Drug Development and the Expansion of Bioinformatic Software

Christopher Long

May 7, 2023

Abstract

In this paper we investigate whether PyMol had a preferential treatment on the drug development process of biologics over small-molecule drugs using a difference of differences regression with dummy variables for the drug type as well as the time periods before and after the introduction of PyMol. The measures for the drug development process that were examined were the time for phase trials to complete, the completion rate of those phase trials, the number of phase trials, and the number of BLAs or NDAs. Overall it seems that the introduction of PyMol did not have a preferential effect on biologics over small-molecule drugs, either equivalent or no effect on both types.

Acknowledgments

I would like to thank Professor Stein for giving me the opportunity to research this question and for the advice she has given throughout. I would also like to thank my family and my friends for everything.

Contents

I. Introduction (4)

- Biological And Small Molecule Drugs (4)
- What Is Pymol (5)
- What Is AutoDock (5)
- The Drug Approval Process (6)

II. Literature Review (8)

- Previous Studies Involving Drug Production (8)

III. Data and Methodology (10)

- Data Sources (10)
- Differences of Differences (10)
- Expected Time To Completion Based on a Conditional (11)
- Hypothesis (12)

IV. Results (13)

- PyMol Influence on NDA and BLA (13)
- PyMol Influence on Conditional Completion Time for Phase Trials (13)
- PyMol Influence on Completion Rate of Clinical Trials (14)
- PyMol Influence on Number of Trials (14)
- Pymol Influence on Growth Rate of Number of Trials (15)
- Conclusion (16)
- Figures (18)
- Bibliography (26)

I. Introduction

Biological and Small Molecule Drugs:

Pharmaceutical drugs are generally broken up into two classifications: biologics and small-molecule drugs. Biologics are characterized as being produced and isolated from a model organism and consist of larger molecules consisting of larger proteins, RNA, or DNA that can be used to elicit a particular desired physiological effect. In contrast, small molecule drugs are produced chemosynthetically; they are generally produced from a series of chemical reactions in a lab that usually have intermediate isolation and purification steps(Favour, Danladi, Makurvet, 2021). Small molecule drugs typically function as ligands that bind to target proteins to inhibit the effects of disease and as a result, are generally smaller than Biologics. In general, the molecular weight of small molecule drugs varies between 0.1 and 1 kDa, while biologics tend to be more than 1 kDa(Favour, Danladi, Makurvet, 2021). Another difference between small-molecule drugs and biologics is the nature of storage. Since biologics are composed of larger proteins or genetic material, they are more sensitive to factors like temperature, light, PH, or salinity which can result in their denaturation and a loss in their efficacy(Favour, Danladi, Makurvet, 2021). Small-molecule drugs are smaller and are usually constrained to a particular conformation or shape and are generally less sensitive to these effects; as a consequence, small-molecule drugs are on average easier to store and transport than biologics as they are more tolerant to changes in their environment(Favour, Danladi, Makurvet, 2021). As a result of biologicals being more complex, having lower tolerances to environmental effects, and their origins in cellular organisms, they are more specific in their target sites than small molecule drugs. Biologics typically need to be administered intravenously rather than orally and

oftentimes stimulate the immune system. As a result of many of these requirements, small-molecule drugs have been preferentially produced over biologicals.

What Is PyMOL:

In 2000, PyMOL was created by Warren Lyford DeLano under DeLano Scientific LLC as an open-source visualization software(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). In 2010 the rights to PyMol were sold to Schrodinger Inc., which is the present owner of PyMol(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). This software allows scientists to view and interact with the three-dimensional structure of biological molecules such as Protein, DNA, and RNA in a myriad of ways. Some of the unique visualization techniques PyMOL possesses is the ability to make movies of these interactions, enabling scientists to get a better understanding of the interaction in real-time(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). Furthermore, PyMol allows viewership of the aforementioned macromolecules in multiple ways that estimate their shape and can help determine functional domains on the macromolecule. Some examples of these different views are using ribbons and chords to illustrate different domains on the macromolecule or viewing the macromolecule in its more ‘true’ globular shape(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). Also, the software allows scientists to highlight particular domains on the macromolecule, enabling scientists to easily differentiate different parts of the macromolecule(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). The software works by taking compiled information about the crystallographic structures of known proteins, as well as the known primary sequences of these molecules to construct a 3d model of the protein(Yuan, S., Chan, H.C.S. and Hu, Z. 2017). By looking at the three-dimensional structure and using a variety of plugins associated with PyMOL, scientists can look for structural motifs and gain insight into the functions of particular regions. Plugins used with PyMOL such as Autodock Vina, SLIDE, or Amber enable scientists to get a better

understanding of the binding affinity of a potential drug as well as the functional relationships between proteins and disease(Lill MA, Danielson ML, 2011). As a result, PyMOL directly allows companies to test and rule out potential drug candidates before testing them in a lab.

