Disentangling the impact of seroconversion age and set-point viral load on ART-free HIV survival

Amelia Bertozzi-Villa

A thesis submitted in partial fulfillment of the
requirements for the degree of

Masters of Public Health

University of Washington

2016

Committee:
Abraham D. Flaxman
Anna Bershteyn
Laura Dwyer-Lindgren
Michael Freeman

Program Authorized to Offer Degree:
Global Health

University of Washington

Abstract

Disentangling the impact of seroconversion age and set-point viral load on ART-free HIV survival.

Amelia Bertozzi-Villa

Chair of the Supervisory Committee:
Abraham D Flaxman, Ph.D
Global Health

Introduction: Prediction of off-treatment HIV survival is important in understanding risk among individuals unaware of their HIV status or unable to access care. Historically HIV survival analyses use age at seroconversion or set point viral load (SPVL) as their predictor of interest, but the relationships and interactions between these two covariates has yet to be rigorously determined. We analyzed (1) the impact of different SPVL estimation methods on survival prediction, (2) the relative effects of age at seroconversion and SPVL on survival, and (3) the effect of interaction terms between the two. All models were run on multiple subsets of the same dataset to test for sensitivity to time period, sample size, debiasing methods, and imputation types.

Methods: We utilized the CASCADE seroconverters dataset, composed of 16,964 eligible participants. We tested two specifications of SPVL: a geometric mean and a nonlinear modeling method. Our central model was a log-linear regression with time to death (from seroconversion) as the dependent variable and age at seroconversion, SPVL, and an AIDS censorship indicator as the independent variables. We tested five variations on this specification, including a null model, age-only, SPVL-only, a two-way age-SPVL interaction, and a three-way age-SPVL-AIDS indicator interaction. Each of these model specifications was tested on 16 different modifications of the CASCADE dataset: pre-1996 or full-timeseries, debiased or nondebiased, imputed or nonimputed, and testing 18, 20, and 22-year imputation upper bound for the imputed datasets. All models were validated and ranked using 10x10-fold cross-validation with root mean squared error (RMSE) as the error metric.

Results: Of the 160 models considered, average RMSE was 3.56 years (range 3.16, 3.98). The nonlinear SPVL method produced estimates with an RMSE 0.29 (0.10, 0.35) years lower than the geometric SPVL method, on average. Models without SPVL performed barely better than the null models. The best-performing model was fit using the nonimputed, nondebiased, pre-1996 dataset, and predicted a 27.0% (95% CI 12.9, 38.8) decrease in survival per tenfold increase in set-point viral load. It did not include a covariate for age. Such a strong impact of SPVL on survival time could have serious implications on mortality for at-risk groups, especially since population-level SPVL has been increasing since the early 1980s. Results were sensitive to the time period modeled.

Conclusion: Our analysis showed that SPVL was more predictive of survival than age at seroconversion, but that the way SPVL is calculated has a large impact on predictive performance. We did not find significant effect of the interaction between age at seroconversion and SPVL. Our work highlights the importance of targeting and treating at-risk populations quickly to avoid adverse effects from globally increasing set point viral loads.

1 Introduction

The sweeping rollouts of antiretroviral therapy (ART) and HIV prevention methods have saved an estimated 19.1 million life-years since the start of the epidemic [1]. However, access to treatment is still limited in many areas [2], and while the WHO's recommendation to universally test and treat people living with HIV would expand access to care for millions, it also adds stress to already-strained health infrastructures. For this reason, the question of what drives off-treatment survival remains a salient one.

Much emphasis has been put on age at seroconversion as a predictor of HIV survival time [3–11]. Generally, these studies found that adults who seroconvert at older ages face worse outcomes and shorter survival times than those who seroconvert when younger. A less commonly used predictor of survival is set point viral load (SPVL) [12–18]. The concept of SPVL comes from natural history models of viral load HIV infection [19], which describe a period of swift increase in viral load immediately after seroconversion (acute phase) followed by a steep decline as the adaptive immune response attacks the virus. Next come years of relatively low viral load (asymptomatic phase), until viral levels increase to the point of a clinical AIDS infection. Some summary measure of viral load during the asymptomatic phase is commonly referred to as the "set point". SPVL has been shown to vary widely between individuals due to both transmitted strain and host immune response [18], and higher levels of SPVL have correlated with survival times in the studies cited above.

While age at seroconversion and viral load have often been used separately as predictors in HIV survival models, few studies [12, 16] have tested them both together, and none explore in-depth the effect of interacting the two variables, or test the sensitivity of SPVL specification in the context of an age-SPVL off-treatment survival model. In this analysis, we address three questions using the collection of seroconverter studies known as the CASCADE dataset: (1) How choice of SPVL metric affects survival estimates, (2) the relative effects of SPVL and age at seroconversion on survival prediction, and (3) the effect of interacting the two covariates. To test the sensitivity of model outputs to sample size, time period, and other factors, we ran the analysis on multiple transformations of the main dataset.

