Classification of Generic Manufacturers and Competition in the Pharmaceutical Industry

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

by

Mark Benmuvhar

Dr. Scott Alberts, Thesis Advisor

Department of Data Science

School of Arts and Letters

2022

TRUMAN STATE UNIVERSITY

Kirksville, Missouri

Acknowledgements

Thank you to my advisor, Dr. Scott Alberts, for your encouragement in moving this project from draft to a defensible thesis.

That thesis gained insight from my committee members, Dr. Scott Thatcher and Dr. An-Di Yim. Your review helped me to find the larger story behind the data.

Thank you to my Dad for buying that first personal computer many years ago and for teaching me to find questions.

Thank you to my Mom for the patience to transcribe code for BASIC games and for the same patience when I took apart the computer.

Thank you to Colin, Tom, Susan, Tom, Catherine, and Rose for your support and belief throughout the program.

Thank you to Ethan and Jaden, whose willingness to listen cleared up so many design challenges.

And thank you to my wife, Andria, who kept goals in focus, inspired me to follow through on this journey, and supported me along the way.

Contents

Abstract **iv**

1 Background **1**

2 Data Resources **6**

2.1 Approved Drug Products with Therapeutic Equivalence Evaluations. 25th to 42nd Editions. 6

2.2 Drug Product Labels 9

3 Statement of Security and Ethical Considerations **11**

4 Methods **12**

4.1 Data Acquisition 12

4.2 Data Cleaning and Preprocessing 14

4.3 Topic Modeling for Exploratory Data Analysis 14

4.4 Predictive Models 17

5 Exploratory Analysis **23**

5.1 Market Characteristics 23

5.2 Topics in Usage Codes 39

6 Results **50**

7 Discussion **73**

8 Conclusion **79**

Abstract

Classification of Generic Manufacturers and Competition in the Pharmaceutical Industry

An Abstract of the Thesis by

Mark Benmuvhar

United States generic and brand pharmaceutical companies balance innovation, profit, and public health under the Drug Price Competition and Patent Term Restoration Act. Patent expiration dates demarcate potential profit and legal risk for generic pharmaceutical companies, underscoring the importance of forecasting competition for limited revenue. For this, modules coded in Python extracted data from Food and Drug Administration product approvals and individual product labels to create sets of brand and generic products. Predictive models using classification methods including gradient boosted random forest algorithms were built in R to evaluate brand product patent counts, patent categories, ingredients, and packaging components as predictors of generic competition before and after brand product patent expiry. Gradient boosted algorithms predicted generic competition before first, before last, or after last patent expiry with significant accuracy ($p < 0.05$) of 83.3%, 68.9%, and 69.4%, respectively. Models demonstrated the feasibility of using non-fiscal data to predict generic competition.

1 Background

The United States pharmaceutical industry provided almost 194 billion daily doses of pharmaceuticals in 2021, representing around $776 billion in wholesale costs and $80 billion in patient out-of-pocket costs (IQVIA Institute for Human Data Science, 2022). Considering the US population of around 332 million people, the wholesale cost, which is shared among insurers and patients, works out to around $2300 per person, while the out-of-pocket costs are around $241 per person. However, new drugs may provide a 3 to 6% reduction in years of life lost from multiple disease states, providing a significant benefit to society in the years following their introduction (Lichtenberg, 2016).

In the United States, pharmaceutical products enter commercial trade after the Food and Drug Administration (FDA) issues a marketing authorization based on either a new drug application (NDA) for a novel "brand" drug product or an abbreviated new drug application (ANDA) for a generic drug product (Drug price competition and patent term restoration act, 1984). A separate pathway exists for complex molecules submitted under a biological licensing application, which is outside the scope of this paper. An NDA documents extensive studies of product safety and efficacy and can represent over a billion dollars of development costs and over a decade of research expenditures (Sun et al., 2022). Branded pharmaceutical products are developed based on medical need, prevalence of disease and technical feasibility (Tamimi & Ellis, 2009). Therefore, branded product manufacturers have a luxury of time to optimize both the manufacturing process and selection of inactive ingredients that comprise the drug delivery system and individual dosage forms. Refinement occurs during preclinical trials and during the numerous production cycles required to produced clinical supplies and demonstrate the commercial manufacturing process (Buckley et al., 2020).

