REDS HTLV Cohort Analytic Dataset
Phases 1-3

This documentation describes the contents of the analytic dataset used for the HTLV Cohort Phase 1-3 analysis (HTLVP123.sas7bdat). This dataset was created from data gathered during Phases 1-3 of the HTLV Cohort Study. The selection of subjects and derivation of variables were guided by the needs of a single specific analysis. While this dataset may be useful for future analytic endeavors, it is of course the responsibility of individual investigators to ensure that these data are suitable for any analyses conducted.

SAS Dataset

The analytic dataset is a SAS dataset, HTLVP123.sas7bdat. The SAS PROC FORMAT code for creating formats (i.e. value labels) for the variables is P123ProcFormat.txt.

Selection Criteria

The following subject inclusion criteria were applied:

- Phase 1 Interview File
- Completed Phase 1 Level I Medical Exam
- Positive Donor or Control
- Known HTLV Status

The analytic file has *n*=1,338 subjects.

Variables of Interest

PERSONID	Person/Subject Identifier

Independent Variables

HTLVTYPE	HTLV Status
P1CENTER	Center
D_TYPE	Donation Type
GENDER	Gender
RACE	Race
P3AGE	Age at Final Phase
P3EDUC	Education Level at Final Phase
DRINKW	Avg Drinks per week while Drinking through Final Phase
PACKYRS	Total Pack-Years through Final Phase
P3IDU	Intravenous Drug Use through Final Phase
P3SEXN	Total Number of Sex Partners through Final Phase

PERDAYS	Period (Number of Days Covered by Dependent Variables)

Dependent Variables

PRSWALK	Symptom: Difficulty Walking
PRSRISE	Symptom: Difficulty Rising
PRSCLMB	Symptom: Difficulty Climbing
PRSURG	Symptom: Urinary Urgency
PRSINC	Symptom: Urinary Incontinence
PRSVURG	Symptom: Post-Void Urgency
PRSIMP	Symptom: Impotence
PRSFEET	Symptom: Burning Feet
PRSGLND	Symptom: Swollen Glands
PRSFEVR	Symptom: Fever
PRSSWET	Symptom: Night Sweats
PRSWTLS	Symptom: Weight Loss
PRSGINC	Symptom: Loss of Bowel Control
PNEU	Acute: Pneumonia in Phase 2/3 [Y/N]
BRON	Acute: Bronchitis in Phase 2/3 [Y/N]
BLAD	Acute: Bladder in Phase 2/3 [Y/N]
KIDN	Acute: Kidney in Phase 2/3 [Y/N]
P3BK	Acute: Bladder/Kidney in Phase 2/3 [Y/N]
PNEUDAYS	Acute: Number of Days from P1DATE until Pneumonia in Phase 2/3
BRONDAYS	Acute: Number of Days from P1DATE until Bronchitis in Phase 2/3
BLADDAYS	Acute: Number of Days from P1DATE until Bladder infection in Phase 2/3
KIDNDAYS	Acute: Number of Days from P1DATE until Kidney infection in Phase 2/3
BKDAYS	Acute: Number of Days from P1DATE until Bladder/Kidney infection in Phase 2/3
P3PNEU	Acute: Number of Times Treated for Pneumonia P1-P3
P3BRON	Acute: Number of Times Treated for Bronchitis P1-P3
P3BLAD	Acute: Number of Times Treated for Bladder Infection P1-P3
P3KIDN	Acute: Number of Times Treated for Kidney Infection P1-P3
P3BLADKI	Acute: Number of Times Treated for Bladder/Kidney Infection P1-P3
ARTH	Chronic: Incident Arthritis in Phase 2/3
THYR	Chronic: Incident Thyroid Disease in Phase 2/3
HTN	Chronic: Incident Hypertension in Phase 2/3
DM	Chronic: Incident Diabetes in Phase 2/3
ASTH	Chronic: Incident Asthma in Phase 2/3
HERP	Chronic: Incident Herpes in Phase 2/3
CANCER	Chronic: Incident Cancer in Phase 2/3
ARTHDAYS	Chronic: Number of Days from P1DATE until incident Arthritis in Phase 2/3
THYRDAYS	Chronic: Number of Days from P1DATE until incident Thyroid in Phase 2/3
HTNDAYS	Chronic: Number of Days from P1DATE until incident Hypertension in Phase 2/3
DMDAYS	Chronic: Number of Days from P1DATE until incident Diabetes in Phase 2/3