2 Methods

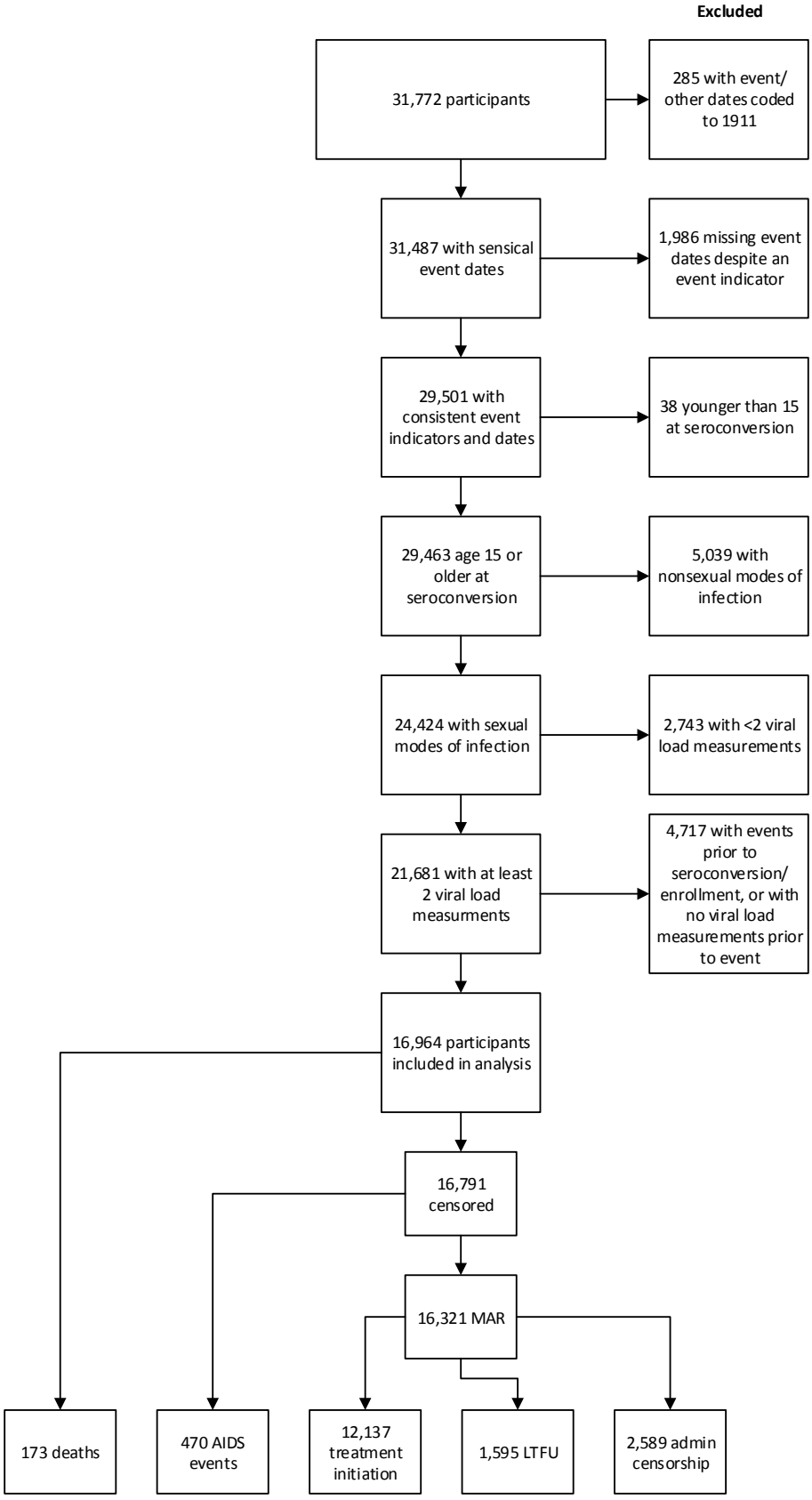
2.1 Data and Data Transformations

Concerted Action on SeroConversion to AIDS and Death in Europe (CASCADE) is a collaboration of 28 cohort studies whose participants have well-estimated dates of HIV seroconversion. We used data from 26 of these cohorts, two from sub-Saharan Africa and the remainder from Europe, Canada, and Australia. The two excluded cohorts were the Greek Haemophilia Cohort and the Amsterdam Cohort Study Among Drug Users, since we retain only those infected by sexual means. Note that there are other injection drug user cohorts included in the dataset, for which we kept only individuals recorded as infected by sexual transmission. Data were extracted on November 26, 2014. All collaborating cohorts received approval from their regulatory or national ethics review boards.

The full dataset includes 31,772 individuals. After excluding participants who seroconverted when younger than 15, those infected by nonsexual means, those with fewer than two viral load measurements, and those with nonsensical event recording (e.g. seroconversion dates in 1911), the final sample included 16,964 participants (Figure 1). The focus of this study is time to death for those who never experienced treatment, but we witness only 173 such deaths–the remainder were censored by treatment initiation, loss to follow-up (LTFU), or study termination/other administrative censorship.

Most of these censorship events can be considered missing at random (MAR), meaning that the mechanism by which individuals are censored does not depend on the outcome of interest (here, time to death). For example, the termination date of a study does not generally depend on the outcomes of individuals in that study. A clear occasion when this assumption does not hold true for the data at hand is when an individual was diagnosed with AIDS prior to censorship. Because AIDS diagnosis indicates that one is extremely ill, patients may initiate treatment after diagnosis as a life-saving or life-extending measure, and as such any censorship event following such a diagnosis must be considered missing not at random (MNAR). The term "AIDS event" is used to refer to

Figure 1: Process of exclusion from survival analysis. LTFU: Lost to follow-up.



such individuals throughout this paper. We do not report estimates for these AIDS censorship events in any results or plots except for regression tables.

Thus, we split our final dataset into three categories (Figure 1): 173 death events, which include any never-treated death captured in the study (with our without a prior AIDS diagnosis); 470 AIDS events, described above and censored at time of AIDS diagnosis; and 16,321 MAR events censored at treatment initiation (12,137), LTFU (1,595), or administrative censorship (2,589), whichever came earliest. Under the assumptions of missingness-at-random, this latter group can be excluded from analysis without generating bias, but at the expense of statistical power. The alternative is to use an imputation algorithm to estimate time-to-death for the MAR group. We test both options, running all analyses on both a small non-imputed group (only death and AIDS events) and a large, fully-imputed dataset. We used the 𝔸MELIA II software package in R [20] to impute event time (for more details see Appendix). Additionally, we tested the following data transformations:

- Pre-1996: Beginning in 1996, the triple-drug regimen known as highly-active antiretroviral therapy (HAART) became widely available. This marked a fundamental shift in access and efficacy of HIV medication, as well as in the way people living with HIV considered their care. We reran all analyses on a subset of the two datasets described above consisting only of individuals whose pre-imputation events were prior to 1996. Note that this could lead to imputed death dates after 1996, but these imputations would be informed using only data from the pre-1996 era.

- Debiasing: In most cases, participants enroll in CASCADE cohorts after seroconversion, and seroconversion date is established using estimation methods well-documented elsewhere [6]. However, this method of study inclusion generates survivorship bias: those who enroll in the study must already have survived long enough to be in the cohort, and as such may be predisposed to a longer life. We tested two versions each of the four datasets described above: one with seroconversion dates as determined by CASCADE, and one with enrollment date substituted for seroconversion date if individuals seroconverted prior to enrollment.

- Upper bound for imputation: If imputation predictions are left unbounded, 𝔸MELIA II

will occasionally predict either a too-short event time (before the censored individual's last appointment) or an infeasibly long event time. The lower bound for each individual is always his/her last logged encounter with the health system, but we tested the effect of 18, 20, and 22 years as plausible upper bounds for off-treatment survival times. (The longest off-treatment time-to-death in our dataset is 17.9 years, and the longest time to an AIDS event is 18.6 years. The longest time to censorship is 28.0 years, but such a long event time was extremely rare– only 5 (0.03%) of the 16,321 censored individuals survived more than 25 years. Instead, we picked an upper bound of the mean survival of those surviving more than 20 years: 22.1 years.) These three upper bounds were tested on each of the four imputed datasets described above (nondebiased full timeseries, nondebiased pre-1996, debiased full time series, and debiased pre-1996).

Our final analysis thus included 16 datasets: four non-imputed, and twelve imputed. The following models specifications were tested on each of them.

