BIOLOGICS PATENT DATABASE

Michael D. Frakes[*]
Melissa F. Wasserman[**]

It is undeniable that the escalating cost of biopharmaceuticals is a critical issue. Congress has held a half-dozen hearings and introduced over a hundred bills to address prescription drug prices. Many of these bills have targeted the patenting practices of firms that manufacturer large-molecule drugs, which are also known as biologics.

Biologics have complex molecular structures and are derived from living organisms. While biologics hold enormous potential to treat an array of diseases, they represent only a small share of all prescription drugs. The vast majority of drugs are so-called small-molecule drugs, which have simple molecular structures that can be chemically synthesized in a laboratory. Although biologics comprise a small proportion of prescriptions, they have an outsize effect on drug expenditures. That is, while biologic drugs represent only 2% of all U.S. prescriptions, they comprise nearly 50% of net drug spending. Fortunately, biosimilars, the generic form of biologic drugs, can reduce drug prices. When a biosimilar enters the market, drug prices decrease, sometimes dramatically. Yet to date, very few biosimilars have entered the U.S. market.

Although several factors contribute to the high cost of biologic drugs and to the delayed entry of biosimilars, one likely culprit is the patenting practices of firms that manufacture biologics. Two patenting practices are especially controversial, both of which stem from the fact that drug manufacturers obtain a series of patents on a Food and Drug Administration (FDA) approved drug, rather than a single patent. The first practice is pejoratively referred to as "evergreening," whereby a pharmaceutical firm seeks additional patents years after filing the primary patent—i.e., the patent on the molecule that forms the basis of the drug product. Given that patent terms run twenty years from filing, these later-filed patents expire later and thereby extend the exclusivity period of the relevant drug. The second controversial patenting strategy is known as patent thicketing. Firms obtain numerous patents associated with a drug product to form a patent thicket surrounding each drug. The primary goal of patent thickets is not necessarily to extend the exclusivity period of a drug but rather to build a dense web of patents for each drug to ensure that at least some patents survive legal challenges against the patent portfolio.

---

[*] A. Kenneth Pye Distinguished Professor of Law and Professor of Economics, Duke University School of Law; Research Associate, National Bureau of Economic Research.

[**] Ph.D. in chemical engineering, Charles Tilford McCormick Professor of Law, Associate Dean for Research. The authors want to acknowledge the support of *Impacts of Enhanced Drug Examination,* National Institute of Health, National Institute on Aging, R01, 1R01AG076586-01A1, 2023-2027.

Much of what we know about the drug-patent landscape is based on research of small-molecule drugs. U.S. law has mandated since 1984 that brand-name manufacturers of small-molecule drugs disclose to the FDA patents that would be infringed if generics were to launch before the expiration of the relevant drug's patents. The FDA publicly disseminates the resulting patent lists in something known as the Orange Book, which scholars have exploited to study the patent practices of firms that manufacture small molecules.

In contrast, we know virtually nothing about the patenting landscape associated with large-molecule drugs or biologics. The law that governs transparency of patent rights works quite differently in the context of biologics than small-molecule drugs. It was not until 2020 that Congress initiated a process by which the FDA would publish the patent lists associated with biologics in something known as the Purple Book, a biologics counterpart to the Orange Book. However, the information to be disclosed under this 2020 Act is initiated by an essentially voluntary process. Our analysis demonstrates that the Purple Book identifies *less than* 4% of the relevant patents associated with approved biologics.

The result is that scholars lack a patent reference to study the most expensive segment of the pharmaceutical industry—biologics. To fill this gap, we built the first comprehensive patent database associated with all 515 FDA-approved biologics which are comprised of over 11,500 patents. Despite the long-felt need for such a database, no one had taken up the task. This is partly due to the time-consuming nature of building such a database and the need for both scientific and legal expertise. We obtained a grant from the National Institute of Health to help build this database. Our database of patent rights associated with biologics will be publicly available and housed by the National Bureau of Economic Research.

Comprehensive patent landscaping was conducted between November 2023 and January 2025 to identify granted patents (currently active and expired) and patent applications (currently pending and abandoned) associated with all 515 FDA biologics licensed and approved by the FDA as of January 2025. Before discussing the search steps that we took to construct the biologics patent database, we first address the theoretical foundations underlying its scope.