ASTHDAYS	Chronic: Number of Days from P1DATE until incident Asthma in Phase 2/3
HERPDAYS	Chronic: Number of Days from P1DATE until incident Herpes in Phase 2/3
CANDAYS	Chronic: Number of Days from P1DATE until incident Cancer in Phase 2/3

Derived Variables

Missing Data

Many of the derived variables are computed from multiple component variables. A decision must be made regarding what is done when one or more of the component items is missing. For the independent variables, missing items resulted in missing derived variables. For example, a subject that reported no intravenous drug use in Phases 1 and Phase 2 but refused to answer the question in Phase 3 would have P3IDU set to missing. For dependent variables, it was decided that missing items would be treated as responses of "no". That is, a subject who reported never having been treated for Pneumonia in Phase1 and Phase 2 but skipped the question in Phase 3 would have PNEU=0. These imputed values may result in a slight downward bias of the data, but the rate of missing items is low enough that this is likely to have little effect.

Note that these rules apply to *missing items* on a completed interview, distinct from an entire interview being missing. The interview itself is designed to accommodate skipped phases, as questions are typically asked "since your last interview...".

Independent Variables

Characteristics that do not change over time (HTLVTYPE, P1CENTER, D_TYPE, GENDER, RACE) were taken from Phase 1 data. Characteristics that may change over time were taken either from the final phase completed (P3AGE, P3EDUC) or from a combination of data from all completed phases (DRINKW, PACKYRS, P3IDU, P3SEXN, PERDAYS).[1]

P3AGE	Integer age (as it would be self-reported) on date of final phase.
P3EDUC	Education level as of final phase. Because the education question was not asked in Phase 2, subjects that completed Phase 1 and Phase 2 (but not Phase 3) have education taken from the Phase 1 data.
DRINKW / PACKYRS	In Phase 1, lifetime drinking history and smoking history were not part of the main questionnaire. They were included in a one page insert. The rate of missing data is unusually high, suggesting that interviewers may have occasionally forgotten to complete the insert. In Phase 2, lifetime drinking and smoking history were included as part of the main instrument.

[1] Because subjects may differ in the number of phases completed, in some sense the "meaning" of these variables may differ across subjects. For example, for subjects that completed only Phase 1 (but not Phase 2 or Phase 3), the variable P3SEXN will indicate the number of sex partners reported in Phase 1. For subjects that completed only Phase 1 and Phase 2, the variable P3SEXN will indicate the number of sex partners reported in Phase 1 and 2 combined.

For subjects who completed Phase 2, we use the Phase 2 lifetime data as the baseline. For those who did not complete Phase 2, we use the Phase 1 lifetime data as the baseline. The baseline smoking and drinking questions ask subjects the age they started smoking/drinking and the age they stopped smoking/drinking. To compute number of years smoking/drinking from these ages, we assume that the subject starts and stops on a birthday. The number of years is then computed as (stop date-start date)/365.25. One artifact of this method is that a subject who reports starting smoking at age 18 and stopping smoking at age 18 will have PACKYRS=0.

DRINKW Average Drinks Per Week while Drinking is computed as a weighted average of: number of drinks per week at baseline; number of drinks per week between baseline and Phase 3.

DRINKW=[(BASETDR*BASEDRINKW) + (P3TDR*P3DRINKW)] / (BASETDR + P3TDR)

where

BASETDR	=Baseline time (years) drinking
BASEDRINKW	=Baseline drinks per week
P3TDR	=Time (years) drinking between baseline and Phase 3
P3DRINKW	=Drinks per week between baseline and Phase 3

Note that this quantity represents average drinks per week *while drinking,* rather than an estimate of the total amount of lifetime drinking, or even average drinks per week. A subject that drinks 1/week for 10 years will have DRINKW=1; and a subject that drinks 1/week for 5 years then 0/wk for 5 years will also have DRINKW=1. Perhaps more surprisingly, a subject that drinks 10/week for 5 years then cuts down to 5/week for 5 years will have DRINKW=7.5, but a subject that drinks 10/week for 5 years then 0/week for 5 years will have DRINKW=10.