In contrast, a generic sponsor of an ANDA does not need to prove safety and efficacy through clinical trials. Instead, they primarily need to demonstrate equivalent performance to the reference drug. Sponsors prove therapeutic equivalency between generic and branded prod-

ucts through a small clinical study in volunteers showing that there are no differences in the rate or amount of drug absorbed into the body. Further, the brand and generic must be pharmaceutically equivalent, having the same active pharmaceutical ingredient, same dosage form, same strengths, and same route of administration (Center for Drug Evaluation and Research (CDER), 2017). Since clinical effectiveness is a function of drug concentration, demonstration of similar drug concentrations makes it unnecessary for generic products to duplicate extensive clinical performance trials required for branded product approval (Midha & McKay, 2009).

Generic sponsors consider profit potential, dosage form complexity, and toxic material handling requirements as potential constraints in new product manufacture (Lee et al., 2016). Because generics require fewer clinical trials and less time to develop, the products may reflect the distinct formulation and manufacturing paradigms available at each company. Attempts to circumvent patents may result in differences between brand and generic formulations. While the resultant generic product demonstrates clinical similarity to the branded product, it may have differences in the manufacturing process or inactive ingredients. However, the cost of 5 years of generic drug development was projected to be around $2.5 million dollars, or around 1/400th the cost of the branded pharmaceutical [slb21].

Reduced spending on pharmaceutical development provides meaningful value to the consumer. While average cost in 2021 for brand pharmaceutical prescriptions were around $21 for commercially insured individuals, the generic cost was around $5, representing a fourfold savings. The combined role of insurance and low cost generic pharmaceutical prices is significant, considering the average out-of-pocket cost for a brand product was around $91 without insurance (IQVIA Institute for Human Data Science, 2022).

Given the scale of pharmaceutical investment, achieving meaningful investment returns is a necessary business objective. Across all pharmaceutical companies, over $83 billion was spent on research and development in 2019, which was around a quarter of corporate costs

(Congressional Budget Office, 2021). These rates are justifiable, as up to 90% of drug molecules are reported to fail in clinical studies (Sun et al., 2022). As a result, the internal return on investment for a cohort of large biopharmaceutical companies was observed to be only 3 to 7% in a longitudinal review conducted by the Deloitte Centre for Health Solutions Deloitte Centre for Health Solutions (2021).

To protect the development investment, brand pharmaceutical companies patent their novel discoveries. In exchange for teaching the public details of the invention, the company receives a 20 year exclusive right to practice the invention (*Contents and term of patents, provisional rights*, 2013). Further, the receipt of additional patents at a later date covering new applications may extend the exclusivity an additional 10 to 20 years due to overlapping terms. Because generic medicines need to mimic the physiological performance and often compositions of the brand product, patents exposed generic products to infringement liability for producing clinical trial materials during development. This is evident in the 1984 case of Roche Products, Inc. vs. Bolar Pharmaceutical Corp, in which federal court found Bolar had infringed on Roche's patents when producing non-commercial materials for preliminary a bioequivalence study (Russo & Johnson, 2014). Although a regulatory pathway existed to approve generic drugs, the legal risks discouraged companies from pursuing generic products.

In 1984, Congress approved into law the Drug Price Competition and Patent Term Restoration Act Drug price competition and patent term restoration act (1984), which established the modern approval pathways. The act established a compromise between protecting intellectual property and enabling a development pathway for generic medicines. Innovative products received additional years of patent protection based on the length of FDA application review that occurred concurrently with the patent. Generic companies gained protection from infringement liability during developmental clinical studies conducted to support product an ANDA. However, upon submitting an application, the generic sponsor had to attest to the absence of patents, expiry of all patents, an intent to withhold marketing until after

patent expiry, or that any existing patents were invalid or not infringed upon.

Under the new law, the act of submitting a drug product with an invalidity or infringement assertion became a surrogate act of infringement. If the generic company made this claim, the branded company could file an infringement lawsuit within 45 days of the FDA's receipt of the application. At that point, the law enjoined the FDA from approving the application until the earlier of 30 months or the resolution of any litigation. If the generic company prevailed in litigation and were the first company to file an abbreviated drug application, they would themselves receive a 6 month exclusivity period to market the generic product before approval of any additional generics could be approved.

Patents allow brand sponsors to sell a new drug at inflated prices due to a monopoly position (Congressional Research Service, 2021), and significant time is granted to defend the validity of product patents. However, significant profit erosion occurs once generic products enter the market after the expiry of patents or exclusivity (Olson & Wendling, 2013), (Reiffen & Ward, 2005). The selling price for the first generic product was observed to be a median of around 76% of the brand product. Meanwhile, the selling price of the next four competitors each decreased the price around 20%. As a result, products with 5 or more generic products competed for a sales price observed to be around 95% of the brand's wholesale product cost (Conrad & Lutter, 2019).