To select the patents to include in our biologics patent database, we utilize the following criterion: we include patents and patent applications that would likely be infringed if a biosimilar were launched (assuming the patent application matriculated into an issued patent) to compete with the focal drug product. Because our inclusion criteria were designed to include those biologics patents that may meaningfully forestall competition, we include only patents that are specifically relevant to biosimilar infringement. If the FDA had only approved the biologic drug alone, we did not include combination patents that claimed an FDA-approved biologic and another chemical compound, as this patent would not be infringed by the launch of a biosimilar. Method of treatment patents for the biologic were only included if the claimed treatment was for an FDA approved indication. Likewise, if a biologic was only approved with a certain route of delivery (i.e. topical, intravenously, etc.) then patents were omitted if they claimed an FDA-approved biologic to a non-FDA approved route of delivery. We emphasize that our decisions here track those taken by Congress in calling for the construction of the Orange Book in the small-molecule context, of course substituting

"patents likely to be infringed by biosimilar entry" with "patents likely to be infringed by generic entry."

The methodology and resources used to build the patent landscape for each drug included the following steps: (1) identifying patents listed on the Purple Book, (2) where applicable and publicly available, identifying patents asserted in litigation in the United States in relation to the biologic, (3) where applicable and publicly available, identifying patents asserted in Patent Trial and Appeal Board (PTAB) proceeding in relation to the biologic, (4) identifying patents that received a patent term extension under 35 U.S.C. §156, which is designed to restore patent term lost during the FDA's review process and (5) conducting patent searches utilizing Orbit Intelligence patent analytic and search software.[1] Any patent that was listed in the Purple Book, or asserted in federal litigation or in a PTAB proceeding, were automatically included in the patent database. However, we took a different approach with the patent searches.

By far the most time-intensive step associated with building the biologics patent database was the last one, conducting the patent searches. The patent searches led to the identification of over 99% of the patents included in our patent database for FDA-approved biologics. Indeed, a key challenge in implementing the patent searches is that patent documents do not contain text fields that directly bear on our inclusion criterion. For instance, assume that we search over Stelara's chemical name ustekinumab and identify a patent with ustekinumab in its claims text. However, further assume that the claim language of the patent speaks to treating indications with ustekinumab that have not been approved by the FDA. Per the above discussion, it would be inaccurate to include that patent as a Stelara patent in our database. Similarly, assume that a careful read of the patent text suggests that the claims speak to treating an FDA approved indication of ustekinumab but only in combination with a second drug that has not been approved by the FDA. For the reasons indicated above, this would further render the patent ineligible for inclusion in our database.

Lacking immediate text-based means of applying our inclusion criterion—e.g., by simply pulling all patents with the word ustekinumab in its claim, it became necessary to take a deeper dive into each candidate patent to assess the application of the inclusion criterion and to screen out failures of this nature. To avoid the risk of inaccurate assessments, we elected not to take an algorithmic machine-based approach. Rather, we chose to read and hand-code all patents that were identified utilizing our search methodology before including them in our database. Of course, given the complexity of patent documents and the time involved with each assessment, it became further necessary to design a search protocol to ensure the feasibility of this validation exercise. In short, we aimed to cast a net that is wide enough to capture all likely patents, but not so wide as to render the validation process unmanageable. To balance these goals, we implemented our search protocol in

[1] One benefit of Orbit Intelligence software is that it allows for the easy grouping of assignees of patents based on the responsible corporate entity—i.e., the entity that filed the biologics license application, including all subsidiaries that may have a different name. For example, searches for biologics owned by Johnson & Johnson would include results for Janssen Pharmaceuticals, Centocor, and Acelion among others. Including all corporate entities with direct ties to the responsible marketing party gives the highest likelihood of finding relevant publications.

different stages, where we traded off on the scope of different search parameters across the stages. Let us elaborate.

Our initial search of all patent types focused on identifying relevant patents owned by the corporate entity that filed the license for the biologic and any additional corporate entity were there was significant evidence of involvement in the development, marketing, or ownership/licensing history of the biologic.[2] This initial search included keywords related to the biologic, including the Trade name (i.e. Humira), the chemical name (i.e. adalimumab), the internal identification code of the biologic (D2E7), and the drug class of the biologic (i.e. Tumor Necrosis Factor). These keywords were cross referenced using Orbit Intelligence with the title, abstract, and claims of the patent documents. If no results were returned within the corporate tree, then the search was expanded to also include matches within the full text of the patent document. In some cases, the text of a document made clear that it read on a particular biologic, but the claims only referenced to "drug" or some other general term. We only included a patent if the claims indicated the biologic in question was the subject of the claim.

