Data Replication ReadMe File

Licensing Life-Saving Drugs for Developing Countries: Evidence from the Medicines Patent Pool

Alberto Galasso (University of Toronto) and Mark Schankerman (London School of Economics)

This readme file comprises of two parts. The first part, which begins on page 2, provides information on the posted datasets and the codes which can be used to replicate the three tables in the paper as well as the two graphs in Figure 1.

The second part of the file, which begins on page 6, focuses on the proprietary data used to construct the three graphs in Figure 2. This second set of data has been obtained under an arrangement that precludes the posting of the variables. For this part of the analysis we provide the information necessary to go from the raw data to the results of the paper (including the code).

All the results in the paper are obtained using STATA (version 14.2). Generating the tables and the figures requires the installation of the "estout" and "esttab" packages. Additional packages which may require installation in older versions of STATA are: "reghdfe" and "estpost".

PART 1 – POSTED DATA

1.1 Overview

Tables.do provides the STATA program that produces the three tables in the paper, using five distinct datasets:

country_drugs.dta

WB_status.dta

HIV_pop.dta

WB_TB.dta

HepC2015.dta

Figure1.do contains the STATA program that produces the two graphs in Figure 1 in the paper, using the dataset

country_drugs.dta

The next subsection provides additional details on the five analysis datasets.

1.2 Datasets description

country_drugs.dta

This dataset contains patent licensing information. It has been manually constructed by the authors using information collected in the portal MedsPal (www.medspal.org) and additional information obtained from the MPP. The file is used to produce Table 1, Table 2, Table 3 and Figure 1. The unit of analysis is a "country-product-year."

Variable definitions:

- country: name of the country
- product: name of the product
- priority_list: dummy=1 if the product is in the 2010 MPP priority list
- disease: variable indicating the main disease targeted by the drug. Values taken are HIV, H-C (hepatitis C), H-B (hepatitis B), TB (tuberculosis) and N/A (not available).

- Year: calendar year. Product-country combinations are included if at least one patent is active (non-expired).
- MPP: dummy =1 if the product-country is included in an MPP license in the year
- MPP_down: number of MPP sublicenses active in the country for the product in the calendar year
- Bilateral: number of bilateral sublicenses active in the country for the product in the calendar year
- Deals: total number of sublicenses (bilateral + MPP) active in the country for the product in the calendar year
- Access: dummy =1 if there is at least one sublicense (bilateral or MPP) active in the country for the product in the calendar year
- ever_MPP: dummy =1 if the product-country is included in the MPP by the end of the sample period.

WB_status.dta

This dataset contains information on the income classification of each country by the World Bank. It is used to produce Table 3. The unit of analysis is a "country-year." It is the STATA version of the file available at http://databank.worldbank.org/data/download/site-content/OGHIST.xls (downloaded in 2020) in which the country names have been manually adjusted to be merged with the country names in our country_drug.dta.

Variable definitions:

- country: name of the country
- Year: calendar year.
- Status: historical World Bank Analytical Classification for each country as of 2020. Values are H= high income, L=low income, LM= lower middle income, UM=upper middle income, and ..=missing classification

HIV_pop.dta

This dataset contains information on the population living with HIV in each country. It has been manually retrieved by the portal https://data.worldbank.org/ in 2019. Country names have been adjusted to be merged to the country_drug.dta dataset. It is used to produce Table 3. The unit of analysis is a "country-year."

Variable definitions:

- country: name of the country
- Year: calendar year
- Hiv_living: estimate of number of people living with HIV/AIDS. This includes Adults (ages 15+) and children (0-14 years) living with HIV. The World Bank data code for this indicator is SH.HIV.TOTL.HG.NUM

WB_TB.dta

This dataset contains information on the total population, and incidence of tubercolosis in each country. It has been manually retrieved by the portal https://data.worldbank.org/ in 2019. Country names have been adjusted to be merged to the country_drug.dta dataset. It is used to produce Table 3. The unit of analysis is a "country-year."

Variable definitions:

- country: name of the country
- Year: calendar year
- Pop_tot: estimate of de facto mid-year population in the country. The World Bank data code for this indicator is SP.POP.TOTL
- incidenceoftuberculosisper100000: Incidence of tuberculosis estimated as the number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population. The World Bank data code for this indicator is SH.TBS.INCD.

HepC2015.dta

This dataset contains information on hepatitis C virus infection in 2015. It has been manually coded from Blach et al. (2017). Country names have been adjusted to be merged to the country_drug.dta dataset. It is used to produce Table 3. The unit of analysis is a country.

- Variable definitions:
  - country: name of the country
  - HepC: Viraemic population (1000s) in 2015. This is obtained from column 2 of Table 1 in Blach, Sarah, et al. "Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study." *The lancet Gastroenterology & hepatology* 2.3 (2017): 161-176.

Part 2 (Proprietary Data)

2.1 Overview

The results presented in the three panels of Figure 2 are obtained using proprietary data which we purchased from IQVIA (www.iqvia.com), a leading provider of data on international sales of pharmaceutical products. Our purchase agreement with IQVIA does not allow us to share the data publicly.

We received the data in May 2020. They contain information on product launches, sale revenues (in local currency converted in US dollars) and sales volume for a subset of products and countries used in our licensing analysis. In Table 1 below we provide the full list of molecules associated with the products in the IQVIA sample. In Table 2 below we provide the full list of countries covered in the IQVIA data. Replication of our results require purchasing from IQVIA sales and launch information for all the products related to the molecules in Table 1 and for the countries listed in Table 2.

