FDA Press Releases and R&D in Pharmaceuticals

By Afraz Arif Khan

This paper attempts to estimate how dangerous drugs approved by the FDA have a significant effect on the R&D of pharmaceutical companies. It also attempts to investigate how lobbying power with some pharmaceuticals favours their respective R&Ds through Priority Review of approved drugs. We eventually discover that the textual data in these reports has no significant bearing on R&D in these firms by making use of the standard DID estimator with a localised shock in 2017 using a quantile scheme for labelling the DID data.

I. Introduction

Most drugs that are approved by the FDA show significant movement in the stock price of pharmaceutical companies. A lot of research has been dedicated to finding out the inherent link between the two but there is no significant research on how drug approval reports by the FDA affects the performance of the firm. One key performance indicator is the R&D of these pharmaceuticals which will serve as a tool for measurement of this performance for the rest of this report.

A novel technique in empirical analysis is textual analysis which has been used by researchers for empirical corporate finance applications since 2010 but despite its established presence, has not dominated many sectors of research. Textual analysis can be factored into **(1) Lexicon Analysis** and **(2) Sentiment analysis**, of which the latter is a more Machine Learning focused approach. For this paper, we employed the lexicon analysis approach which mostly used word frequency in reports as indicators for how dangerous certain drugs can be. This paper takes inspiration from multiple branches of literature. There are 2 main branches of literature that this paper directly contributes to. The first one is how textual analysis is used in empirical corporate finance applications. One paper (Bellstam, Bhagat and Cookson 2017) made use of textual analysis and webcrawling techniques to measure the innovation of firms and found a positive correlation of it with total assets. This technique of webcrawling employs the LDA (Latent Dirichlet Allocation) corpus method and a word rate of $\Delta_{\text{sentiment}}$ where sentiment is determined by positive/negative words loaded into a dictionary. This approach directly relates to the measure of *dangerous* in FDA reports but word rates are not necessary for a DID approach as measures of how 'dangerous' some drugs are can be done by using indicators. One common use of textual analysis in empirical studies is in patents and in particular innovation in patent reports (Tseng, Lin and Lin 2007). Textual analysis is also used in financial reporting (Lang and Stice-Lawrence 2015), according to this paper textual analysis is correlated with firm liquidity, analyst following and mutual fund ownership.

The second branch of literature is how drug approvals at the FDA affect the performance of the firm. One paper investigated how R&D of pharmaceuticals is proportional to their firm size and how this determines approval success rates and economic returns (DiMasi 2014). They found that lower sized firms that develop small molecule drugs have high approval rates. Other papers like (Krieger, Li and Papanikolaou 2018) investigate drug molecular composition and use it to determine how it affects investment in pharmaceuticals particularly with varying degrees of risk aversion amongst investors. While this paper is more inherently focused on the qualitative effects of the side effects of drugs and the performance of the firm, the idea to use features of a drug to investigate financial repercussions remains very much intact for this paper. There was also a study conducted on clinical transparency in the FDA related to the size of the firm (Miller et al. 2017). This paper makes use of large pharmaceuticals and determines that clinical transparency is high when patients are tested. this in turn served as inspiration for the use of Firm Size as the control variable for this paper as Firm Size has large explanatory power amongst pharmaceuticals.

The third branch of literature is the relative safety of drugs that are approved by the FDA. There is perhaps no way to quantitatively measure the "safety" of drugs but recent advances in medical journals coupled with historical reviews all universally agree that there are life threatening and potentially fatal side effects of some medicines that could assign them a "dangerous" label. One paper (Alshammari 2016) in particular examines a historical review of how certain medicines in the United States had their safeties revised because of their prolonged use and (eventual) fatalities, like sulfanilamide. In fact the FDA itself (FDA 2020*b*) issues press releases every year based on further testing where they enforce the withdrawal of certain drugs from the market. Despite this scrutiny, there are still some drugs (like cancer treatments) that are approved by the FDA and are still dangerous to use (as revealed by the webcrawler later in the paper). As a result this paper also serves as an empirical tool to highlight those drugs that were approved despite their unequivocal "dangerous levels" and are currently open for use in the market.

