Respiratory Illness Study and Data Description

The respiratory data used for some course exercises is disguised data from a longitudinal study of people with a certain chronic respiratory illness. The goal of the study was simply to follow these people over time, not to compare them with a control group or treat their illness in any way. Information was collected at a baseline visit, and after that participants were contacted approximately once per year to see how they were doing and collect a small amount of additional information.

Participants enrolled in the study over a time span of several years, so when the study stopped after about six years, various numbers of people had completed different numbers of follow-up visits. That is, some participants enrolled near the end of the study and therefore had only the baseline visit, while others enrolled near the beginning of the study and had as many as five follow-up contacts. The participant count for each follow-up year steadily decreases not because people dropped out of the study, but because decreasing numbers of people had reached their anniversary date for each follow-up visit when the study itself was stopped.

The data comes in the form of four data sets, for each of which the corresponding data collection form is provided for your reference. The data were actually collected with an online data entry system, from which the forms might differ in very small ways.

Form	**Data set name**	**Details**
Demographic Form	DEM	Contains one observation per study participant and should be considered to define the study population.
Consent Documentation Form	CDF	Contains information on each participant's consent to be in the study, which can change over time. Thus, it can contain multiple observations per participant. In general, participants should be used for analysis only if their last CDF record has the value F, P, or D for variable CDFA2.
Clinical Procedures Form	CLN	Contains one observation for most study participants.
Annual Follow-up Form	FUP	Contains one or more observations for most study participants.

Each data set is sorted by key variables FAKEID VISIT FSEQNO. No data set should contain duplicates when these three variables are considered. These variables mean the following:

FAKEID	The study participant identifier. An ID variable that uniquely identifies each participant and can be used to link across data sets. Is a character variable but has values such as 0001, 0087, 1042, and 1577 in which all characters should be digits between 0 and 9.
VISIT	The participant visit identifier. A numeric variable with values 1-6 that indicates at which participant contact the data was collected. 1 signals the baseline visit, with values 2-6 representing follow-up years 1-5 respectively. Only data set FUP contains VISIT values greater than 1.
FSEQNO	A within-visit sequence identifier. This numeric variable is a record sequence indicator and contains values 0 and above. In general, all FSEQNO values are 0 except in data set CDF. If a participant modified his or her consent to be in the study or his or her desired level of involvement, a new CDF record with VISIT=1 was entered but with the next highest FSEQNO value. Thus, the participant's final CDF record (if last.visit) when sorting by

	FAKEID VISIT FSEQNO is the one that should be used when determining whether that participant's data should be used for analysis.

Here's an example of a valid collection of participant records in the CDF data set:

FAKEID	VISIT	FSEQNO	CDFA1	CDFA2	CDFA3
0319	1	0	C	F	N
0319	1	1	M	W	X

This participant actually decided to withdraw from the study (that's what CDFA2='W' means), so we would not want to use this participant's data in any report or analysis. But the important point here is that there are duplicate records in CDF for participant 0319 when we look only at FAKEID or only at FAKEID VISIT, but not when we look at FAKEID VISIT FSEQNO. So this participant's CDF records are absolutely fine, and their last record is the one that we would want to look at when deciding about eligibility for analysis.

Note that all date variables were removed from the data sets, along with many other variables, especially from the clinical procedures and follow-up data sets.