README: *Orange Book* patent and exclusivity data

July 17, 2018

Contents

1 Overview

This document describes our construction of digital versions of the US Food and Drug Administration (FDA)'s *Orange Book* patent and exclusivity tables for years 1985-2016 (no *Orange Book* was published in 1986). PDF versions of the *Orange Books* were obtained via a FOIA (Freedom of Information Act) request, and data from these PDF files was either hand-entered or parsed in order to create the digital files. Descriptions of the folders and files in this directory are in Section 2. Descriptions of the digital files' contents can be found in Section 3. Descriptions of the data construction can be found in Section 4.

2 Directory structure and files

The directory that houses the raw and digitized data is comprised of the following folders and files:

- `/1_orange_book_PDFs/` contains the full FDA *Orange Books*, obtained via a FOIA request, for years 1980-2016 (Patent and Exclusivity tables begin in 1985). This folder also contains excerpts of the PDFs that were sent to a data entry firm for hand-entry. Also within this directory is documentation of the FOIA request.
- `/2_hand_entered_by_firm_excel/` contains the raw Excel files as entered by the data entry firm.
- `/3_cross_check_sources/` contains the Stata files and PDFs that were used for cross-checking the data entry firm's output.
- `/4_clean_exclusivity_tables_stata/` contains the clean data files as well as code and other intermediate files used in creating the clean files, including the following:
  - The subfolder `/scripts/` contains the .do file that creates the clean data sets and the .log file.

- The subfolder `/corrected_discrepancies_excel/` contains hand-entered corrections made during data construction. These files should not be deleted or altered in any way.
- There are three subfolders – `/temp/`, `/txt/`, and `/exported_discrepancies_excel/` that are created when the .do file is run. These subfolders can safely be removed after the .do file completes.
- Running the file **create_final_data.do** creates the clean Stata files. The code was written for Stata 15 running on a Linux operating system.
- Each of the files contains data for all *Orange Books* 1985-2016 (excluding 1986, for which there is no *Orange Book*). If you want exclusivity data only for a particular edition, simply open the data file and keep only that year.

3 Description of digital files

The final set of digital files consists of four data files. The **FDA_drug_patents.dta** file records active (at the time of each *Orange Book* edition's creation) patents associated with FDA-approved drugs. This file can be linked with the **FDA_patent_use_codes.dta** file in order to obtain the text description of a patent's "use code" for all patents that have such associated codes.

The **FDA_drug_exclusivity.dta** file records exclusivity periods granted by the FDA. This file can be linked with the **FDA_drug_exclusivity_codes.dta** file in order to obtain the text description of the exclusivity codes.

Throughout this document we use the term "key variables" to describe those variables that identify unique observations within a data set.

3.1 FDA_drug_patents.dta

For the purposes of understanding the structure of the exclusivity tables, Figure 1 provides an example of the patent and exclusivity information available for the drug Avodart from page ADA 53 of the 2010 *Orange Book*. The data includes the drug's application number ("N021319"), product number ("001"), active ingredient(s) ("Dutasteride"), and trade name ("Avodart"). Under these four identifying pieces of information are listed three active patents (that is, active in 2010, when this version of the *Orange Book* was published), along with the expiration dates and any codes associated with each patent. In this case, two of the patents are associated with "DS" (indicating a drug substance claim) and "DP" (indicating a drug product claim) and two are associated with use codes (U-476 and U-477). The meanings of these codes can be found in **FDA_patent_use_codes.dta**. There are no entries under "PATENT DELIST REQUESTED," indicating that the drug sponsor has not requested that any of the patents be removed from the listing. The last two columns, "EXCLUSIVITY CODE(S)" AND "EXCLUSIVITY EXPIRATION DATE," concern FDA-granted exclusivity, not patent information, and will be discussed in Section 3.3 below.