What Is AutoDock:

AutoDock Vina is a molecular modeling software published in 1989 that calculates the energetics of molecular docking. AutoDock Vina is essentially a software that uses the principles of thermodynamics to predict the favorability of binding between two molecules, specifically between proteins and some ligands. The software pulls from a library like ZINC, KEGG, or PDB information about the protein or ligand of interest, and then proceeds to calculate the favorability of the interaction between the two structures(Lill MA, Danielson ML, 2011). This software considers the enthalpic and entropic interactions that occur when a molecule interacts with a structure of interest. The software works by running through various conformations of the structure of interest and running through various simulations of the interactions that are expected to occur between the two structures(Lill MA, Danielson ML, 2011). The software then calculates the expected value of the binding energy for these various environments and conformations(Lill MA, Danielson ML, 2011).

The Drug Approval Process:

Drug approval has been broken up by the FDA into 5 distinct steps: Discovery and Development, Preclinical Research, Clinical Research, FDA Drug Review, and FDA Post-market Drug Safety Monitoring(Commissioner, Office of the). Discovery and Development are characterized rather loosely as the phenomena in which researchers pinpoint a compound as having the potential for a desired effect and elect to pursue an investigation into its efficacy for its action as a drug. A new compound is usually selected as the result of some new understanding

of the enzymatic mechanisms that underlie a particular disease. Preclinical research is essentially when researchers, adhering to GLP regulations, gather evidence surrounding the metabolic effects of the candidate drug, evaluate whether the candidate is carrying out the desired effect, and gain a better understanding of the risks the candidate poses to individuals(Commissioner, Office of the). In this stage, candidates are evaluated in In-Vitro or In-Vivo using a model non-human organism(Commissioner, Office of the). The next stage is clinical research, which is when the candidate, after being evaluated for its efficacy and safety, is tested in humans with the underlying disease. However, before clinical trials for a candidate can proceed, an IND (Investigational New Drug Application) must be submitted to the FDA which contains results from animals studies conducted in the preclinical stage, the relative toxicity of the drug, information on how the drug was made, information about the investigators, and protocols for the planned clinical trials. Once this is approved by the FDA, clinical trials can proceed for the candidate(Center for Drug Evaluation and Research). Phase 1 Trials are relatively small and simply evaluate the safety and dosage of the candidate drug(Commissioner, Office of the). Phase 2 Trials are increasingly large containing a few hundred people and are primarily to establish the efficacy of the drug and gain insight into the side effects of the drug(Commissioner, Office of the). Phase 3 Trials range from a few hundred to a few thousand people and establish a better basis for the efficacy of the drug and the drug's potential acute and chronic side effects(Commissioner, Office of the). After clinical trials have been conducted, the next phase is FDA Drug Review(Commissioner, Office of the). During this stage, researchers must submit an NDA (New Drug Application) or a BLA (Biologic License Application) to the FDA, provided the candidate is a Small Molecule Drug or a Biologic respectively(Commissioner, Office of the). These applications include all of the data collected on the candidate from the pre-clinical trials to

the Phase 3 clinical trials(Commissioner, Office of the). Once these applications are approved, the drug is allowed to go to market. Once the drug is on the market, the final stage is FDA Post-Market Drug Safety Monitoring(Commissioner, Office of the). This stage is also sometimes referred to as a Phase 4 clinical trial(Commissioner, Office of the). During this final stage, a drug is essentially under surveillance for any potential health hazards that were not discovered during clinical trials. This surveillance goes on for many years and looks into side effects that take a long time to manifest(Commissioner, Office of the). If a drug is flagged during this time as a potential health hazard, the FDA can pull it from the market(Commissioner, Office of the).

II. Literature Review

Previous Studies Involving Drug Production:

Most economic studies looking at drug development have been to estimate the underlying costs of production and the probability of a drug candidate succeeding. The most notable studies concerning this were a series of studies conducted by DiMassi. DiMassi has published several papers estimating the various components contributing to the cost of production of pharmaceutical drugs. DiMassi concluded that the largest contribution to the cost of production was the opportunity cost of capital(DiMassi et al. 2003). Meaning, the largest cost of production is when capital is tied up in projects for an extended period pursuing potential products that may or may not lead to a deliverable drug. It was estimated that the total cost of production for a drug was around 802 million in 2000 dollars(DiMasi et al. 2003). Another study similarly tried to estimate the cost of development and obtained a similar value(Adams CP, Brantner VV. 2006). It was also estimated that the time between the start of clinical trials to the time it took for a drug to get approval was 90.3 months or around 7.5 years(DiMasi et al. 2003). It was also estimated that

the probability of entering Phase 1, 2, and 3 trials was 100%, 71%, and 31% respectively(DiMasi et al. 2003). In a later paper, it was estimated again that the probabilities of entering a Phase 2 trial, entering a Phase 3 trial, applying for an NDA/BLA, and receiving an NDA/BLA were 59.52%, 35.52%, 61.95%, and 90.35% respectively(DiMasi et al. 2016).