An additional initial search was conducted for patents related to delivery of biologic drugs through approved delivery methods (infusion, injection, etc.)—i.e., device patents—only. This additional search focused on the same corporate entities and the same keywords as the initial search. However, this additional search for device patents was broader in one aspect and narrower in another aspect than the initial search. It was narrower in that it required the word "device" or other delivery methods related to the biologic (i.e., inhaler, etc.) to be in the claims of the patent document. The additional search of device patents was broader than the initial search in that the keywords (i.e., Trade name, the chemical name, the internal identification code of the biologic, and the drug class of the biologic) were not limited to matches within the title, abstract, and claims of the patent documents but instead within the full text of the patent document.

The second patent search of all patent types was also broader in one aspect and narrower in another aspect than the initial search. The second patent search across all patent types did not restrict the patent search to be within the corporate entities that either filed the license and/or were involved in the development, marketing or ownership of the biologic. However, the second search was narrower in that it did not include keywords related to the drug class of the biologic (i.e., Tumor Necrosis Factor), as doing so returned far too many non-relevant results. That is, the keywords in our second search included those related to the Trade name (i.e., Humira), the chemical name (i.e., adalimumab), and internal identification code of the biologic (D2E7), which were cross-referenced using Orbit Intelligence with title, abstract, and claims of the patent documents. Importantly, for this second stage, we searched patents assigned to any assignee, not just those that are part of the corporate entity that filed the biologic license.

Finally, we conducted a third search of all patent types using biosequences (nucleotides or amino acids) for the FDA-approved biologic to find any results where the

[2] Additional parties are sometimes involved in the development, marketing, or ownership of a biologic that are not the reference product sponsor. These additional entities were identified by either a general internet search, the use of Adisinsight database by Springer, or the use of LexisNexis database.

biologic was only referred to in the claims of the patent as a numbered sequence rather than by its chemical name, trade name, or internal identifier. Sequences were obtained from Drugbank Online or Kegg Drug in most instances. If sequences could not be found in those two sources, then previously found publications were searched for reported sequences that were designated as matching that of the biologic by name. Patent documents were searched for inclusion of the matching sequence within the claims section of the document. A matching sequence was determined to be either a full sequence match to the reported sequence or one that comprised part of the sequence. Some genetic mutations were included if there was enough other evidence to indicate that the mutant in question was indeed a variant of the indicated biologic and the mutant had a greater than 95% match with the reported sequence of the biologic.

Although the methodology utilized to build our patent biologics database is comprehensive, we acknowledge that it will not find every relevant patent that could possibly be asserted in biologics patent litigation. Perhaps most saliently, our methodology will not identify biologic-agnostic manufacturing patents unless these patents were either asserted in federal patent litigation, asserted before the PTAB, subject to a patent term extension under 35 U.S.C. § 156, or listed in the Purple Book.

Every patent or patent application included in our database was categorized into one of six different patent types. When determining the type of patent or patent application, we evaluated the main claim as reported by Orbitz Intelligence and, if no main claim was reported, we evaluated the first independent claim.[3] Every patent or patent application was assigned to only one type. Those types are defined as follows: main compound, method of treatment, method of manufacturing, formulation, or device.

We used US patent numbers and US patent-application numbers to merge our dataset with other numerous public sources, such as the Patent Examination Research Dataset (PatEx)[4], to collect various patent characteristics for each patent in our biologics database. In the Online Appendix, we briefly describe the various patent characteristics collected from numerous public sources for each patent in our biologics database. Beyond the patent-type flags just discussed, such information includes identifying patent information, claims characteristics (including claims text and linguistic summary statistics), identities of the patentee(s) and assignee(s), patent citation statistics, and terminal disclaimer filings, among other things.

[3] Orbitz Software states the main claim is determined by the U.S. Patent & Trademark Office and they are unaware of the criteria utilized to identify the main claim. Our review of the main claims of patents suggests they are in fact the first independent claim of the patent.

[4] PatEx is published by the U.S. Patent and Trademark Office.