

# The NLS Investigator

William Howard Beasley (Howard Live Oak LLC, Norman)  
Joseph Lee Rodgers (University of Oklahoma, Norman)  
Kelly Meredith (University of Oklahoma, Norman)  
David Bard (University of Oklahoma Health Sciences Center, OKC)

April 21, 2012

## Abstract

This vignettes will be useful to behavior genetic researchers interested in using the National Longitudinal Survey of Youth (NLSY79) or Children (NLSYC) data. To fit biometrical models to the NLSY or NLSYC requires that they be extracted from the appropriate online NLS database into a usable file format. The extracting software is called the NLS Investigator. We describe how to use the NLS Investigator to select and download variables. In subsequent vignettes, we show how to reformat the data extract into file structures that support behavior genetic research, and how to fit biometrical models. The following steps are not specific to R, but rather precede the use of any analytic statistical software.

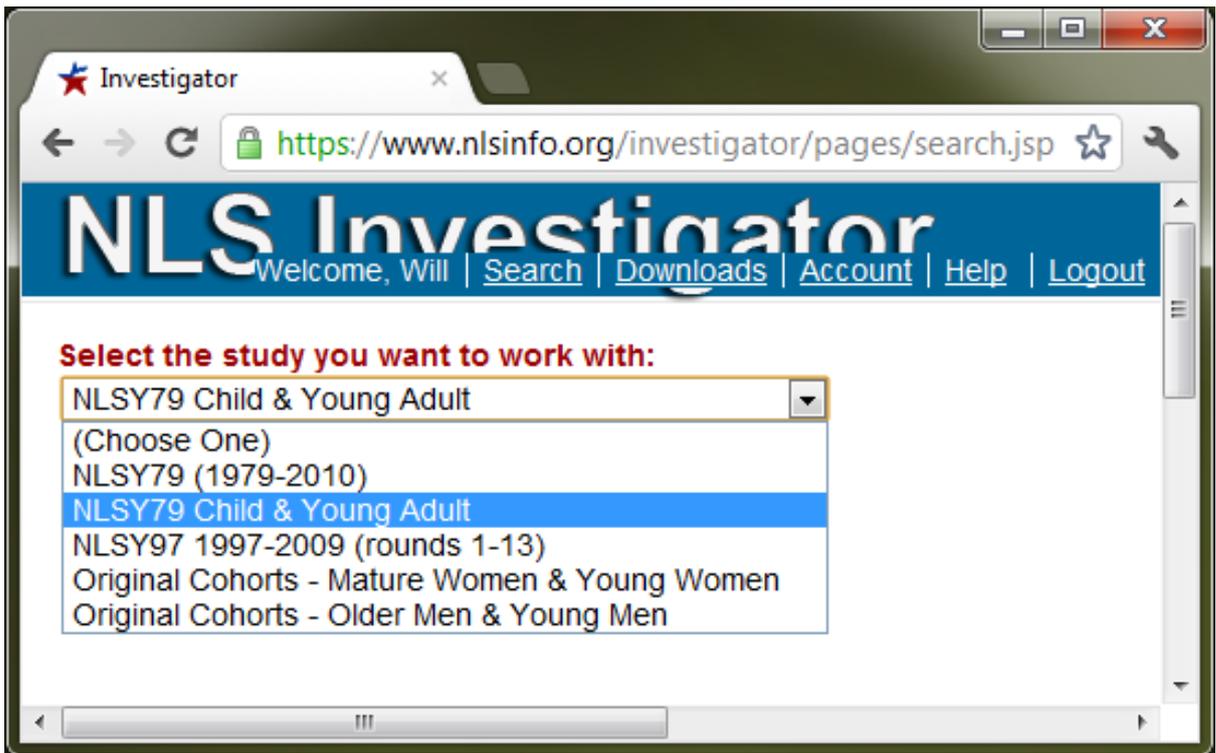
## 1 Terminology

This package considers both generations of the NLSY79. The ‘first generation’ refers to subjects in the original NLSY79 sample (<http://www.bls.gov/nls/nlsy79.htm>). The second generation of subjects are the biological offspring of the original females -*i.e.*, those in the NLSY79 Children and Young Adults sample ( <http://www.bls.gov/nls/nlsy79ch.htm>). The NLSY97 is a third dataset that can be used for behavior genetic research (<http://www.bls.gov/nls/nlsy97.htm>), although this vignette focuses on the two generations in the NLSY79.