Instead of examining the potential costs of output, our focus will be on how the development process of particular drugs has changed over time in response to the introduction of new development tools. The Particular tool of interest for our study is PyMol and the aspects of development that are of interest are whether the number of clinical trials has increased, whether the time for these trials to complete has changed, and whether the number of drugs getting approval has changed as well.

Most papers that have been published about this question have been merely speculative and hypothetical. The consensus among scholars has been relatively divided across all of the core questions. The first source of division is whether or not the increased accessibility of software tools of this nature will have any effect at all on the costs of the production of drugs in general. Some people believe that the effect of such software will make the costs of production cheaper, while others believe that the production procedures that have been established for discovering and creating drugs have been too established to change significantly with the increase of these tools. The second source of division follows from the first class of individuals that do believe there is an effect of software on the cost of drug production and its frequency. This is the nature of whether or not there is a significant difference between how this software has affected small molecule drugs versus biologics.

III. Data and Methodology

Data Sources:

Data for this project was collected from three sources, Clinicaltrials.gov, the CDER datasheet provided by the FDA, and Pubmed. From clinicaltrials.gov, data about Phase 1, 2, and 3 trials between 1985 and 2021 was collected. This data contained the start dates of the trials, the end dates of the trials, the names of the drugs used in the trials, the status of the trial (as either being completed, suspended, or terminated), and the classification of the drugs used within that study as either Biologicals or Small molecule drugs.

From the FDA website, the CDER data sheet was used to collect the number of drugs that received approval for use between 1985 to 2021 (https://www.fda.gov/media/135307/download). The CDER datasheet reported whether these drugs were under a Biologic License Application (BLA), meaning the drug was a biological, or New Drug Application (NDA), meaning the drug was a small molecule drug. Finally, the Pubmed search engine was used to collect the number of publications that were published in a given year that contained a query word in it. The words that were queried using this engine were Pymol and Autodock respectively. This was used as a measure of how much the particular software was being used and talked about in the scientific community as well as within clinical trials themselves

Differences of Differences:

The Differences of Differences regression is used to determine whether there is an effect of treatment between a control group and a treatment group over time. Under this study, the Small molecule drugs were viewed as a control group for the Pymol software. This is because, since Pymol is used for seeing ligand binding and visualization of complex biological molecules like proteins or nucleic acid structures, it would be expected to have a larger effect on the

production of biologics than Small Molecule Drugs. The differences of differences regression take four forms that will be described below.

The Small Molecule Drug before the appearance of Pymol is represented by:

$$Yi = \beta 0$$

The Biologic before the appearance of Pymol is represented by:

$$Yi = \beta 0 + \beta 1(Biologic)$$

The Small Molecule Drug after the appearance of Pymol is represented by:

$$Yi = \beta 0 + \beta 2(Pymol)$$

The Biologic before the appearance of Pymol is represented by:

$$Yi = \beta 0 + \beta 1(Biologic) + \beta 2(Pymol) + \beta 3(Biologic * Pymol)$$

The value of β3 is the difference between the effect Pymol had on the number of Biologics versus the number of Small Molecule Drugs approved by the FDA. In order to perform this regression, the number of biologics and small-molecule Drugs approved by the FDA were organized by year. Two dummy variables were constructed both obtaining either a value of 0 or 1. The first variable was "Biologic" and it was assigned 1 if the number of drugs approved were biologics and 0 if the number of drugs approved that year were Small Molecule Drugs. The second was a dummy variable called "Pymol" and it took a value of 0 for all of the years before Pymol was released and a value of 1 for all of the years following its release.

Expected Time to Completion:

When determining the expected time to completion for a phase trial, it is necessary to truncate and condition the data being used. In order to evaluate changes in the time it takes for a

phase trial to be completed over time across Biologicals and Small molecule drugs, it was necessary to take the expected value of the time it took for trials that were completed within a given period. For this, we used 10 years, up until the start date of trials that occurred of the same length of that period, 10 years, before our most recent data on completed trials. This is necessary because there is a selection bias that occurs as we get closer to the most recent dates. For instance, the population of trials that started 10 years ago and were completed consists of trials that could end between 0 and ten years, and the average value of those trials results from this range. However, of the trials that started 2 years ago and were completed, these trials can only be a maximum of two years to complete because the other trials taking more than 2 years to complete that started the same year will have not finished yet. By selecting only a particular expected value for trials based on a condition, we can control the bias in the population of studies we are examining.