2.2 Model Specification

At its core, our model is a log-linear regression with time-to-event (from seroconversion) as the dependent variable and age at seroconversion and SPVL as the independent variables. Since the AIDS event censorships described in the previous section are not MAR and thus cannot be imputed, they are accounted for directly by an indicator variable within the model:

$$log(time\ to\ event) = \beta_0 + \beta_1 age + \beta_2 SPVL + \beta_3 I(AIDS) + \epsilon \tag{1}$$

We tested five variations on this central model (Table 1): a null model (only intercepts for death and AIDS events), age only, SPVL only, a two-way interaction between age at seroconversion and SPVL, and a three-way interaction between age, SPVL, and event type.

There is no standard for how SPVL should be defined. Many assume that viral load stays roughly

Model Term	Null	Age-Only	SPVL-Only	Central	Two Way	Three Way
Intercept	X	X	X	X	X	X
I(AIDS)	X	X	X	X	X	X
Age		X		X	X	X
SPVL			X	X	X	X
Age*SPVL					X	X
Age*I(AIDS)						X
SPVL*I(AIDS)						X
Age*SPVL*I(AIDS)						X

Table 1: Model Specifications. SPVL: Set point viral load

constant over much of the asymptomatic phase, and estimate SPVL as a geometric or arithmetic mean of values [17, 18]. Others approximate the same idea by just the first or second viral load sampled after seroconversion [15]. Yet others argue that the assumption of a near-constant viral load is erroneous, and that viral load is better modeled as a linear increase over the course of the asymptomatic phase [21].

We tested three methods of determining the SPVL metric: a simple geometric mean from 6 months after seroconversion until the event, a nonlinear regression method, and a hybrid of the two (see Appendix). The hybrid method produced results almost identical to the geometric mean method and was not included in further analysis. Thus, every model in Table 1 that included SPVL was tested twice, for a total of 10 model specifications.

2.3 Validation

To determine which of the 160 models described above (10 model specifications x 16 data transformations) is most predictive, we ran 10x10-fold cross-validation using root mean squared error (RMSE) as the error metric. That is, we randomly split the original dataset of 16,964 participants into 10 groups. We ran our entire analysis 10 times, holding out one group each time. We used the model outputs from each analysis to predict survival time for each individual in the held-out dataset with an observed AIDS or death event, and calculated the RMSE between the observed and predicted values. We repeated this entire process 10 times, with a different random split of observations each time, for a total of 100 test-train splits. The reported RMSE is the mean across

these 100 tests.

Note that the data for cross-validation was split into testing and training sets prior to imputation, and error calculated based only on observed death/AIDS events. This substantially reduces the number of testable events in each testing set, but ensures that the values in those sets remain unused by any part of the modeling process.

3 Results

3.1 Data

Of the 16,964 individuals included in our analysis, 14,133 were men and 2,830 were women. 12,335 men and 1 woman were infected through homosexual interactions, the remainder through heterosexual encounters. Mean age at seroconversion was 33.7 years (range: 15.5-79.3). Mean SPVL for the geometric mean method was 4.31 log10(units/mL of blood) (1.07, 7.71) and for the nonlinear model method was 4.31 (-0.48, 7.40). Debiasing (setting seroconversion time equal to enrollment time) was performed for 13,374 individuals who seroconverted before their study enrollment, with an average change of 0.89 years (<0.1, 24.0).

The pre-1996 subset consisted of 376 individuals. In this subset, 33 (8.8%) seroconverters had logged death events, 54 (14.4%) had AIDS events only, and 289 (76.9%) were censored. 83.7% of the 289 missing at random events were treatment initiation, compared to 74.4% in the sample as a whole. However, these treatment initiation events must be considered fundamentally different from those that came after 1996, as they would have consisted of one or two early-stage drugs, rather than the triple-drug therapy introduced in 1996.