Standard terminology is to refer second generation subjects as ‘children’ when they are younger than age 15 (NLSYC), and as ‘young adults’ when they are 15 and older (NLSY79-YA); though they are the same respondents, different funding mechanisms and different survey items necessitate the distinction. This cohort is sometimes abbreviated as ‘NLSY79-C’, ‘NLSY79C’, ‘NLSY-C’ or ‘NLSYC’.

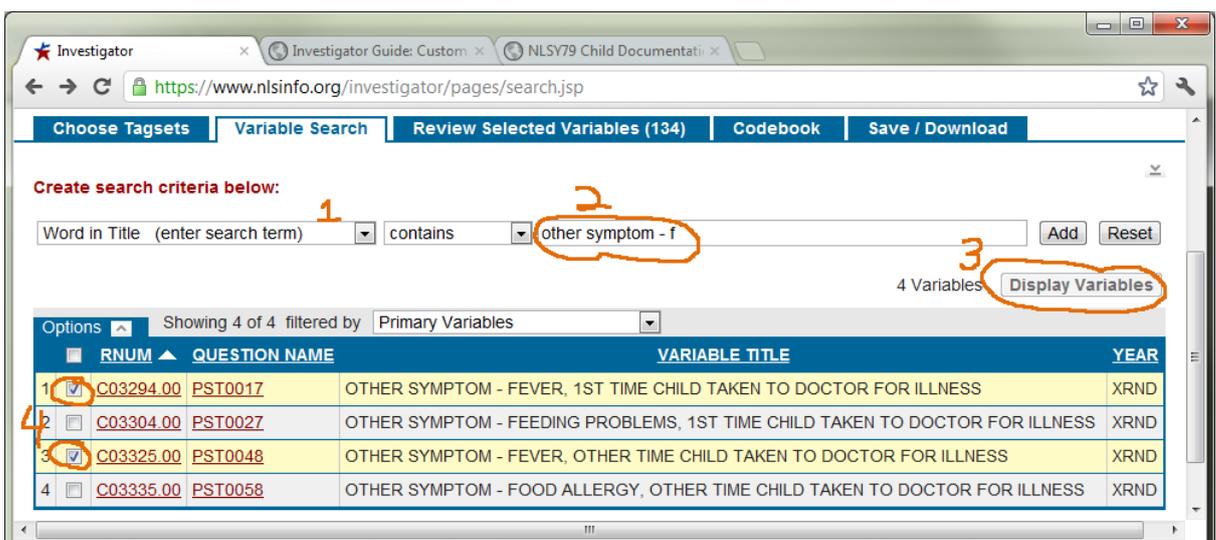
## 2 Steps to Select Variables and Download Data

1. Browse to <http://www.nlsinfo.org/investigator/>. Select the ‘REGISTER’ link in the top right, and create a personal account. If you have already registered, Log In and proceed to the next step.
2. Select your desired cohort, in the dropdown box titled, “Select the study you want to work with:”. In the screenshot below, second generation of the NLSY79 sample is selected.