Hypothesis:

For the following study, the small molecule drugs were set as the control group. This is because Pymol should not have as much of an effect on the production frequency of small molecule drugs as it relates specifically to the visualization of larger molecules like proteins or long chains of nucleic acids. Likewise, we made our treatment group to be the class of biological drugs for the converse reasoning. We hypothesize that the software will have a larger impact on the production of biological drugs because the software allows producers to better understand the structure of macromolecules, ascertain their function, and identify which compounds can elicit an effect on them.

IV. Results

PyMol Influence on NDA and BLA

The results of the differences of differences regression, Figure D, suggest the introduction of PyMol may have had a preferential increase in the number of BLAs over the number of NDAs since the coefficient of the interaction term between the PyMol and Biologic dummy variables is positive and statistically significant. The value for the coefficient on the PyMol*Biologic interaction term has a value of 7.0061 with a standard deviation of 3.079. Due to the coinciding expansion of the usage of the PyMol software at this time, shown in Figure B, and the preferential increase in the number of BLAs after PyMol's release, there seems to be some evidence that the PyMol software played a role in this increase.

PyMol Influence on Conditional Completion Time for Phase Trials:

The results of the regression, Figure D, suggest the introduction of PyMol may have resulted in a reduction in the amount of time it takes for a trial for a drug or biological to complete. However, it does not indicate that the introduction of this software had a preferential effect on Biologicals over Small Molecule Drugs. The coefficient on the PyMol*Biologic interaction term for Phase 1, 2, and 3 trials was -0.6803, -0.2472, and 0.4227 with standard deviations of 0.969, 0.841, and 0.621 respectively. This result suggests that there was no preferential effect on the difference in time it takes for phase trials to complete whether they study small-molecule drugs or biologics. The coefficient on simply the PyMol variable for the Phase 1, 2, and 3 trials are -1.7086, -1.5689, and -1.4330 with standard deviations of 0.638, 0.444, and 0.355. These results combined with the expansion of the PyMol software shown in Figure B suggest that the time for trials to complete may have been reduced in part as a result of the PyMol software. However, there are many other factors that coincide with the introduction of

the PyMol software that could coincide with this development such as increased computing power or better laboratory tools.

PyMol Influence on Completion Rate of Clinical Trials:

The results of the regression as shown in Figure D indicate there was no preferential effect of the introduction of PyMol on the completion rate of phase trials for biologics over small-molecule drugs. The coefficients on the PyMol*Biologic interaction term for the completion rate of Phase 1, 2, and 3 trials were -0.0086, 0.0221, and 0.0308 with standard deviations of 0.029, 0.032, and 0.041 respectively. None of the coefficients on the PyMol*Biologic interaction term are statistically significant, indicating there was no preferential effect on the completion rate of trials on biologics. One unique suggestion of the regression is the completion rate for clinical trials actually decreased after the introduction of PyMol. The coefficients on the PyMol variable for the completion rate of Phase 1, 2, and 3 trials were -0.0711, -0.0648, and -0.0585 with standard deviations of 0.013, 0.022, and 0.024. These results on the surface indicate that the introduction of PyMol reduced the completion rate of phase trials. It does not make much logical sense for the completion rate to go down as a result of a beneficial tool being introduced. It is more likely that there are unaccounted variables associated with this time frame that resulted in this trend. The PyMol variable is a dummy variable that takes a value of 1 for all of the years after 2000, so there are likely other factors that coincide with the post-2000 time period that may result in fewer phase trials being completed. Examples of these possible effectors are legislative changes or an increased willingness to take on risky projects.

PyMol Influence on Number of Trials:

The results of the regression, Figure D, suggest the introduction of PyMol increased the number of drugs and biologicals being studied in clinical trials but that there was no preferential

effect on the number of biologics being studied over the number of small-molecule drugs being studied. The coefficient on the PyMol*Biologic interaction term for the number of Phase 1, 2, and 3 trials are -798.2833, -986.3000, and -692.1833 with standard deviations of 188.518, 146.901, and 88.048 respectively. This indicates that after the treatment effect of PyMol on the number of biologics being studied was actually negative. This means that small-molecule drugs were being studied more than biologics after the introduction of PyMol. The coefficient on the PyMol variable for the number of Phase 1, 2, and 3 trials was 955.0833, 1154.5167, and 788.9500 with standard deviations of 186.33, 145.416, and 86.539 respectively. These results indicate the treatment effect of PyMol on small-molecule drugs and biologics overall was a positive effect, which was consistent with expectations. One thing to note is small-molecule drugs have existed for a longer period of time so the effect of PyMol on the raw number of trials may be masked due to the fact that there are just fewer biologics overall compared to small-molecule drugs since biologics are newer species of drugs.