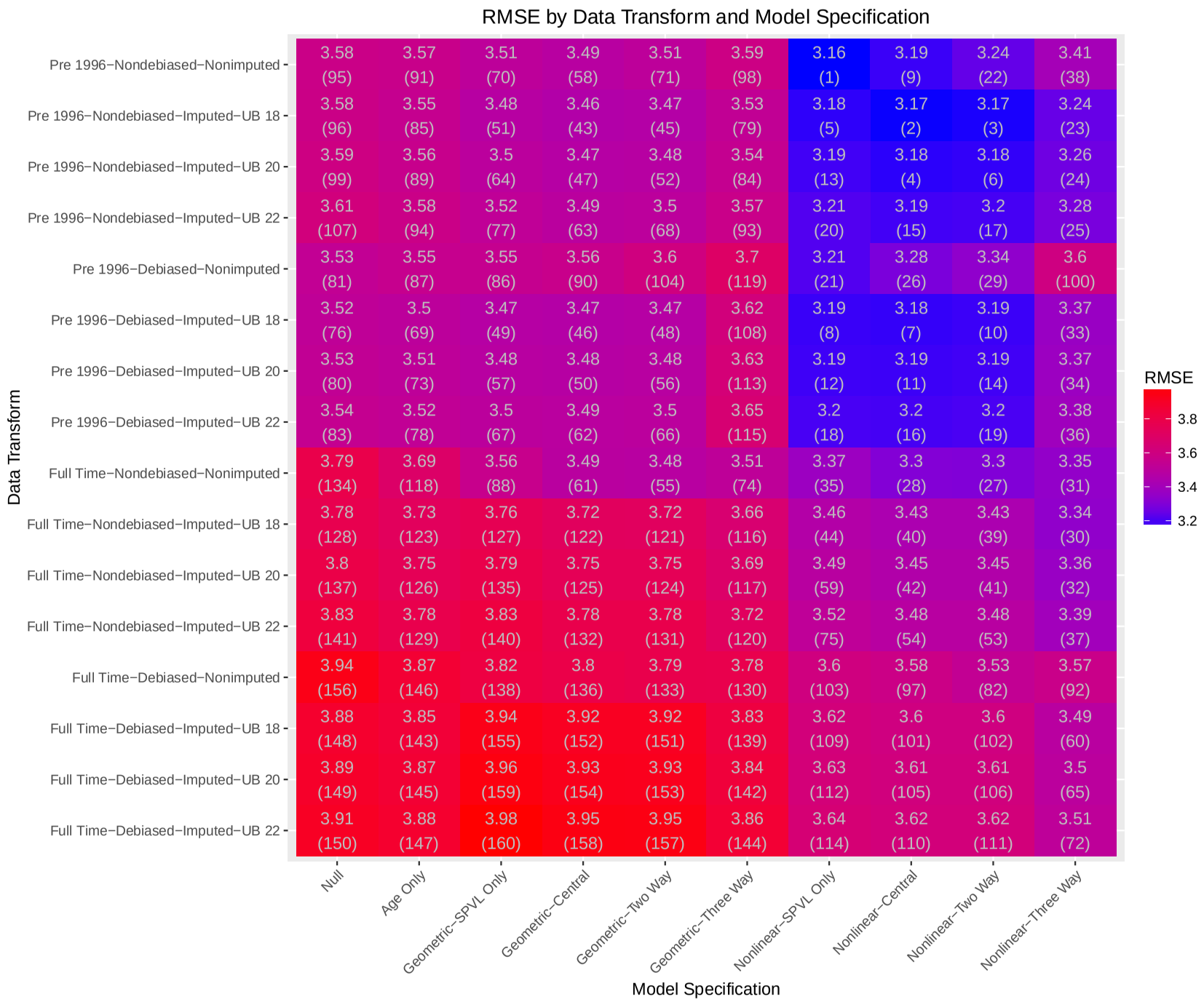
3.2 Models

A total of 160 models were considered, with an average RMSE of 3.56 years (range: 3.16, 3.98). There are clear patterns in the performance of different data-transformation-model specification combinations (Figure 2). The best-performing models utilize the nonlinear SPVL covariate, have no more than a two-way interaction, and utilize data subset prior to 1996. The worst-performing models utilize the geometric SPVL covariate and full-timeseries, imputed data with debiasing. Null or age-only models performed worst within each dataset, with age-only models granting, on average, a 0.03-year (-0.02, 0.1) decrease in RMSE from their respective null model. Results are highly sensitive to SPVL type: Nonlinear SPVL models have an average RMSE of 3.37 (3.16, 3.64), geometric SPVL models a mean RMSE of 3.65 (3.46, 3.98), and no-SPVL models (age-only and null) a mean RMSE of 3.69 (3.50, 3.94). Thus, moving from a no-SPVL model to a geometric model gives on average a 0.04-year decrease in RMSE, and moving to a nonlinear model reduces RMSE by an additional 0.29 (0.10, 0.35) years.

To explore these features in more depth, we take as a baseline the null, full-timeseries model without debiasing and without imputation, since this was the least-adjusted dataset with the simplest model specification. This model was ranked 134th, with an RMSE of 3.78. Twenty-six models performed worse than this baseline (Figure 2, bottom left). All were full-timeseries datasets, all but four had imputed values, all but four were debiased, and all either used the geometric method for estimating SPVL or did not use SPVL as a covariate.

Of the 10 worst-performing models, 9 utilized the full-timeseries, debiased, imputed datasets with geometric SPVL and the SPVL-only, central, or two-way model specification. (The tenth, ranked 156th, was a null model with a full-timeseries, debiased, nonimputed dataset.) These nine models had an average RMSE of 3.94 (range: 3.92, 3.98), a decrease of 0.16 years (0.14, 0.20) from the baseline model (Figure 2, bottom left). Holding everything constant except SPVL type in these models (i.e. moving to the nonlinear SPVL models) resulted in an average RMSE of 3.62 (3.60, 3.64), an improvement of 0.33 years (0.32, 0.34) from the geometric SPVL models and of 0.16 years (0.14, 0.18) from the baseline (Figure 2, bottom right). Holding everything constant except time

Figure 2: Heatmap of Root Mean Squared Errors (RMSEs) for every data transform-model specification combination. Ranking listed below RMSE. UB: Upper bound; SPVL: Set point viral load.



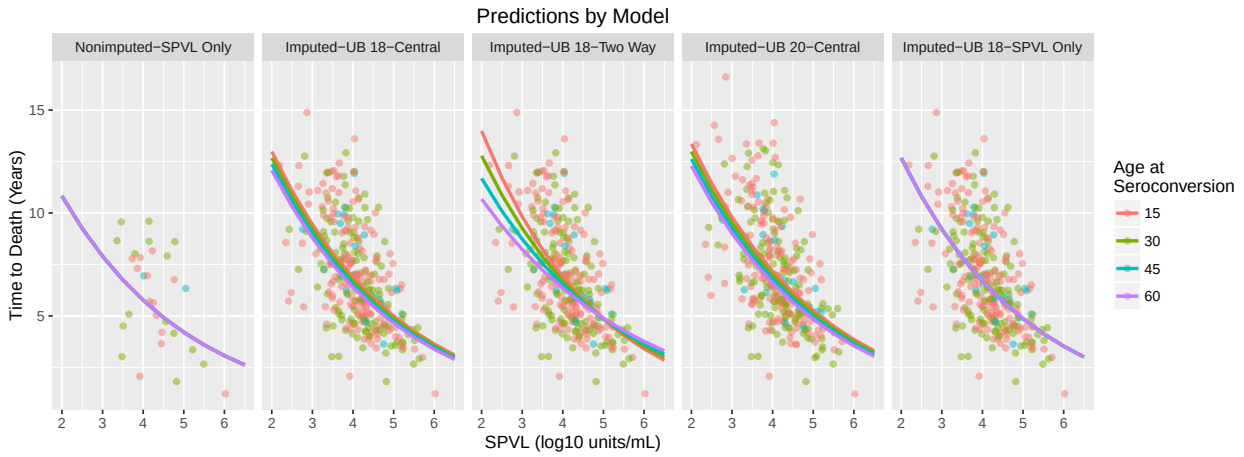
period for the worst-performing models (i.e. moving to the pre-1996 dataset) resulted in an average RMSE of 3.48 (3.47, 3.52), an improvement of 0.46 years (0.45, 0.48) from the full-timeseries models and of 0.30 years (0.26, 0.31) from the baseline model (Figure 2, upper left). Making both changes (using nonlinear SPVL and the pre-1996 dataset) resulted in an average RMSE of 3.19 (3.18, 3.20), an improvement of 0.75 years (0.73, 0.78) from the worst performers and of 0.59 (0.58, 0.60) from the baseline (Figure 2, upper right).

Nondebiased datasets consistently performed better than debiased, with an RMSE of 3.50 (range: 3.16, 3.83) in the former group and 3.60 (3.18, 3.98) in the latter. Only four nondebiased models performed worse than baseline, and none had RMSEs higher than 3.83. The remaining nondebiased models include the six top-ranked (RMSE 3.16-3.18, improvements of 0.60-0.62 years from baseline). Nonimputed models performed almost indistinguishably from imputed models overall, with an average RMSE of 3.53 (3.16, 3.94) for the former and of 3.55 (3.16, 3.98) for the latter.

Three-way interaction models performed differently depending on the dataset tested. In pre-1996 datasets, the three-way specification consistently performed worse than two-way, central, or SPVL-only specifications with on average a 0.13 (range: 0.06, 0.26) year increase in RMSE over the two-way model. In full-timeseries datasets (with one exception) the three-way models performed best, with an RMSE on average 0.08 years (0.04, 0.11) lower than the two-way models. The one exception was the nondebiased nonimputed model, which behaved similarly to the pre-1996 datasets.