The raw data we received were in excel format. We used these raw data to construct analysis variables related to entry by generics, yearly sale revenues and volumes. The next subsection provides the details on how each variable was constructed.

Table 1. Molecules in IQVIA data	
Abacavir	Lamivudine
Abacavir/Dolutegravir/Lamivudine	Lamivudine/Nevirapine/Zidovudine
Abacavir/Lamivudine	Lamivudine/Raltegravir
Abacavir/Lamivudine/Zidovudine	Levofloxacin
Amikacin	Linezolid
Atazanavir	Lopinavir/Ritonavir
Atazanavir/Cobicistat	Maraviroc
Atazanavir/Ritonavir	Moxifloxacin
Bedaquiline	Nevirapine
Bedaquiline/Pretomanid/Linezolid	Nilotinib
Bendamustine	Ombitasvir/Paritaprevir/Ritonavir
Bevacizumab	Oseltamivir
Bictegravir/Tenofovir	Raltegravir
Cabotegravir	Ravidasvir
Cobicistat	Ribavirin
Cobicistat/Elvitegravir/Emtricitabine/Tenofovir	Rilpivirine
Daclatasvir	Ritonavir
Darunavir	Rituximab
Darunavir/Cobicistat	Saquinavir
Darunavir/Emtricitabine/Cobicistat/Tenofovir	Simeprevir
Dasabuvir	Sofosbovir/Velpatasvir
Dasabuvir+Ombitasvir/Paritaprevir/Ritonavir	Sofosbuvir
Dasabuvir/Ombitasvir/Paritaprevir/Ritonavir	Sofosbuvir+Daclatasvir
Dasatinib	Sofosbuvir/Velpatasvir
Delamanid	Sofosbuvir/Velpatasvir/Voxilaprevir
Didanosine	Sofosbuvir/ledipasvir
Dolutegravir	Stavudine
Dolutegravir/Lamivudine	Sutezolid
Dolutegravir/Rilpivirine	Tenofovir
Doravirine	Tenofovir/Emtricitabine
Doravirine/Lamivudine/Tenofovir	Tenofovir/Emtricitabine/Efavirenz
Efavirenz	Tenofovir/Emtricitabine/Elvitegravir/Cobicistat

Elvitegravir	Tenofovir/Emtricitabine/Rilpivirine
Emtricitabine	Tenofovir/Lamivudine
Emtricitabine/Rilpivirine/Tenofovir	Tenofovir/Lamivudine+Nevirapine
Entecavir	Tenofovir/Lamivudine/Dolutegravir
Etravirine	Tenofovir/Lamivudine/Efavirenz
Fosamprenavir	Trastuzumab
Glecaprevir/Pibrentasvir	Valganciclovir
Imatinib	Zidovudine
Isoniazid/Rifapentine	Zidovudine/Lamivudine

Table 2. List of countries covered by IQVIA for markets	
Algeria	Lebanon
Argentina	Mexico
Bangladesh	Morocco
Belarus	Pakistan
Bosnia	Peru
Brazil	Philippines
Chile	Russia
China	S. Africa
Colombia	Saudi Arabia
Dominican Republic	Sri Lanka
Ecuador	Tunisia
Egypt	Turkey
French West Africa	UAE
India	Uruguay
Indonesia	Venezuela
Jordan	Vietnam
Kazakhstan	

2.2 Dataset description and variable construction

The STATA dataset used to perform our analysis is called **IQVIA_country_drugs.dta**. The code to generate the three graphs in Figure 2 is provided in the file **IQVIA_figures.do**

IQVIA_country_drugs.dta

This dataset contains market outcome information. It is used to produce Figure 2. The unit of analysis is a "country-product-year."

- Variable definitions:

- country: name of the country
- product: name of the product
- year: calendar year. Product-country combinations are included if at least one patent is active (non-expired) as in the licensing dataset.
- MPP: dummy =1 if product-country is included in an MPP license in the year
- one_gen: dummy = 1 if at least one generic producer has launched the product in the country by year *t*.
- logSales: logarithm of (1 +) the product-specific sales revenue generated in the country in year *t* by all firms selling the product.
- zeroSales: dummy=1 if sales revenue for the product in the country, in year t is equal to zero.
- logVolume: logarithm of (1+) total units of product sold in the country in year *t*.

The variables one_gen, logSales, and logVolume are manually constructed from the raw IQVIA data using the following approach.

one_gen:dummy equal to one if the IQVIA data indicate that at least one generic manufacturer has entered the market for the product by year t. Generic manufacturers are identified using the "originator/generic" tag provided by IQVIA. The entry date is reported in the IQVIA data in the variable "launch date".

logSales: this measure is constructed using the variable "sales LC USD" in the IQVIA raw data. It contains the total sales for the product in each country-year. The sales measure is identical to the variable in the IQVIA data in the case in which only one firm sells the product in the country. It is equal to the sum of the sales in the case in which multiple firms sell the focal product in the country in year t. In the empirical analysis we use the natural logarithm of the variable (adding 1 to all observations) and include a dummy for observations in which there are no sales (zeroSales).

logVolume: this measure is constructed using the variable "volume (units)" in the IQVIA raw data. We also exploit the information on the number of vials or tablets contained in each pack, which is reported in the variable "pack" of the IQVIA data. Total volume is computed multiplying

the number of packets sold by the number of units in each packet. In the case in which multiple firms sold the same product in the country during the focal year, we sum the volume at the product level, as in the case of sales revenue. In the empirical analysis we use the natural logarithm of the variable (adding 1 to all observations) and include a dummy for observations in which there are no sales (zeroSales).