Pymol Influence on Growth Rate of Number of Trials:

The results of the regression, Figure D, suggest the introduction of PyMol had no preferential effect on the growth rate of the number of biologics being studied or the growth rate of the number of drugs being studied. The coefficient on the PyMol*Biologic term for the growth rate of Phase 1, 2, and 3 trials are -0.5064, 0.1115, and -0.1991 with standard deviations of 0.380, 0.590, and 0.350 respectively. None of the coefficients on the PyMol*Biologic term were statistically significant, indicating the treatment of PyMol did not have a preferential effect on the growth rate of studies on biologics. The coefficient on the PyMol term for the growth rate of Phase 1, 2, and 3 trials are 0.0689, -0.5193, and -0.3977 with standard deviations of 0.190, 0.355, and 0.184 respectively. The only coefficient that was statistically significant above the

95% level was the coefficient on the Phase 3 growth rate with a value of -0.3977. Given the other coefficients were not statistically significant, this result is not very well supported. It seems for some reason after 2000 the growth rate of Phase 3 trials has gone down. While this may be the result in part of the introduction of the Pymol software, it seems more likely that some other phenomena are occurring.

Conclusion:

In general, the aforementioned results suggest PyMol did not have a preferential positive effect on the drug development process of biologics over small-molecule drugs. The coefficient on the PyMol*Biologic was statistically insignificant for all of the regressions except the number of drugs approved by the FDA and the number of Phase 1, 2, and 3 trials. While the regression on the number of drugs approved by the FDA indicates a positive treatment effect of PyMol on the approval of biologic drugs, the number of Phase 1,2, and 3 trials indicate a negative treatment effect of PyMol on the number of trials involving biologics, contradicting the hypothesis that PyMol would result in a preference for investigating biological drugs.

While the hypothesis of PyMol having a preferential effect on the development of biologics has little support, there seems to be some indication that the introduction of PyMol had a positive effect on both drug types. However, when examining Figures A and C, the perspective that PyMol was the main treatment responsible for the increase in both drug types should be viewed with some skepticism. Figures A and C illustrate that many of the positive trends in the development of these drug entities pre-date the introduction of PyMol. Furthermore, one would expect that the effects of PyMol should have a delayed effect on drug development rather than acting instantaneously as our regression model assumes. The combination of the positive trends in development and the time dependency for the diffusion of PyMol into the production process

indicates that there is likely not a strong connection between the development process and the PyMol software.

Figures:

A)