The rest of this paper is organised as follows. We commence with a discussion on the institutional background of the FDA and the drug approval process followed by a discussion on the data collection methodology and the chosen independent and dependent variables for the proposed identification strategy. Later on we highlight the identification strategy which is a standard DID estimator with artificially constructed (through the quantile approach) treated and control groups including an exogenous shock in 2017. We then discuss the results of the regression and finally we conclude. The appendix has more information on descriptive statistics, regression output and a short section on the webcrawler to enable its use for further research applications.

II. Institutional Background & Data

A. Institutional Background

The U.S Food and Drug Administration is perhaps the most well known authority behind the approval of new and innovative medicines that are used to treat all types of diseases/conditions. Pharmaceutical companies present in the U.S operate by developing a drug internally through various trials and stages and then submitting it to the FDA for approval. If a drug is granted FDA approval then it can be marketed for use by the general public. This approval process in the FDA has many stages (FDA 2020*a*). First, the FDA runs animal trials and then runs 3 phases of human trials. After testing these drugs, companies then file an NDA (New Drug Approval) which is a formal process of asking the FDA to approve the drug for marketing in the United States. If companies reach this stage of the process then after the 60 day review period of the NDA, the FDA dispatches analysts to review the drug development process and inspect the labels of these drugs along with their respective manufacturing facilities. After this process the FDA formally approves the drug and issues a press release a few days after the formal approval detailing useful information on the drug (like treatment, dosage, side effects etc.). It is common knowledge that pharmaceutical companies are well known for spending a lot of money on R&D as the development of new and innovative drugs is a core part of their business. This in turn makes the R&D of these pharmaceuticals a suitable metric for the performance of the firm based on drug approval (or potentially information on the drug's approval).

B. Data

There are 2 sources of data for this experiment. We make use of (**1**) FDA Press Release data for textual analysis and (**2**) R&D data of pharmaceutical companies. The FDA press release data is available on the main page of the FDA website including archives up till 2013. R&D data is available on Compustat.

1) **FDA Press Release Data**: Textual analysis techniques using Python were utilised to extract relevant information from each FDA press release. This textual data was accessed and analysed using an algorithm that extracted important information such as the date of the press release, the approved pharmaceutical's name and the side effects of the approved drug. The FDA data is not available before 2013 as most web archive links are unresponsive or decommissioned so the full data set is from 2013-2020. There are 2 main variables of interest:

    a) *Dangerous*: This is an indicator variable that is used to measure the severity of side effects in many drugs. Universally, the worst side effect that any drug can have is fatality/death. Another serious side effect is "suicidal" which is attributed to chemical imbalance in the

brain. By using webscraping techniques the paragraph discussing side effects of each press release is captured and textual analysis determines if the word "death" or "suicidal" occurred in the localised side effects paragraph. This data is robust because the FDA web pages are partitioned in such a way that it is possible to extract those paragraphs that obey a hierarchical structure where each paragraph deals with a specific topic (side effects, dosage, approval to a company, priority review, comparative analysis).

b) *Priority Review*: As discussed in the institutional background section some firms opt for accelerated approval or priority review. This is usually due to the severity of some illnesses and the need for a drug to treat them in the market. Companies apply for this type of approval and are dubbed as "priority review" or "accelerated approval" companies so these drugs skip the measurement of drug effectiveness and receive NDA approval straight away. This is an interesting variable to investigate as these drugs have not been totally vetted and could be subject to severe problems. So the *priority review* variable represents an indicator which finds the word "priority review" or "accelerated approval" in each drug approved press release.

2) **R&D Data**: After extracting each pharmaceutical name from the FDA press release the corresponding names were then matched with their associated ticker symbols and extracted from Compustat. The R&D data used in the regression is scaled by total assets to minimise the standard error.