The top-performing model utilized the SPVL-only model specification and the pre-1996, nonimputed, nondebiased dataset. It predicted a 27.0% (95% CI: 12.9, 38.8) decrease in mortality per tenfold increase in SPVL. The top ten models all utilizes the pre-1996 dataset and predicted an effect on SPVL similar to the best model (for two-way interaction models, the effect of SPVL was calculated for age 33.7, the mean age in the dataset). Seven of the top ten models included a term for age at seroconversion, ranging in effect size from a 10.5% (-29.1, 47.7) decrease in survival time for every decade of age at seroconversion to a 1.7% (-36.1, 51.9) increase in survival time. In none of these models was age at seroconversion significant. See Figure 3 and Table 2 for results from the top five models. Full model outputs are available upon request.

Figure 3: Predictions from top five models. All datasets are pre-1996, nondebiased. UB: Upper bound; SPVL: Set point viral load.



4 Discussion

This analysis strove to find the most effective framework possible for predicting off-treatment HIV survival by answering three questions: (1) is there a best way to estimate SPVL? (2) what are the relative effects of age at seroconversion and SPVL on survival? (3) does the interaction between the two matter? We ran analyses on a variety of datasets to test the sensitivity of results to the data used.

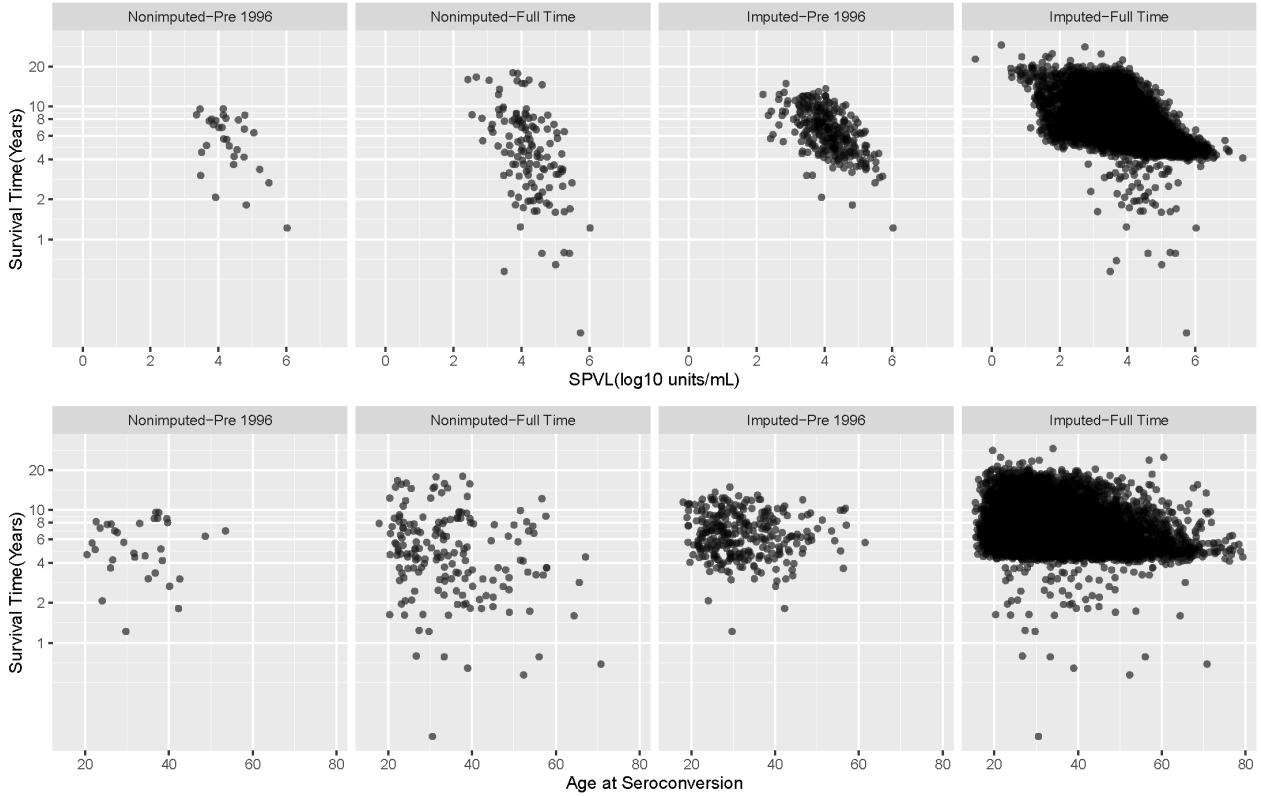
The answer to the first question is an unequivocal yes. Modeling the asymptomatic phase of infection as a linearly increasing function of viral load, rather than assuming approximately constant viral load values over the course of the asymptomatic phase, leads to better predictions for every model tested (mean RMSE difference between the two models 0.29 years, range 0.10, 0.35). This finding supports the assessment by Vidal et al. [21] and Masel et al. [22] that most patients' viral load progression is not well described by random fluctuation around a single value. This is not to suggest that some marker of asymptomatic viral load is irrelevant to survival–both this analysis and those by Fraser et al. [18] show that such a viral load measurement is both highly variable across individuals and predictive of survival– but our results suggest that the estimation framework pioneered by Pantazis et al. [14] and replicated here is a more appropriate marker than the traditional aggregate measures. On modern computers, there is little added computational cost of one method over another.

Ranking	Data Transform	Model Term	SPVL Only	Central	Two Way
1	Nonimputed	Intercept	3.01		
			(2.24, 3.79)		
		I(AIDS)	-0.24		
			(-0.45, -0.04)		
		SPVL	-0.32		
			(-0.49, -0.14)		
2	Imputed-UB 18	Intercept		3.22	
				(2.83, 3.62)	
		I(AIDS)		-0.43	
				(-0.54, -0.31)	
		Age		0	
				(-0.01, 0)	
		SPVL		-0.32	
				(-0.41, -0.22)	
3	Imputed-UB 18	Intercept			3.5
					(2.12, 4.87)
		I(AIDS)			-0.43
					(-0.54, -0.31)
		Age			-0.01
					(-0.05, 0.03)
		SPVL			-0.38
					(-0.72, -0.05)
		Age*SPVL			0
					(-0.01, 0.01)
4	Imputed-UB 20	Intercept		3.24	
				(2.78, 3.7)	
		I(AIDS)		-0.47	
				(-0.58, -0.35)	
		Age		0	
				(-0.01, 0)	
		SPVL		-0.31	
				(-0.41, -0.21)	
5	Imputed-UB 18	Intercept	3.18		
			(2.78, 3.58)		
		I(AIDS)	-0.43		
			(-0.54, -0.32)		
		SPVL	-0.32		
			(-0.42, -0.22)		

Table 2: Top five regression outputs. All data transformations are pre-1996 and nondebiased, all SPVL methods are nonlinear. UB: Upper bound; SPVL: Set point viral load.