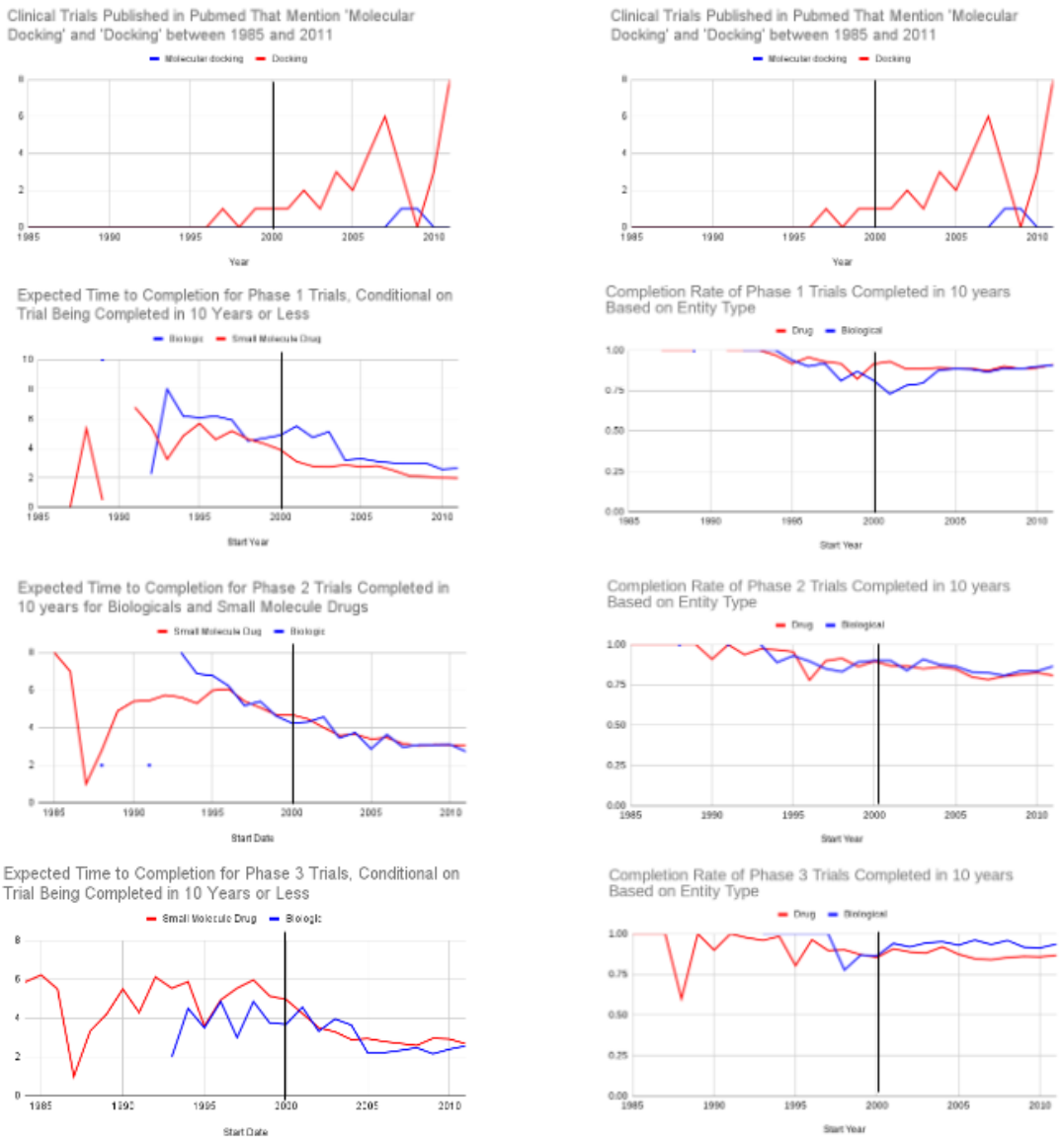

Figure A) The left column of figures displays data from two sources. The graph on the top of the left column labeled “Clinical Trials published in Pubmed That Mention ‘Molecular Docking’ and Docking” displays the number of papers that were recovered from a PubMed search for papers published in each year between 1985 and 2011 that were classified as Clinical Trials and contained the terms “Docking” in red or “Molecular Docking” in blue anywhere within their text. The bottom three graphs in the left column of figures display the expected time to

completion for phase trials for either biologics or small molecule drugs conditional on being completed in 10 years for trials that started between 1985 and 2011. In order to calculate this value clinical trial data was downloaded from clinicaltrials.gov. Trials were then differentiated on whether the “Intervention” was either a “Biological” or a “Drug”. The length of a trial was calculated by subtracting the “Completion Date” from the “Start Date”. Then the average time for completion for trials that started each year between 1985 and 2011 was calculated for each phase(1,2, or 3) trial for either the “Biological” in blue or the “Drug” in red that were completed in 10 years or less. A black line was displayed across all of the figures marking the year 2000, which corresponds to the release date of PyMol.

In the right column the same graph labeled “Clinical Trials published in Pubmed That Mention ‘Molecular Docking’ and Docking” is placed at the top. Beneath this graph is the completion rate of trials between 1985 and 2011 that studied either Biologics or Drugs. Using data collected from clinicaltrials.gov, the completion rate was calculated by separating the trials by whether the “Intervention” was a “Biological” or a “Drug” and based on the years that those trials started. The completion rate was calculated by taking the number of trials for each start date between 1985 and 2011 that were classified as completed and dividing it by the sum of trials that were completed, suspended or terminated that started in the same year. The completion rate for Biologics is shown in blue and the completion rate for small-molecule drugs is shown in red. A black line was displayed across all of the figures marking the year 2000, which corresponds to the release date of PyMol.

B)

Figure B) Displays two graphs. The top graph labeled “Clinical Trials published in Pubmed That Mention ‘Molecular Docking’ and Docking” displays the number of papers that were recovered from a PubMed search for papers published in each year between 1985 and 2021 that were classified as Clinical Trials and contained the terms “Docking” in red or “Molecular Docking” in blue anywhere within their text. The bottom graph was also composed from the search engine on PubMed. The bottom graph is labeled “Number of Clinical Trials Published and Number of Papers Mentioning PyMol per Year on PubMed”. This graph displays the number of papers published in PubMed between 1985 and 2021 that were either classified as “Clinical Trials” in red or contained the word PyMol anywhere in their text in blue.

