freedom allows the model to capture more complicated trends but may also lead to the fitting of spurious associations, particularly with small sample sizes. In general, as with any regression model, the larger the sample size, the more degrees of freedom one can include in

the model. A popular rule of thumb suggests that there should be at least 10–20 events per degree of freedom [19].

Figure 1B shows the predicted 5-year probability of death after ART initiation as a function of age using splines and varying the degrees of freedom. One degree of freedom indicates that age was included in the model assuming linearity. Two degrees of freedom fits two parameters for age, and so forth. In this example where the sample size is large and there are many events ($n=1259$), results are essentially the same regardless of whether one chooses 2, 3, or 7 degrees of freedom, although with 7 degrees of freedom there appears to be some over-fitting to artifacts in this particular dataset (i.e., we probably do not really believe the extra wiggle in the curve).

The mathematics behind restricted cubic splines are not overly complicated, and are summarized in the Supplemental Material. In short, the smoothed curves are fit by selecting “knots” or points where curves come together. One may think of fitting lines that best fit the data between these knots. Smoothness around the knots is introduced with a cubic polynomial although the curves are “restricted” or forced to be linear at the tails of the predictor variables to avoid unstable estimates. Smoothed curves could also be fit using polynomials (e.g., age, age², age³, etc.), and restricted cubic splines are similar, although they tend to be more stable than polynomials in the extremes of the predictor variable (e.g., very young and very old ages) because they are restricted to be linear at the extremes [20]. There are other flexible and stable options for relaxing linearity, including fractional polynomials [21–22], but we will limit our discussion to splines in this manuscript. Fitting splines requires selecting the location of the knots, but in most cases it is sufficient to use default locations in the statistical software, which are often based on percentiles of the predictor variable. Selecting the number of knots may be done many ways including using cross-validation, AIC, likelihood ratio tests, and visual inspection. In general, we prefer to *a priori* specify the number of degrees of freedom based on the number of events/sample size, the perceived relative importance of variables, and the likelihood of non-linear associations. Variables that are thought to be more influential on the outcome or more likely to have non-linear associations are assigned more degrees of freedom. One can then visually inspect the robustness of the model fit to this choice. Two to four degrees of freedom (corresponding to three to five knots) is generally sufficient [23]. A reader who is interested in more details is referred elsewhere [14, 19].

Interpreting results using splines is paramount for their practical use. A regression fit will result in estimated coefficients for each parameter used in the splines. Other than including them in technical appendices, in almost all cases, one does not present these estimated coefficients – their interpretation is essentially meaningless. Instead, one should use linear combinations of these coefficients to obtain predicted values and interpret results using these predicted values. This is typically straightforward in standard statistical software.

The best way to interpret results from splines is to use a figure. For example, consider the Supplementary Figure. This uses the same data and analysis as before and it shows the predicted 5-year survival for each predictor in the model holding all other variables constant at representative levels. Estimates and pointwise 95% confidence intervals are shown. The

relationships between continuous variables and mortality are clearly demonstrated. This plot shows predicted 5-year survival, but it is also simple to plot log-hazards, hazards, or hazard ratios.

There are some settings, such as in abstracts, text, or tables, where specific contrasts (e.g., hazard ratios) are desired. Such contrasts are easy to construct using splines and interpretation is straightforward. Consider Figure 1C, which shows the hazard of death as a function of age holding all other variables constant at representative levels. Again, these are results from the same multivariable models with age included using restricted cubic splines with 4 knots (solid curve) and age included with categorization (dashed lines). To compute hazard ratios after categorization, one selects a reference category (e.g., 18–29 years) and compares the hazard of mortality for other categories (e.g., 50 years) with the hazard of mortality in the reference category. In this case the hazard ratio is $1.21/0.69=1.76$. After fitting a model including splines, one may compute hazard ratios in a similar manner by comparing specific ages with a single reference age. For example, one could choose 30 years as the reference age and compute hazard ratios by comparing the hazard of death at select ages with the hazard at 30. The hazard ratio for 50 versus 30 years, for example, is $0.99/0.68=1.46$. Any age may be compared to any other age without model re-fitting.

Tables with hazard ratios and confidence intervals using splines may be easily constructed. Table 1 is an example. Hazard ratios and 95% confidence intervals are given for all predictor variables at select predictor values. The right column contains p-values. These p-values are from likelihood ratio tests with the same number of degrees of freedom as the splines. They may be thought of as p-values corresponding to a test that the variable contains predictive information. In this case, age, CD4, and year of ART initiation are each highly predictive of mortality in the multivariable model. Notice that these p-values do not correspond to specific hazard ratios. One could construct p-values testing whether a specific hazard ratio equals 1 (e.g., testing the null hypothesis that the hazard of death is higher for a person starting ART at 50 years than someone starting ART at 30 years). These p-values would correspond with the 95% confidence intervals that are reported and are therefore typically not needed or of interest. Finally, one may also formally test for non-linearity. This is done by contrasting the model fit using splines with a model fit assuming linearity for a specific variable (i.e., 1 degree of freedom) using a likelihood ratio test. Such a test may be interesting – age, CD4, and year of ART initiation all contain evidence that their relationships with death are non-linear ($p<0.02$ for each) – but lack of evidence of non-linearity is not necessarily a reason to simply fit a model assuming a linear relationship [19].

Conclusion

We have provided a brief introduction to splines and their interpretation. Splines provide a scientific and preferable alternative to the categorization of continuous variables. We therefore hope to see more of their use in studies of HIV/AIDS.