Figure 1: **Excerpt of *Orange Book* Patent and Exclusivity Data**

APPL/PROD NO	PATENT NO	PATENT EXPIRATION DATE	PATENT CODES	PATENT DELIST REQUESTED	EXCLUSIVITY CODE(S)	EXCLUSIVITY EXPIRATION DATE
DUTASTERIDE - AVODART						
N021319 001	5565467	Nov 20, 2015	DS DP		I-565	Jun 19, 2011
	5846976	Sep 17, 2013	U-476			
	5998427	Sep 17, 2013	DS DP U-477			

Note: Data is excerpted from page ADA 53 of the 2010 *Orange Book*.

Table 1: **Example data records for Avodart patent information, 2010 *Orange Book***

*edition**	*active_ingredient*	*trade_name*	*application_type*	*application_number**	*product_number**	*patent_number**	*use_code**	*ped_extension**	*patent_expiration*	*DS*	*DP*	*delist_requested*
2010	DUTASTERIDE	AVODART	NDA	021319	001	5565467		0	20nov2015	1	1	0
2010	DUTASTERIDE	AVODART	NDA	021319	001	5846976	U-467	0	17sep2013	0	0	0
2010	DUTASTERIDE	AVODART	NDA	021319	001	5998427	U-477	0	17sep2013	1	1	0

Table 2: **Variables in FDA_drug_patents.dta**

Key variable	**Variable**	**Description**	**Coverage**					
			1985	**1987***	**1988-2003**	**2004-2008**	**2009**	**2010-2016**
X	*edition*	Year of *Orange Book* (1985-2016)	X	X	X	X	X	X
	active_ingredient	Active ingredient(s)**		X	X	X	X	X
	trade_name	Drug trade name		X	X	X	X	X
	application_type	NDA or ANDA						X
X	*application_number*	Six-digit FDA application number	X	X	X	X	X	X
X	*product_number*	Three-digit FDA product number (usually indicates different dosage sizes within an application number)	X	X	X	X	X	X
X	*patent_number*	USPTO-granted patent number. Usually utility patents, but sometimes reissue or design patents (prefixed with "RE" or "D").	X	X	X	X	X	X
X	*use_code*	Functional description of patent, where available.***			X	X	X	X
X	*ped_extension*	Indicator for whether the patent has received a 6-month exclusivity extension for performing pediatric studies.****	X	X	X	X	X	X
	patent_expiration	Patent expiration date (including any pediatric extensions).	X	X	X	X	X	X
	DS	Indicator for whether the patent has any drug substance claims.				X	X	X
	DP	Indicator for whether the patent has any drug product claims.				X	X	X
	delist_requested	Indicator for whether the drug sponsor has requested patent be removed from listing.					X	X

Notes: When a variable is not available for an edition, an entry of "N/A" is provided in the data file; this is why the indicator variables are given as strings ("0", "1", or "N/A").
* There is no *Orange Book* for the year 1986.
** Only a single active ingredient per drug is listed for years 1985-2004. From 2005 onward, all active ingredients are listed, each separated by a semicolon.
*** That is, *use_code* is given as "N/A" in 1985-1987, but may be populated or missing from 1988 onward.
**** Although we list pediatric exclusivity as non-missing for all editions, in practice the program was not started by the FDA until June 1998. Therefore, *ped_extension* takes on the value "0" for years 1985-1998 and "0" or "1" thereafter.

The three example patent records shown in Figure 1 would be digitized as in Table 1. Note that this table shows all the variables available in the **FDA_drug_patents.dta** file, although not all variables are available for each edition of the *Orange Book*. The key variables are marked with asterisks: *edition*, *application_number*, *product_number*, *patent_number*, *use_code*, and *ped_extension*. The first four of these variables – *edition*, *application_number*, *product_number*, and *patent_number* – identify a patent associated with a particular drug in a particular *Orange Book* edition. The variable *use_code* is also needed to uniquely identify an observation, because sometimes a patent is associated with more than one use code. Similarly, the variable *ped_extension* is also needed to uniquely identify an observation, because sometimes a single patent is listed both with and without a pediatric extension indicator.[1]

A description of each variable's meaning and coverage can be seen in Table 2.