C)

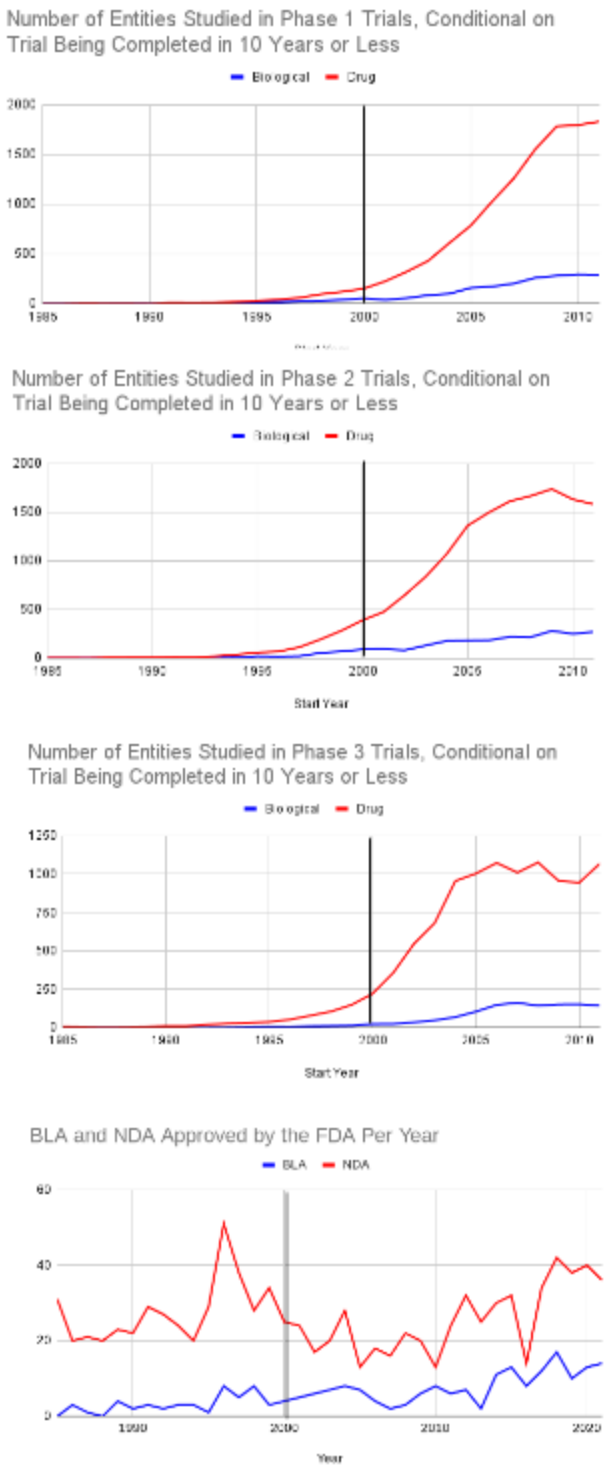

Figure C) The top three graphs display the number of phase trials that started in each year that were completed in 10 years. This data was taken from clinicaltrials.gov. The trials were separated based on whether or not the “Intervention” was a “Biological” or a “Drug”. The length of a trial was calculated by

subtracting the completion date from the start date. Trials that were completed in 10 years were then counted for each year between 1985 and 2011. The trials studying Biologics are shown in blue and the ones studying small molecule drugs are shown in red. The bottom graph displays the number of NDA and BLA done by the FDA between 1985 and 2021. This data was collected from the FDA website (https://www.fda.gov/media/135307/download).

D)

Table 1 Differences of Differences Regression Parameters for Pymol

Dependent Variable	Beta 0	Biologic	Pymol	Biologic*Pymol
Number of Drugs Approved by FDA	27.8000 ** (2.162)	−24.7333 ** (2.250)	−2.2091 (2.886)	7.0061* (3.079)
Duration of Phase 1 Trials Completed in 10 Years	4.3133 ** (0.619)	1.7771 (0.917)	−1.7086 ** (0.638)	−0.6803 (0.969)
Duration of Phase 2 Trials Completed in 10 Years	5.1155 ** (0.414)	0.1799 (0.811)	−1.5689 ** (0.444)	−0.2472 (0.841)
Duration of Phase 3 Trials Completed in 10 Years	4.4594 ** (0.312)	−0.3405 (0.541)	−1.4330 ** (0.355)	0.4227 (0.621)
Completion Rate of Phase 1 Trials Completed in 10 Years	0.9654 ** (0.013)	−0.0320 (0.026)	−0.0711 ** (0.013)	−0.0086 (0.029)
Completion Rate of Phase 2 Trials Completed in 10 Years	0.9886 ** (0.008)	−0.0071 (0.020)	−0.0648 ** (0.022)	0.0221 (0.032)
Completion rate of Phase 3 Trials Completed in 10 Years	0.9313 ** (0.023)	0.0199 (0.038)	−0.0585* (0.024)	0.0308 (0.040)
Number of Drugs in Phase 1 Trials Conditioned Complete in 10 Years	25.6667 ** (9.617)	−17.4667 (10.129)	955.0833 ** (186.333)	−798.2833 ** (188.518)
Number of Drugs in Phase 2 Trials Conditioned Complete in 10 Years	54.4000* (21.436)	−41.8667 (22.137)	1154.5167 ** (145.416)	−986.3000 ** (146.901)
Number of Drugs in Phase 3 Trials Conditioned Complete in 10 Years	35.4667 ** (11.333)	−32.4000 ** (11.393)	788.9500 ** (86.539)	−692.1833 ** (88.048)
Growth Rate Number of Drugs in Phase 1 Trials Completed in 10 Years	0.1957 (0.184)	0.4911 (0.368)	0.0689 (0.190)	−0.5064 (0.380)
Growth Rate Number of Drugs in Phase 2 Trials Completed in 10 Years	0.6830 (0.353)	−0.1367 (0.585)	−0.5193 (0.355)	0.1115 (0.590)
Growth Rate Number of Drugs in Phase 3 Trials Completed in 10 Years	0.4752 ** (0.141)	0.2234 (0.306)	−0.3977* (0.184)	−0.1991 (0.350)