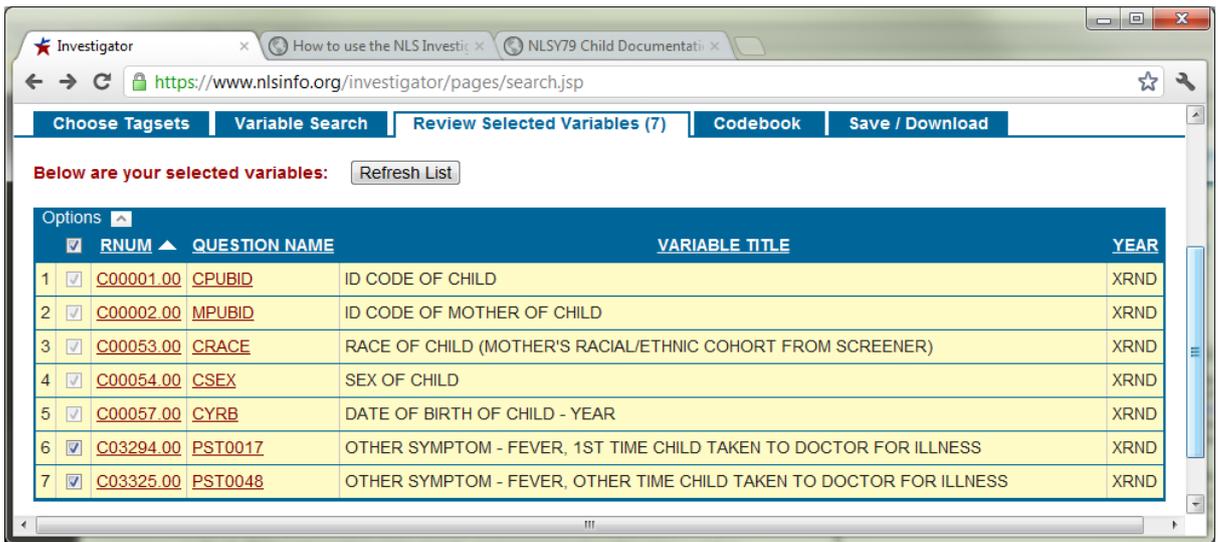
3. Select the variables. There are tens of thousands of variables in some cohorts, and selecting the correct ones can require careful attention and a few tricks. A thorough tutorial begins on the NLS Investigator page: [http://www.nlsinfo.org/InvestigatorGuide/investigator\\_guide\\_TOC.html](http://www.nlsinfo.org/InvestigatorGuide/investigator_guide_TOC.html). To better leverage the NLSY's extensive variable set (and to avoid mistakes), we recommend that researchers dedicate time to this tutorial. However for the purposes of this vignette, we'll simply select a few easy variables.

First, in the 'Variables Search' tab, select 'Word in Title (enter search term)'. Second, type "other symptom - f" in the textbox. Third, clicking the 'Display Variables' button should retrieve at least four NLSYC variables whose title starts with "other symptom - f". Fourth, supposing we care about only their fevers, click their two corresponding checkboxes. (If you're curious, the 'XRND' value for year stands for [cross round](#); XRND variables are calculated by the NLS staff, and typically come from the subject's most recent survey).



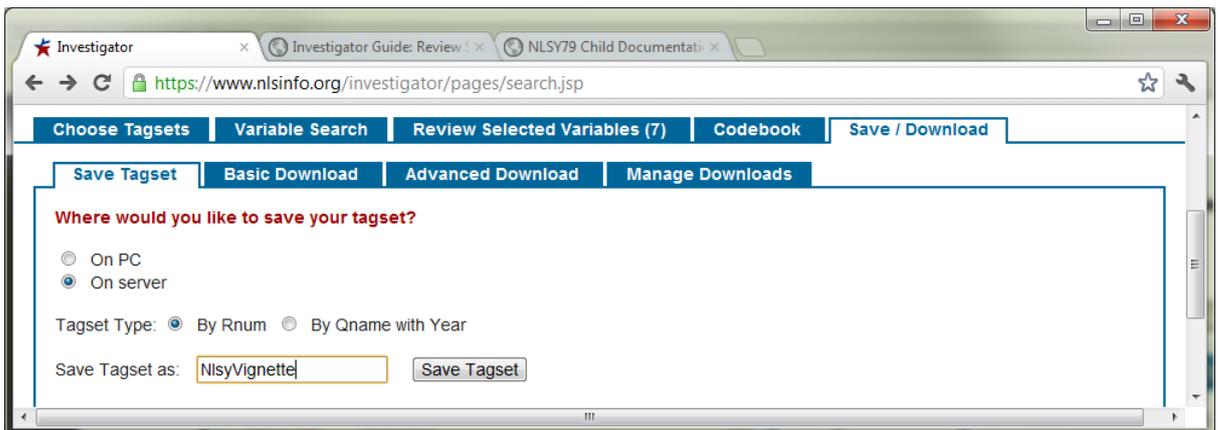
4. Review your selected variables by clicking the corresponding tab. Notice that several important variables are automatically included in every dataset. In real research these steps are iterated many times, as you repeated select, then review, then save, then re-select, then re-review, then... But

we'll move on, because these seven variables are good enough for an example.