Subjects reporting drinking less than one drink per year are treated as non-drinkers.

PACKYRS Pack-Years is computed as the (total number of years smoking regularly at baseline)*(number of cigarettes smoked per day/20) + (total number of years smoking regularly between baseline

and Phase 3)*(number of cigarettes smoked per day/20). When cigarettes smoked per day was reported as "less than one" it was set to 0.5.

P3SEXN Total number of sex partners through final phase.

In Phase 1 some subjects could not specify the number of lifetime sex partners. The following recodes were used:

"not very many"	=missing
"numerous"	=6
"greater than 200"	=200
"1000 or more"	=1000

In Phase 3 the sex partner question asked for number of sex partners since previous interview, not number of *new* partners. Thus, it is possible that some partners included in Phase 1 or Phase 2 were double-counted in Phase 3.

P3IDU History of intravenous drug use through final phase. This variable is coded into three levels: Current IDU; Past IDU; Never IDU. Classification as "Current" indicates that IDU was reported on the final phase completed. The Phase 1 questionnaire asked about lifetime IDU. A subject that completed Phase 1 only, and reported lifetime IDU, is coded as Current IDU. A subject that reported lifetime IDU in Phase 1 and reported no IDU in Phase 2 and Phase 3 would be coded as Past IDU.

PERDAYS The Phase 1 dependent variables ask about the past 5 years. PERDAYS is the number of days covered for the subject. It is 5 * 365.25 + (final interview date – Phase 1 interview date).

Dependent Variables

Symptoms

PRSWALK, PRSRISE, PSCLMB, PRSURG, PRSINC, PRSVURG, PRSIMP, PRSFEET, PRSGLND, PRSFEVR, PRSSWET, PRSWTLS, PRSGINC

The symptom variables are coded 1/0 indicating whether or not a symptom was reported in any phase. Female subjects have PRSIMP set to missing. These variables are only computed for subjects that completed Phase 3.

Acute [Y/N]

PNEU BRON BLAD KIDN P3BK

The acute [Y/N] variables indicate whether or not a condition was reported in Phase 2 or Phase 3 (regardless of whether or not it was reported in Phase 1). A small number of subjects reported number of times treated = "chronic/continuous". For these cases the acute [Y/N] variable was set to missing. These variables are only computed for subjects that completed Phase 2 or Phase 3.

Acute [DAYS]

PNEUDAYS, BRONDAYS, BLADDAYS, KIDNDAYS, BKDAYS

The acute [DAYS] variables indicate the number of days from the Phase 1 interview date until the date of first treatment in Phase 2 / Phase 3 or date of censoring (final interview date). When month of first treatment was missing, it was taken from the interview month. Subjects reporting number of times treated as "chronic/continuous" have this variable set to missing. Four subjects that have a date of first treatment prior to the Phase 1 interview date are set to missing. These variables are only computed for subjects that completed Phase 2 or Phase 3.

Acute [TIMES]

P3PNEU, P3BRON, P3BLAD, P3KIDN, P3BLADKI

The acute [TIMES] variables indicate the total number of times a subject reported being treated for a condition across all phases. Subjects reporting the number of times treated as "chronic/continuous" have variable set to missing. These variables are computed for all subjects.

Chronic [Y/N]

ARTH, THYR, HTN, DM, ASTH, HERP

The chronic [Y/N] variables indicate an incident case of the condition. These variables are computed for subjects that reported never having had the condition at Phase 1, and completed either Phase 2 or Phase 3.

Chronic [DAYS]

ARTHDAYS, THYRDAYS, HTNDAYS, DMDAYS, ASTHDAYS, HERPDAYS

The chronic [DAYS] variables indicate the number of days from Phase 1 interview date until date of first incident case of the condition in Phase 2 / Phase 3 or date of censoring (final interview date). When month of first treatment was missing, it was taken from the interview month. Four subjects that have a date of diagnosis prior to the Phase 1 interview date are set to missing. These variables are computed for subjects that reported never having had the condition at Phase 1, and completed either Phase 2 or Phase 3.