The regressions in the figure use four variables Beta 0, Biologic, PyMol, and Biologic*PyMol. Beta 0 is an empty column that represents the Y-intercept of the regression, or in this case, the number that corresponds to our control group of small-molecule drugs. Biologic

is a dummy variable that takes a value of 0 if the Y variable is related to small-molecule drugs and 1 if related to biologics. The coefficient on this term represents the difference between the corresponding Y value for biologics and small-molecules. The PyMol variable takes on a value of 0 for all years before 2000 and a value of 1 for all years after 2000. The coefficient on this term shows the effect PyMol has on the Y variable for both biologics and small-molecule drugs. The Biologic*PyMol variable is an interaction term between the variables PyMol and Biologic. The coefficient on this term shows the difference between the effects PyMol had on the Y variable corresponding to biologics and small-molecule drugs.

The regression "Number of Drugs approved by the FDA" was a regression of these variables on the NDAs and BLAs of the FDA between 1985 to 2021 obtained from the CDER data set linked in Figure C. The regressions "Duration of Phase [insert number] Trials Completed in 10 Years" was a regression of these variables on the data used the graphs "Expected Time to Completion for Phase [insert number] Trials, Conditional on Trial Being Completed in 10 Years or Less" in Figure A. The regressions "Completion Rate of Phase [insert number] Trials Completed in 10 Years" was a regression of these variables on the data used in the graphs "Completion Rate of Phase [insert number] Trials Completed within 10 years" in Figure A. The regressions "Number of Phase [insert number] Trials Conditioned Complete in 10 Years" was a regression of these variables on the data used the graphs "Number of Entities Studied in Phase [insert number] Trials, Conditional on Trial Being Completed in 10 Years or Less" in Figure C. The regressions "Growth Rate Number of Phase [insert number] Trials Completed in 10 Years" was a regression of these variables on the growth rate of the number of trials between each year for both drug types. The growth rate was calculated by taking the percent change, in decimal

form, between years of the data used in the graphs “Number of Entities Studied in Phase [insert number] Trials, Conditional on Trial Being Completed in 10 Years or Less” in Figure C.

Works Cited

1. DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, *22*(2), 151-185. https://doi.org/10.1016/S0167-6296(02)00126-1
2. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million dollars? Health Aff (Millwood). 2006 Mar-Apr;25(2):420-8. doi: 10.1377/hlthaff.25.2.420. PMID: 16522582.
3. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May;47:20-33. doi: 10.1016/j.jhealeco.2016.01.012. Epub 2016 Feb 12. PMID: 26928437.
4. Favour, Danladi, Makurvet.Biologics vs. small molecules: Drug costs and patient access. Medicine in Drug Discovery, Volume 9, 2021, 100075, ISSN 2590-0986, https://doi.org/10.1016/j.medidd.2020.100075.
5. Lill, M.A., Danielson, M.L. Computer-aided drug design platform using PyMOL. J Comput Aided Mol Des 25, 13–19 (2011). https://doi-org.libproxy.berkeley.edu/10.1007/s10822-010-9395-8
6. Commissioner, Office of the. "The Drug Development Process." *U.S. Food and Drug Administration*, FDA, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.

7. Center for Drug Evaluation and Research. “Investigational New Drug (IND) Application.” *U.S. Food and Drug Administration*, FDA, https://www.fda.gov/drugs/types-applications/investigational-new-drug-ind-application.

8. Yuan, S., Chan, H.C.S. and Hu, Z. (2017), Using PyMOL as a platform for computational drug design. WIREs Comput Mol Sci, 7: e1298. https://doi-org.libproxy.berkeley.edu/10.1002/wcms.1298