In regards to the second question, our results suggest that SPVL (as we define it) is a considerably stronger predictor of survival time than age at seroconversion. Across the datasets tested, the age-only models had RMSEs barely better than the null models, whereas SPVL-only models performed

Figure 4: Age at seroconversion and SPVL vs survival time for each dataset. All datasets shown are nondebiased and have an upper bound of 18 years, when relevant. SPVL: Set point viral load.



on-par with the best performing models. Furthermore, the confidence intervals on the estimate of the coefficient on age crossed zero in 73 (65.1%) of the 112 models that included age.

This finding is consistent with our data (Figure 4): while a relationship between viral load and survival time is apparent even in the models without imputed data, there is a much less clear relationship between survival and age at seroconversion. Given that age at seroconversion is a well-established predictor of HIV mortality [3–11], this result is surprising. We posit several possible explanations:

1. The analyses cited above did not include SPVL, which acts as a confounder in the age-mortality relationship. Given the poor performance of the age-only models in our analysis, this does not seem to be a large factor. The correlation between age and SPVL in our full model was only 0.11 for the nonlinear SPVL and 0.08 for the geometric SPVL.

2. Our dataset had insufficiently varied ages to capture the effect of age on survival. Many of the

studies cited above were designed specifically to explore the effects of HIV infection at older ages, and as such sampled heavily from older seroconverters and modeled age as a binary variable indicating whether a patient was or wasn't older than 50. The individuals in our study were overwhelmingly in their 20s or 30s, with only 6.3% of the full sample and 3.5% of the pre-1996 sample over the age of 50. Perhaps age at seroconversion begins to matter dramatically more over a certain age threshold that our sample could not capture.

3. Confounding by treatment. Most of the studies cited above assessed outcomes for patients who were receiving treatment. It is plausible that treatment effects and age at seroconversion (as well as time form seroconversion to treatment, the presumed stabilization of viral load while on treatment, and other factors) could interact in such a way as to make age at seroconversion a stronger predictor of mortality than SPVL.

4. Otherwise incomparable datasets. The CASCADE data consist almost entirely of men who have sex with men infected in Europe, where the subtype of HIV known as clade B is dominant. It is possible that study results from sub-Saharan Africa (where the epidemiologically distinct clade C of HIV is dominant) or of populations infected by other routes may experience different relationships between age, SPVL, and off-treatment survival.

The answer to the third question appears to be no. Most interaction term models predicted results little different from those predicted by more parsimonious models for the ages most commonly found in the dataset (although predictions varied widely for the very young or very old, including estimates of both protective and detrimental effects of age at seroconversion on survival as SPVL increases). Furthermore, very few of these effects were significant: Only 7 (10.9%) of the 64 models with interaction terms had confidence intervals that did not cross zero. This is scarcely more than the amount we would expect to see by chance, since we ran multiple comparisons on the same dataset. As such, this analysis finds no compelling reason to include interaction terms between the two covariates in future models. However, it is interesting to note that our interaction models fairly consistenly predicted a (nonsignificant) protective effect of age as SPVL increases. We would have expected the opposite. The nature of the interaction between age at seroconversion and SPVL as it relates to survival is an open area for further research.

The sensitivity of model performance to the dataset used, especially the time series included, is striking (though not unexpected). Running the same model specifications on the pre-1996 datasets compared to the full-timeseries datasets produced on average a 0.26-year (range: 0.13, 0.34) decrease in RMSE, a larger change than that seen for any other data transformation. This finding highlights the importance of carefully considering and testing the data used in any regression analysis.

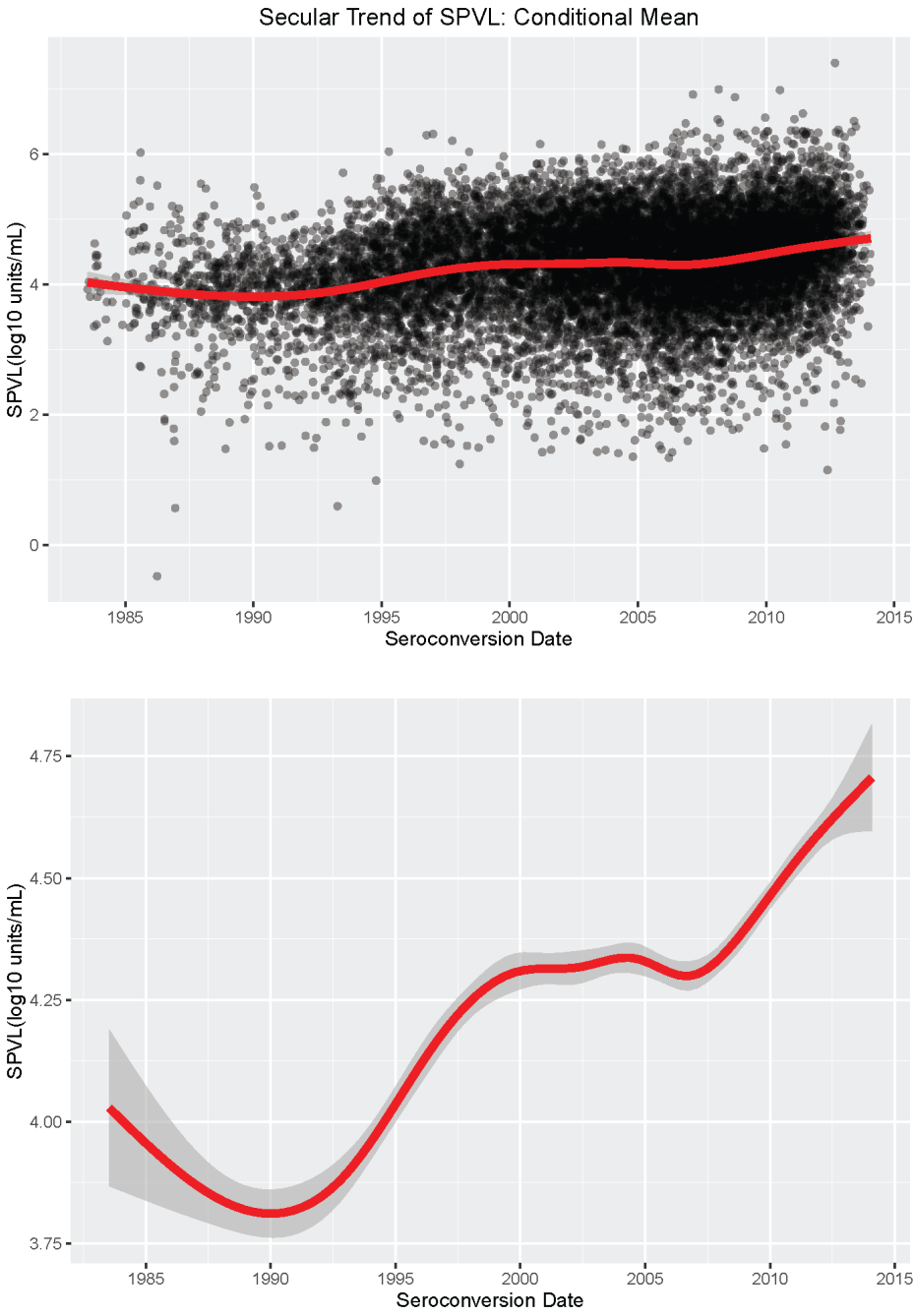
A strong correlation between SPVL and survival could have important policy implications. Pantazis noted in 2014 [23] that SPVL at the population level has been increasing since the beginning of the epidemic. We also find this pattern: an almost tenfold increase in population-level SPVL from 1990 to 2014 (Figure 5). Later work [24] shows that this phenomenon is likely due to selective pressure on the virus by drug regimens: while HIV naturally selects for strains of intermediate virulence [18, 25] to simultaneously optimize transmissibility and duration of infection, widespread use of antiretrovirals incentivizes selection of more virulent, shorter-lived HIV strains.