- When the dataset is complete, it is time to save the tagset. A **tagset file** is simply metadata of the desired variables. The file identifies the variables, *but does not contain actual data values*. As your project evolves over time (because you're adding variables, or refreshing your dataset after a new survey is released), it's convenient to keep the metadata distinct from the real data.

The actual values are contained in the **data file**, which is discussed in the next two steps. These issues are covered further in the [official tutorial](#).



- The first step of the download process is to create the data file on the NLS server. Click the 'Save/Download' tab, and then the 'Advanced Download' tab. As far as the NlsyLinks package is concerned, only the the 'Comma-delimited datafile...' box needs to be checked. Then provide a 'Data filename'. Finally, click the 'Start Download' button.

Choose Tagsets | Variable Search | Review Selected Variables (130) | Codebook | Save / Download

Save Tagset | Basic Download | **Advanced Download** | Manage Downloads

**Select the options for your custom download:**

**Create Download of Data**

- Tagset (list of selected variables)
- SAS® control file (includes the datafile of selected variables)
- SPSS® control file (includes the datafile of selected variables)
- STATA® dictionary file of selected variables
- Codebook of selected variables
- Short Description File
- Comma-delimited datafile of selected variables (to be read in Excel, etc.)
  - Column headers -- Use  Reference Number  Question Name (Does not guarantee uniqueness)

**Create Frequency / Table**

**Apply Universe Restrictors** ([How to use Universe Restrictors](#))

Data filename:   (status will appear under 'Manage Downloads' tab)

7. The second step of the download process is to transfer the zip file to your local computer. Click the maroon 'download' hyperlink.

Choose Tagsets | Variable Search | Review Selected Variables (130) | Codebook | Save / Download

Save Tagset | Basic Download | Advanced Download | **Manage Downloads**

**Download Status:**

All downloads are available. Please click a download link below to begin downloading.

**All Available Downloads:**

<input type="checkbox"/>	Date	Study	Name	Size	Download
1 <input type="checkbox"/>	2012-02-23 22:03:20	NLSCYA	NlsyVignette	473.4K	<a href="#">download</a>

8. Open the zip file and extract the \*.csv file to a location that the vignette examples have permission to read. Then try some of the NlsyLinks vignette examples at <http://cran.r-project.org/web/packages/NlsyLinks/>.

### 3 Using Multiple Tagsets

Tagsets were introduced in Step 5, but are discussed more thoroughly here. Tagsets make large projects more manageable in two ways. First, they save effort and reduce errors because you don't have to re-select all the variables every time you revisit the NLS Investigator. You can save and load the tagsets during subsequent sessions. In fact, you can save multiple tagsets on the NLS server.

Second, using *multiple* tagsets provides a convenient approach to organize and compartmentalize your variables. It's not unusual for some complicated longitudinal studies to use hundreds of NLSY variables. We've found it easier to manage five tagsets of 100 variables, than one tagset of 500. Since all NLSY tagsets (and therefore their datasets) automatically include the subject ID, merging the multiple datasets later is trivial with statistical software.

In R, use the `merge` function; the `by` argument should be 'R0000100' for NLSY79 subjects, 'C0000100' for NLSYC subjects, and 'R0000100' for NLSY97 subjects. For example, a study about intelligence and teenage fertility would have one tagset containing the intelligence variables, and a second tagset containing the fertility variables. These two tagsets eventually could produce two CSV data files called

IQ.csv and Fertility.csv, located in the BGRResearch directory. The R code to read and merge these two datasets could be as simple as