All analyses were performed using R statistical software, with heavy use of the `rms()` package. Complete code for all results, as well as a simulated dataset on which one may

implement the code, is posted at <http://biostat.mc.vanderbilt.edu/ArchivedAnalyses>. Similar Stata code is also posted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of Funding: This work was supported in part by the United States National Institutes of Health, grant numbers U01 AI069923, R01 AI093234, and P30 AI110527.

References

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006; 25:127–41. [PubMed: 16217841]
2. Taylor J, Yu M. Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*. 2002; 83:248–63.
3. Vickers AJ, Lilja H. Cutpoints in clinical chemistry: time for fundamental reassessment. *Clin Chem*. 2009; 55:15–17. [PubMed: 19028819]
4. Streiner DL. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *Can J Psychiatry*. 2002; 47:262–66. [PubMed: 11987478]
5. van Walraven C, Hart RG. Leave 'em alone: why continuous variables should be analyzed as such. *Neuroepidemiology*. 2008; 30:138–39. [PubMed: 18421216]
6. Heavner KK, Phillips CV, Burstyn I, et al. Dichotomization: 2×2 ($\times 2 \times 2 \times 2 \dots$) categories: infinite possibilities. *BMC Med Res Methodol*. 2010; 10:59. [PubMed: 20573189]
7. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharm Stat*. 2009; 8:50–61. [PubMed: 18389492]
8. Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *Am J Epidemiol*. 1992; 136:464–74. [PubMed: 1415166]
9. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol*. 2006; 59:798–801. [PubMed: 16828672]
10. Chen H, Cohen P, Chen S. Biased odds ratios from dichotomization of age. *Stat Med*. 2007; 26:3487–97. [PubMed: 17066378]
11. Howe CJ, Cole SR, Westreich DJ, et al. Splines for trend analysis and continuous confounder control. *Epidemiology*. 2011; 22:874–5. [PubMed: 21968779]
12. McGowan CC, Cahn P, Gotuzzo E, Padgett D, Pape JW, Wolff M, et al. Cohort Profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme. *Int J Epidemiol*. 2007; 36:969–76. [PubMed: 17846055]
13. Carriquiry G, Fink V, Koethe JR, et al. Mortality and loss to follow-up among HIV-infected persons on long-term antiretroviral therapy in Latin America and the Caribbean. *Journal of the International AIDS Society*. 2015; 18:20016. [PubMed: 26165322]
14. de Boor, C. *A Practical Guide to Splines*. Springer; 2001.
15. Friedman JH. Multivariate adaptive regression splines. *The Annals of Statistics*. 1991; 19:1–141.
16. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*. 1992; 87:942–951.
17. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*. 1990; 85:941–949.
18. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine*. 1989; 8:551–561. [PubMed: 2657958]

19. Harrell, FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; 2001.
20. Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute*. 1988; 80:1198–1202. [PubMed: 3047407]
21. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*. 1994; 43:429–467.
22. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*. 1999; 28:964–974. [PubMed: 10597998]
23. Stone CJ. Comment: Generalized additive models. *Statistical Science*. 1986; 1:312–314.

Figure 1.

A. Association between age and estimated 5-year survival probability after starting ART including age using categories (dashed lines), as a continuous variable assuming linearity (gray curve), and as a continuous variable using restricted cubic splines with 3 degrees of freedom (black curve). The model is adjusted for the following variables (set at representative levels): sex (male), age (40 years), clinical stage at ART initiation (not AIDS), CD4 at ART initiation (200 cells/ μ L), year of ART initiation (2007), initial ART regimen (NNRTI), and site (Peru). **B.** Association between age and estimated 5-year survival probability after starting ART fit using restricted cubic splines with 1, 2, 3, and 7 degrees of

freedom (df). **C. Computing hazard ratios.** The model shows the hazard of death times a constant as a function of age at ART initiation categorizing age (dashed lines) and keeping age as a continuous variable using restricted cubic splines with 3 degrees of freedom (solid curve). The constant is the inverse of the hazard of death at the mean linear predictor, which is the default value for R's Predict.cph function in the rms package; other constants would result in identically shaped curves and the exact same hazard ratios (see Supplementary Material). The hazard is plotted on a logarithmic scale.

Table 1

Association between predictors and the hazard of death after ART initiation. *

	Adjusted Hazard Ratio (95% Confidence Interval)	p-value
Male sex	1.09 (0.96–1.22)	0.18
Age at start of ART (years)		<0.001
20	1.01 (0.77–1.32)	
30 (ref)	1	
40	1.11 (0.99–1.25)	
50	1.46 (1.25–1.70)	
60	2.06 (1.75–2.44)	
AIDS at start of ART	1.70 (1.50–1.93)	<0.001
CD4 at start of ART (cells/μl)		<0.001
50	1.98 (1.64–2.40)	
100	1.50 (1.25–1.82)	
200	1.08 (0.91–1.27)	
350 (ref)	1	
Year of starting ART		<0.001
2000	1.04 (0.75, 1.45)	
2002	1.07 (0.88, 1.30)	
2004	1.08 (1.01, 1.16)	
2006 (ref)	1	
2008	0.78 (0.70, 0.86)	
2010	0.60 (0.51, 0.71)	
Initial Regimen		0.37
NNRTI (ref)	1	
Boosted PI	1.17 (0.94, 1.45)	0.16
Other	1.07 (0.78, 1.46)	0.67

* A similar table, except using multiply imputed data, was published previously [13].

ART: antiretroviral therapy; AIDS: acquired immunodeficiency syndrome; NNRTI: non-nucleoside reverse transcriptase inhibitor; PI: protease inhibitor