The WHO's recommendation of universal testing and treatment for persons living with HIV has two relevant implications for the coming years: many more people will go on treatment, and some highly vulnerable populations will prove extremely difficult to reach with widespread and effective medication. These populations may eventually experience higher set point viral load due to the aforementioned selection pressure, leaving them at an even higher risk of mortality than they previously faced. Our analysis highlights the need to proactively identify and target these groups for treatment to minimize their risk.

This work has several limitations. Although our full sample was almost 17,000 individuals strong, only a small subset of those individuals experienced the outcome of interest–death without ever going on treatment. From a public health perspective, this reflects the widespread efficacy of lifesaving HIV treatment programs, and we laud that achievement, but it did force a trade-off between severely limiting sample size and imputing a large fraction of our data. Models showed considerable sensitivity to the time period selected for analysis, but less sensitivity to other data transformations (debiasing, imputation, etc).

The assumption of missingness-at-random for treatment initiation is not universally agreed-upon.

Figure 5: Secular trend of SPVL in the population. Trend shown without data below for visibility. SPVL: Set point viral load.



While Pantazis et al. [14] and others make this assumption, some have argued that treatment initiation should be considered a missing not-at-random form of censorship.

Finally, the composition of our dataset prevented us from stratifying effectively on gender, and may have prevented us from finding a significant effect of age on survival. We look to other analyses on more diverse populations to address these questions.

5 Conclusion

Our best model predicted a 27.0% (95% CI 12.9, 38.8) decrease in survival per tenfold increase in set-point viral load, and did not include a coefficient for age at seroconversion. The most appropriate method of SPVL calculation assumes a linear increase in viral load over time, not a steady state. SPVL is a better predictor of survival than age at seroconversion, and the interaction between the two is negligible. The strong correlation between SPVL and survival, combined with a secular increase in population-level SPVL and the rollout of universal test and treat HIV care, may put vulnerable populations at higher risk of premature HIV mortality.

6 Acknowledgements

Many thanks to my committee members Abie, Anna, Laura, and Mike for thoughtful input and good conversation. Thanks also to my collaborator Christian with whom I hope to publish this work, to Becca for the solidarity, to Mom and Clara for collectively being my rock, to Tyler for endless patience and love, and to Mouse and Bug for important editorial decisions.

7 Appendix

7.1 Details of Imputation

𝔸MELIA II is a multiple-imputation software package that uses a bootstrapped expectation-maximization (EM) algorithm to "fill in" missing values in datasets based on other characteristics of those datasets [20]. Each run of the imputation will return a user-specified number of datasets, each with identical observed values and with varied imputed values representing the uncertainty in the imputed estimates. Further analyses are run on each imputed dataset separately, and results combined at the end. The EM algorithm estimates by borrowing strength from other covariates in the dataset, and users are encouraged to include in the imputation at least all variables included in any regressions they intend to run.

For each data transformation described in the main text, we imputed 10 datasets, including the following variables in addition to our variable of interest (time to death): patient identifier, time to event (death or censorship), age at seroconversion, geometric SPVL, and nonlinear SPVL. Some missing SPVL values were also imputed (see below). Bounds on imputation outputs were specified as detailed in the main text. Note that these imputed datasets are separate from the test-train splits used for cross-validation.

After obtaining model coefficients and standard errors from each imputed dataset, we followed the suggested procedures in Honaker [20] to combine model outputs into final results.

7.2 SPVL Determination

7.2.1 Nonlinear Determination of SPVL

To calculate the nonlinear SPVL covariate, we adopted a method from Pantazis et al. [14] in which viral load trajectory after seroconversion (in log10 space) is modeled as an exponential decline followed by a linear increase until AIDS. The minimum of this function is considered the set point.

The model is specified as follows:

$$f(t) = log10vl(t) = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)t + \beta_2 e^{-\beta_3 t} \tag{2}$$

where β_0 and β_1 are the fixed-effect intercept and slope for the linear part of the model, γ_0 and γ_1 are person-level random effects for the intercept and slope, β_3 describes the rate of exponential decline, and β_2 described the magnitude of that decline (both are fixed effects). One can then calculate an individual's time to set point (TTS) from the following equation:

$$tts(\beta_1, \beta_2, \beta_3, \gamma_1) = -\frac{1}{\beta_3} log\left(\frac{\beta_1 + \gamma_1}{\beta_2 \beta_3}\right) \tag{3}$$

and find that individual's SPVL by plugging this TTS back into the fitted model: $spvl = f(tts)$. Similarly, a population-level SPVL can be found by including only fixed effects from the nonlinear model into the TTS equation.

The nonlinear model had β_0 of 4.25 (SE 0.008), β_1 of 0.16 (0.002), β_2 of 1.15 (0.023), and β_3 of 16.0 (0.54).

7.2.2 Hybrid Method for SPVL Determination

The geometric method for SPVL determination picks 6 months after seroconversion as the beginning of the asymptomatic phase. To test how arbitrary this value was, we calculated the population-level TTS, and calculated the SPVL as the geometric mean from that time, rather than from 6 months after seroconversion.

The global TTS for the nonlinear model was 3.6 months. Taking a geometric mean from this time point produced results very similar to those of the geometric mean from 6 months after seroconversion, and as such this metric was excluded from further analysis.

7.2.3 Imputation of SPVL Values

Those patients without viral load measurements more than 6 months after seroconversion (full timeseries: 2,233, 13.2%; pre-1996: 26, 6.91%) had an indeterminate SPVL according to the geometric method. Those patients for which the nonlinear model predicted a decline in viral load (full timeseries: 3,103, 18.3%; pre-1996: 78, 20.7%) had an indeterminate SPVL according to the nonlinear method. In the imputation datasets, these individuals had an SPVL imputed along with time to death. In the non-imputation datasets, those with observed death or AIDS events and without SPVL measurements were excluded from the model (geometric method: full timeseries 46, 7.15%, pre-1996 1, 1.15%; nonlinear method: full timeseries 116, 18.4%, pre-1996 13, 114.9%).