3.2 FDA_patent_use_codes.dta

The **FDA_patent_use_codes.dta** file includes three variables: *edition*, *use_code*, and *use_code_desc*. The variables *edition* and *use_code* are the key variables that can be used to link this file to the **FDA_drug_patents.dta** file in order to interpret the meanings of the *use_code*s.

3.3 FDA_drug_exclusivity.dta

The exclusivity information for the drug Avodart excerpted in Figure 1 of this document would be recorded into the data file as shown in Table 3. The key variables are marked with asterisks: *edition*, *application_number*, *product_number*, *exclusivity_code*, and *exclusivity_expiration*. The expiration date is a key variable because sometimes a single drug product will have the same exclusivity code listed with multiple expiration dates. The FDA explained to us that this is possible because a single drug may be granted the same exclusivity multiple times; for example, a combination drug consisting of two New Chemical Entities (NCEs) may be entitled to two NCE exclusivities. The constructed variable *observation_count* records how many times a combination of key variables appears for a drug product in the raw data. This is almost always "1," but sometimes it is higher as the same exclusivity codes with the same expiration dates are sometimes listed multiple times for a drug product. The FDA explained to us that this is because a single drug may receive multiple grants of the same exclusivity, and those exclusivity periods may end on the same date.

A description of each variable's meaning and coverage can be seen in Table 4.

Table 3: **Example data records for Avodart exclusivity information, 2010 *Orange Book***

*edition**	*active_ingredient*	*trade_name*	*application_type*	*application_number**	*product_number**	*exclusivity_code**	*exclusivity_expiration**	*observation_count*
2010	DUTASTERIDE	AVODART	NDA	021319	001	I-565	19jun2011	1

[1] Pediatric extensions are indicated by the presence of "*PED" following the patent number in the PDFs. This suffix is removed in the clean data files and replaced with the *ped_extension* indicator variable.

Table 4: **Variables in FDA_drug_exclusivity.dta**

Key variable	Variable	Description	Coverage		
			1985	1987*-2009	2010-2016
X	*edition*	Year of *Orange Book* (1985-2016)	X	X	X
	active_ingredient	Active ingredient(s)**		X	X
	trade_name	Drug trade name		X	X
	application_type	NDA or ANDA			X
X	*application_number*	6-digit FDA application number	X	X	X
X	*product_number*	3-digit FDA product number (usually indicates different dosage sizes within an application number)	X	X	X
X	*exclusivity_code*	Reason for FDA-granted exclusivity	X	X	X
X	*exclusivity_expiration*	Expiration date of FDA-granted exclusivity	X	X	X
	observation_count	Number of exact duplicates of observation in *Orange Book*	X	X	X

Notes: When a variable is not available for an edition, an entry of "N/A" is provided in the data file.
* There is no *Orange Book* for the year 1986.
** Only a single active ingredient per drug is listed for years 1985-2004. From 2005 onward, all active ingredients are listed, each separated by a semicolon.

3.4 FDA_drug_exclusivity_codes.dta

The file **FDA_drug_exclusivity_codes.dta** file includes three variables: *edition*, *exclusivity_code*, and *exclusivity_code_desc*. The variables *edition* and *exclusivity_code* are the key variables that can be used to link this file with the **FDA_drug_exclusivity.dta** file in order to interpret the meanings of the *exclusivity_code*s.

There is one observation in **FDA_drug_exclusivity.dta** that does not successfully link between these two files: the exclusivity code I-55 for the 1990 edition. We assume this was a typo in the 1990 *Orange Book*.

4 Description of data construction

Our digitized data files were constructed and cross-checked through different methods depending on the availability of publicly available cross-checking sources. The PDFs for the years 2005-2016 were of very high visual quality and were converted to text files and parsed into the final data files. Random audits of this data have suggested that the data for these years seems to be of very high quality.

The PDFs for the years 1985-2004 were not of high enough quality to be reliably parsed into text files. A data entry firm, Digital Data Divide (DDD; https://www.digitaldividedata.com/), was instead hired to enter the tables and abbreviation lists for these editions. To improve the quality of these hand-entered data, we cross-checked the firm's entries with data files, generously provided to us by Bhaven Sampat and Scott Hemphill, that were hand-entered from publicly-available editions of the *Orange Books*.