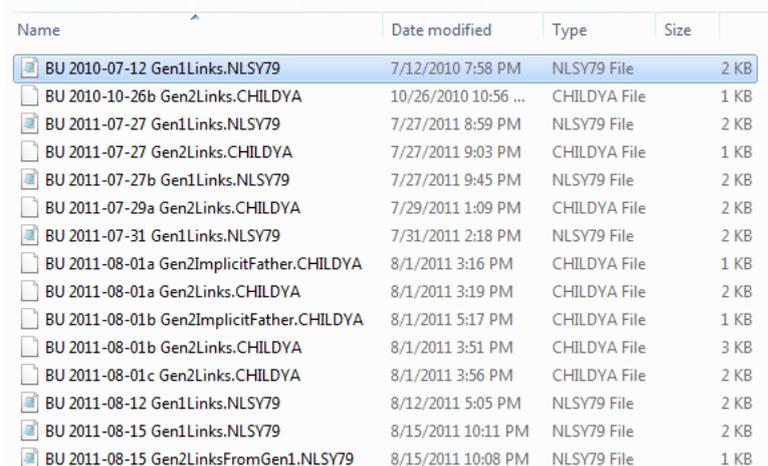
```
dsIQ <- read.csv('C:/BGRResearch/IQ.csv', header=TRUE)
dsFertility <- read.csv('C:/BGRResearch/Fertility.csv', header=TRUE)
ds <- merge(dsIQ, dsFertility, by="C0000100")
```

In SAS, the merge function (and its BY argument) behaves similarly as above. Notice that the SAS documentation states, “Before you can perform a match-merge, all data sets must be sorted by the variables that you want to use for the merge”, which is C0000100 in this example.

If you’re importing CSV files specifically for the NlsyLinks package, you’ll find functions like ReadCsvNlsy79Gen1 and ReadCsvNlsy79Gen2 more convenient for later analyses.

## 4 Tagset History

We recommend saving the tagsets on both the NLS server and your local machine. Tagsets are not guaranteed to be retained on the NLS server more than 90 days. Locally save when variables are added or removed from the tagset in a consistent location, and name the file to reflect the current date. The files can be read with any simple text editor. The directory below contains the partial evolution of four distinct tagsets (i.e., ‘Gen1Links’, ‘Gen2Links’, ‘Gen2ImplicitFather’, ‘Gen2LinksFromGen1’).



Name	Date modified	Type	Size
BU 2010-07-12 Gen1Links.NLSY79	7/12/2010 7:58 PM	NLSY79 File	2 KB
BU 2010-10-26b Gen2Links.CHILDYA	10/26/2010 10:56 ...	CHILDYA File	1 KB
BU 2011-07-27 Gen1Links.NLSY79	7/27/2011 8:59 PM	NLSY79 File	2 KB
BU 2011-07-27 Gen2Links.CHILDYA	7/27/2011 9:03 PM	CHILDYA File	1 KB
BU 2011-07-27b Gen1Links.NLSY79	7/27/2011 9:45 PM	NLSY79 File	2 KB
BU 2011-07-29a Gen2Links.CHILDYA	7/29/2011 1:09 PM	CHILDYA File	2 KB
BU 2011-07-31 Gen1Links.NLSY79	7/31/2011 2:18 PM	NLSY79 File	2 KB
BU 2011-08-01a Gen2ImplicitFather.CHILDYA	8/1/2011 3:16 PM	CHILDYA File	1 KB
BU 2011-08-01a Gen2Links.CHILDYA	8/1/2011 3:19 PM	CHILDYA File	2 KB
BU 2011-08-01b Gen2ImplicitFather.CHILDYA	8/1/2011 5:17 PM	CHILDYA File	1 KB
BU 2011-08-01b Gen2Links.CHILDYA	8/1/2011 3:51 PM	CHILDYA File	3 KB
BU 2011-08-01c Gen2Links.CHILDYA	8/1/2011 3:56 PM	CHILDYA File	2 KB
BU 2011-08-12 Gen1Links.NLSY79	8/12/2011 5:05 PM	NLSY79 File	2 KB
BU 2011-08-15 Gen1Links.NLSY79	8/15/2011 10:11 PM	NLSY79 File	2 KB
BU 2011-08-15 Gen2LinksFromGen1.NLSY79	8/15/2011 10:08 PM	NLSY79 File	1 KB

## 5 Notes

This package’s development was largely supported by the NIH Grant 1R01HD65865, “NLSY Kinship Links: Reliable and Valid Sibling Identification” (PI: Joe Rodgers; Vignette Construction by Will Beasley).

These screenshots were taken February and March, 2012 with Google Chrome 17 and Windows 7 SP1 Enterprise.