References

[1] IHME. *Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 19902013: a systematic analysis for the Global Burden of Disease Study 2013 - The Lancet*. 2014. URL: http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)60844-8/abstract (visited on 06/07/2016).

[2] UNAIDS. *Access to antiretroviral therapy in Africa: Status report on progress toward the 2015 targets.* Status Report. 2013. URL: http://www.unaids.org/sites/default/files/media_asset/20131219_AccessARTAfricaStatusReportProgresstowards2015Targets_en_0.pdf.

[3] Ilan Asher et al. "Characteristics and Outcome of Patients Diagnosed With HIV at Older Age". In: *Medicine* 95.1 (Jan. 2016). ISSN: 0025-7974. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4706254/ (visited on 06/07/2016).

[4] Jim Todd et al. "Time from HIV seroconversion to death: a collaborative analysis of eight studies in six low and middle-income countries before highly active antiretroviral therapy". eng. In: *AIDS (London, England)* 21 Suppl 6 (Nov. 2007), S55–63. ISSN: 1473-5571.

[5] A. G Babiker et al. “Age as a determinant of survival in HIV infection”. In: *Journal of Clinical Epidemiology* 54.12, Supplement 1 (Dec. 2001), S16–S21. ISSN: 0895-4356. URL: http://www.sciencedirect.com/science/article/pii/S0895435601004565 (visited on 10/30/2015).

[6] “Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. Collaborative Group on AIDS Incubation and HIV Survival including the CASCADE EU Concerted Action. Concerted Action on SeroConversion to AIDS and Death in Europe”. eng. In: *Lancet (London, England)* 355.9210 (Apr. 2000), pp. 1131–1137. ISSN: 0140-6736.

[7] Daniel H. J. Davis et al. “Early diagnosis and treatment of HIV infection: magnitude of benefit on short-term mortality is greatest in older adults”. eng. In: *Age and Ageing* 42.4 (July 2013), pp. 520–526. ISSN: 1468-2834.

[8] N. Han et al. “HIV and aging: insights from the Asia Pacific HIV Observational Database (APHOD)”. eng. In: *HIV medicine* 16.3 (Mar. 2015), pp. 152–160. ISSN: 1468-1293.

[9] Amy C. Justice et al. “Predictive accuracy of the Veterans Aging Cohort Study index for mortality with HIV infection: a North American cross cohort analysis”. eng. In: *Journal of Acquired Immune Deficiency Syndromes (1999)* 62.2 (Feb. 2013), pp. 149–163. ISSN: 1944-7884.

[10] Rebecca A. Legarth et al. “Long-Term Mortality in HIV-Infected Individuals 50 Years or Older: A Nationwide, Population-Based Cohort Study”. eng. In: *Journal of Acquired Immune Deficiency Syndromes (1999)* 71.2 (Feb. 2016), pp. 213–218. ISSN: 1944-7884.

[11] Jalal Poorolajal et al. “Predictors of progression to AIDS and mortality post-HIV infection: a long-term retrospective cohort study”. eng. In: *AIDS care* 27.10 (2015), pp. 1205–1212. ISSN: 1360-0451.

[12] T. R. Sterling et al. “Initial plasma HIV-1 RNA levels and progression to AIDS in women and men”. eng. In: *The New England Journal of Medicine* 344.10 (Mar. 2001), pp. 720–725. ISSN: 0028-4793.

[13] F. de Wolf et al. "AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion". eng. In: *AIDS (London, England)* 11.15 (Dec. 1997), pp. 1799–1806. ISSN: 0269-9370.

[14] N. Pantazis et al. "Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs". en. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.2 (Apr. 2005), pp. 405–423. ISSN: 1467-9876. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9876.2005.00491.x/abstract (visited on 10/30/2015).

[15] Ludo Lavreys et al. "Higher set point plasma viral load and more-severe acute HIV type 1 (HIV-1) illness predict mortality among high-risk HIV-1-infected African women". eng. In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 42.9 (May 2006), pp. 1333–1339. ISSN: 1537-6591.

[16] Catherine Chirouze et al. "Viremia copy-years as a predictive marker of all-cause mortality in HIV-1-infected patients initiating a protease inhibitor-containing antiretroviral treatment". eng. In: *Journal of Acquired Immune Deficiency Syndromes (1999)* 68.2 (Feb. 2015), pp. 204–208. ISSN: 1944-7884.

[17] Ramy A. Arnaout et al. "A simple relationship between viral load and survival time in HIV-1 infection". In: *Proceedings of the National Academy of Sciences of the United States of America* 96.20 (Sept. 1999), pp. 11549–11553. ISSN: 0027-8424. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC18071/ (visited on 10/30/2015).

[18] Christophe Fraser et al. "Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.44 (Oct. 2007), pp. 17441–17446. ISSN: 0027-8424. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2077275/ (visited on 10/30/2015).

[19] Ping An and Cheryl A. Winkler. "Host genes associated with HIV/AIDS: advances in gene discovery". eng. In: *Trends in genetics: TIG* 26.3 (Mar. 2010), pp. 119–131. ISSN: 0168-9525.

[20] James Honaker, Gary King, and Matthew Blackwell. *AMELIA II*. Dec. 2015. URL: https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf.

[21] C. Vidal et al. “Lack of evidence of a stable viral load set-point in early stage asymptomatic patients with chronic HIV-1 infection”. eng. In: *AIDS (London, England)* 12.11 (July 1998), pp. 1285–1289. ISSN: 0269-9370.

[22] J. Masel et al. “Fluctuations in HIV-1 viral load are correlated to CD4+ T-lymphocyte count during the natural course of infection”. eng. In: *Journal of Acquired Immune Deficiency Syndromes (1999)* 23.5 (Apr. 2000), pp. 375–379. ISSN: 1525-4135.

[23] Nikos Pantazis et al. “Temporal trends in prognostic markers of HIV-1 virulence and transmissibility: an observational cohort study”. en. In: *The Lancet HIV* 1.3 (Dec. 2014), e119–e126. ISSN: 23523018. URL: http://linkinghub.elsevier.com/retrieve/pii/S2352301814000022 (visited on 06/07/2016).

[24] Joshua Herbeck et al. “Evolution of HIV virulence in response to widespread scale up of antiretroviral therapy: a modeling study”. en. In: *bioRxiv* (Mar. 2016), p. 039560. URL: http://biorxiv.org/content/early/2016/03/17/039560 (visited on 06/07/2016).

[25] Christophe Fraser et al. “Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective”. eng. In: *Science (New York, N.Y.)* 343.6177 (Mar. 2014), p. 1243727. ISSN: 1095-9203.