4.1 Summary of sources used

Table 5 summarizes the entry method for each *Orange Book* edition as well as details of how each edition was cross checked.

Table 5: **Data sources and cross-checking sources**

Edition	Source	Entry method	Check/cross-check source
1985-1999	FOIA request	Hand-entry by DDD	Stata files hand-entered by Bhaven Sampat and Scott Hemphill from publicly-available scans of *Orange Books*
2000-2004	FOIA request	Hand-entry by DDD	Parse computer-readable versions obtained from FDA via Wayback Machine (http://archive.org/web/)
2005-2016	FOIA request	Parse	Random quality checks

4.2 Summary of Stata code

Below we provide a broad overview of the Stata code **create_final_data.do** used to create the patent and exclusivity tables (**FDA_drug_patents.dta** and **FDA_drug_exclusivity_codes.dta**) as well as the use/exclusivity code crosswalks (**FDA_patent_use_codes.dta** and **FDA_drug_exclusivity_codes.dta**).

4.2.1 Tables

Step 1: Parse tables for 2005-2016. The first section of Stata code converts the 2005-2016 editions from PDFs to .txt files. It then imports these .txt files into Stata and parses them into a usable format. Other than some final formatting that occurs later in the code, this completes the creation of the patent and exclusivity tables for 2005-2016.

Step 2: Prepare publicly-available cross-checking sources. The next section of code prepares Stata files and PDFs obtained from publicly available sources that will be used in cross-checking the data entered by Digital Data Divide. It can be split into two subsections: (A) Parsing PDF files for editions 2000-2004 and (B) cleaning hand-entered Stata files for 1985-1999.

Step 3: Internally harmonize Digital Data Divide-entered data. The next major chunk of code uses regular expressions to find errors in the data as entered by Digital Data Divide. For example, an application number must be six numeric characters. Any entries for application numbers that do not fit this format are exported to an Excel workbook for manual review. Corrections are made in this exported workbook and re-imported and corrected as needed.

Step 4: Cross check internally-harmonized patent data. The next section of code cross-checks the internally-harmonized Digital Data Divide data with the publicly-available cross-checking sources prepared in Step 2. This is done in three broad steps: First, we find "product groups" that appear in only the Digital Data Divide data (the "base" data) or in the comparison data. By product group, we mean an application number by product number combination. Product groups appearing in only the base or comparison source are exported to an Excel file for manual review, where changes, additions, or deletions to be made are noted. Second, after reconciling product groups across sources, we find combinations of key variables that appear only in one source but not the other. We export an Excel workbook of this set of discrepancies and manually code corrections until the two sources can be merge together perfectly on the key variables. Third, after making these corrections, we inspect differences on non-key variables (such as active ingredient, trade name, or patent expiration) and manually code corrections in an Excel workbook.

Step 5: Cross check internally-harmonized exclusivity data. This step is largely the same as Step 4, but for the FDA-granted exclusivity data (instead of patent data). First, product groups (application number by product number groups) appearing in only one source are identified and resolved. Second, combinations of key variables appearing in only one source are identified and corrected. Last, differences on non-key variables are resolved.

4.2.2 Abbreviation lists

Step 1: Parse abbreviation lists for 2005-2016 The first section of code converts the 2005-2016 abbreviation lists from PDF to .txt files. It then takes these .txt files and parses then into crosswalks between use/exclusivity codes and the codes' meanings.

Step 2: Identify entries missed by Digital Data Divide for 1985-2004 This section searches for codes that the data entry firm failed to enter for the years 1985-2004. It finds a small number of omissions and imports a hand-coded Excel workbook of additions.

Step 3: Clean files entered by Digital Data Divide Last, errors in exclusivity codes are identified by using regular expressions and corrections are made. No cross-checking source is used to check these data as we are not aware of any that is available.