III. Empirical Strategy

The difference in differences (DID) estimator is used for this specification. The full regression is given as:

$$Y_{it} = \beta_0 + \beta_1 d_i p_t + \beta_2 d_i + \beta_3 p_t + \gamma X_{it} + \epsilon_{it}$$

Where:

1) i: Firm

2) t: Year (quarterly)

3) Y_{it}: R&D of firm i at time t

4) d_i: Indicator of "treated" firm

5) p_t: Post Event Indicator

6) X_{it}: Control Variables like:

a) Firm Size

b) Leverage

Here we make use of an exogenous shock in 2017, where FDA analysts report a severe reduction in warning letters after Donald Trump's presidency. Warning letters are a measure of how some dangerous drugs are approved by the FDA. This is in turn related to the appointment of Scott Gottlieb (FDA 2020*c*) who in turn pushed for expedited approvals under Trump's administration. This regression specification was used on 2 different hypotheses:

1) **Hypothesis 1**: Do dangerous drugs have any effects on the R&D of pharmaceutical companies?

2) **Hypothesis 2**: Do drugs that receive priority review/accelerated approval have any effects on the R&D of pharmaceutical companies?

To test both these hypotheses the coefficient of interest is β_1 in both regressions. The main deviation from a standard DID approach is the quasi-nature of the "Treated" group. Instead of having a suitable "control" group this regression can be performed by a quintile approach. A common technique in quintile approaches is splitting the "dangerous" data into 2 groups namely *treated: dangerous* > 0 and *untreated: dangerous* $= 0$

IV. Results

From our analysis of table B1 we can see that almost half the firms have drugs that opt for priority review and most drugs are not as dangerous. Recall that a dangerous score of "1" means either "death" or "suicidal" occurred in the side effects paragraph of the FDA report and a score of "2" means both of them occurred which makes the drug very dangerous. The other descriptive statistics are quite standard and are in line with what we expect. From our analysis in table C1 and table C2 we find no significant results following the exogenous shock in 2017. A few reasons explain these results:

1) The dataset does not include values before 2013 and we only have 303 observations. A larger dataset would solidify either the insignificance of the results or could help explain more on the (weak) statistical significance of the treated group in table C1. It is interesting to note that there is a (weak) statistical significance behind the dangerous levels in these drugs with the R&D but even if we were to accept these results, they do not show a large change in the R&D of the firm (about 0.5% increase).

2) The main reason behind there being no significant change in the R&D of these pharmaceuticals lies ultimately with how the R&D budget is allocated ex-ante at the beginning of each financial year. This allocation usually represents a long term plan that the company makes in the development of

a new drug. The outcome of a report for a single drug intuitively does not guide R&D decisions of a pharmaceutical due to the fact there are multiple drugs in the pipeline, the complexity of the pipeline and the inherent timing of allocating this budget in contrast to the release of an FDA report.

3) It is also clear that drugs that received priority review were not affected by the 2017 shock. The priority review of the drug usually happens ex post so there is no significant way to tell how it affected the R&D ex-ante. An alternative empirical strategy would involve aggregating these priority reviews and utilising a time series to see how this shock would affect predicted priority reviews in the future. It would also suffice to examine cross-sectional data as some pharmaceuticals manufacture drugs that inherently deal with life threatening issues (that do not have an immediate cure) and therefore file for priority review. Unfortunately, the data on priority reviews is confined to the 2013-2020 sample and an efficient predictor for seeing how this modifies R&D would require more yearly data due to the reasons listed in point 2.

V. Conclusion

In this paper we were able to construct an efficient textual analysis tool on FDA reports from press release data. We created a webcrawler that efficiently circumvents the web content on the FDA by outputting values of interest (such as side effects, priority review) with the potential of capturing more data with slight modifications. We were able to measure how these reports impacted the performance of pharmaceuticals and based on our analysis we found no significant changes in the R&D of these firms. We therefore conclude that either a presence of more data or drug approvals alone would solidify our findings and we leave this open to future work.

REFERENCES

Alshammari, Thamir M. 2016. "Drug safety: the concept, inception and its importance in patients' health." *Saudi Pharmaceutical Journal*, 24(4): 405–412.

Bellstam, Gustaf, Sanjai Bhagat, and J Anthony Cookson. 2017. "A text-based analysis of corporate innovation." Working paper.

DiMasi, Joseph A. 2014. "Pharmaceutical R&D performance by firm size: Approval success rates and economic returns." *American journal of therapeutics*, 21(1): 26–34.

FDA. 2020*a*. "FDA Approval Process." https://www.fda.gov/drugs/drug-information-consumers/fdas-drug-review-process-continued.

FDA. 2020*b*. "Safety of Drugs." https://www.fda.gov/drugs/drug-safety-and-availability/drug-safety-communications.

FDA. 2020*c*. "Scott Gottlieb." *https://www.fda.gov/about-fda/fda-leadership-1907-today/scott-gottlieb*.

Krieger, Joshua L, Danielle Li, and Dimitris Papanikolaou. 2018. "Missing novelty in drug development." National Bureau of Economic Research.

Lang, Mark, and Lorien Stice-Lawrence. 2015. "Textual analysis and international financial reporting: Large sample evidence." *Journal of Accounting and Economics*, 60(2-3): 110–135.

Miller, Jennifer E, Marc Wilenzick, Nolan Ritcey, Joseph S Ross, and Michelle M Mello. 2017. "Measuring clinical trial transparency: an empirical analysis of newly approved drugs and large pharmaceutical companies." *BMJ open*, 7(12).

Tseng, Yuen-Hsien, Chi-Jen Lin, and Yu-I Lin. 2007. "Text mining techniques for patent analysis." *Information processing & management*, 43(5): 1216–1247.

Webcrawler

The webcrawler is free to use and open source. It is available for download on this link. The webcrawler itself makes use of Beautiful Soup in Python, a common tool for webscraping documents on the internet. The FDA press releases can be found here. They are available in archive format from 2013-2020, other websites do have links prior to this date range but most of the data is fragmented and incomplete.

This webcrawler can be used with slight modifications to generate more interesting data in these FDA reports and can even be modified to find more information on FDA press releases on animals, food and criminal investigations. For more modifications on the functional performance of the webcrawler kindly consult the GitHub page which has more information on the functional overlay of the program.

Descriptive Statistics

Table B1—Descriptive Statistics

Stat	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
scaled R&D	303	0.012	0.013	0	0	0.03	0
Size	303	3.557	1.600	1.816	1.816	5.181	5.181
leverage	303	2.002	2.814	−12.417	0.582	3.146	13.633
Dangerous	303	0.244	0.495	0	0	0	2
Priority	303	0.469	0.500	0	0	1	1

Tables

Table C1—Dangerous

	Dependent variable:
	Y
d	0.006^{*}
	(0.003)
p	0.002
	(0.002)
dp	-0.003
	(0.004)
X	-0.001^{***}
	(0.0003)
Constant	0.015^{***}
	(0.002)
Observations	303
R^2	0.064
Adjusted R^2	0.051
Residual Std. Error	0.012 (df = 298)
F Statistic	5.064^{***} (df = 4; 298)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table C2—Priority Review

	Dependent variable:
	Y
d	0.001
	(0.002)
p	0.002
	(0.002)
dp	−0.001
	(0.003)
X	−0.001***
	(0.0003)
Constant	0.016***
	(0.002)
Observations	303
R^2	0.051
Adjusted R^2	0.038
Residual Std. Error	0.012 (df = 298)
F Statistic	3.970*** (df = 4; 298)
Note:	*p<0.1; **p<0.05; ***p<0.01