Considering litigation risks and effect of competition to decrease price, generic products have the best opportunity for profitability and highest legal risk before patents expire and other generics enter the market. A lesser opportunity for profit and lower risk occurs when a generic can reach the market after all regulatory exclusivity periods expire but before other companies have eroded the sale price. The period between first and last patent expiry presents an interesting opportunity for product introduction, as the risk of litigation may mitigate the benefits of entering the market before other competitors. However, other competitors may be dissuaded from entering the market due to these risks, resulting in a higher

market share.

Because of the time required for product development and regulatory review, identification of development projects requires accurate forecasting of the competitive landscape three to six years after project inception. Equally, predicting competition during patent lifespans may lend critical insight in timing the starting points for new development projects. Therefore, a significant benefit for improving predictions exists in performing data analysis on historical patterns in pharmaceutical product approvals, patent protection, and regulatory exclusivity.

Other researchers have reported factors influencing generic competition. Between 2012 and 2017, decreasing rates of generic product substitution in new markets were influenced by authorized generics, market revenue, and first-to-file status (Rome et al., 2021). Separately, generic companies may introduce a product at risk of adverse court decisions during unresolved litigation (Drake et al., 2021). When companies entered markets at risk, drug substance and drug product patents were significantly absent. While the number of patents was not significantly associated with this decision, market sales below $50 million were a significant feature where companies entered at risk. Market competition levels, market size, and dosage form were used to predict the presence or absence of drug shortages in a separate study (Dave et al., 2018). Low priced drugs were observed to have a higher risk of shortages between 2008 and 2014. While this study considered a binary classification, Beall et al. (2018) evaluated drug substance patents, patent term restoration, pediatric exclusivity periods, and statutory limitations of product approvals as variables predicting the fiscal quarter of product entry following brand product approval. In a cohort of top selling drugs between 2000 and 2014, products having only secondary patents unrelated to the drug substance were observed to have earlier competition than products with a drug substance patent.

From the prior studies, details about the brand product's patents and dosage forms could be used to in predictive models. Because patents protect brand monopolies, the number and subject of said patents may estimate the magnitude of obstacles forestalling generic

approvals. Details available in the product labeling may also support predictions. Dosage forms and ingredients used in the brand product may reflect underlying complexity that makes a product more difficult to copy. This may be a function of either patented compositions or complexities of the dosage form itself. Required packaging designs may reflect unique handling requirements that are difficult for generic companies to mimic.

Evaluation of the FDA's published records of drug product approvals, patents, and labeling information will enable a comprehensive review of public details across a broad time range. This will supplement previous studies, which relied on subsamples limited by contemporaneous products and market size. Last, building a classification model for the presence or absence of competition will establish a screening tool for determining potential markets with first-to-file opportunity for generic companies without regard to specific timing of entry.

2 Data Resources

2.1 Approved Drug Products with Therapeutic Equivalence Evaluations. 25th to 42nd Editions.

Each year, the FDA publishes the commonly referenced "Orange Book" (United States Food and Drug Administration, 2005 to 2022), an official record of all drug products approved for marketing in the United States. This book establishes which products are generic and substitutable for each innovative brand product. It contains a record of all products discontinued from marketing. Finally, it details which patents the brand products claim for each drug product. These specific patents may form the basis for any infringement lawsuits brought against generic applicants in accordance with the Patent Term Restoration Act. Details of the patents also include any regulatory exclusivity clauses that accompany each patent. These sections differ in format and content, which may be described as follows.

Common details in each section include the drug substance name, unique application num-

ber, and product code, which enumerates differing product strengths or other variations. Together, these establish a unique key for correlating distinct records across sections.

The Prescription Drug Product section includes the following fields:

Route of Administration: The FDA uses specific terms for how a product is introduced to the body. This broadly differentiates products taken by mouth (oral), injected into the bloodstream (intravenous), and similar demarcations.

Dosage Form: Products intended for each route of administration may take different forms. For example, orally administered products comprises both liquids, tablets, and capsules taken by mouth.

Drug Product Name: Commercial names for the product may either be an invented trade name for the brand product, such as the cholesotrol drug Lipitor, or the name of the drug substance used in the generic product, such as acetaminophen, found in the pain treatment Tylenol. Because the product names are not unique, they can be ignored in the analysis for this report.

Reference Designation: For each unique drug substance, route of administration, and dosage form, the FDA must designate one or more products as the reference product for subsequent generic products to test against. In certain cases, a separate product receives designation as a reference standard due to discontinuance of the original reference product or the replacement of the reference product with newer products that follow modern application requirements.

Therapeutic Equivalence Code: Where multiple product sources exist for a drug substance and route of administration, the FDA assigns a two letter equivalence code, such as "AB". Generally, the first letter represents whether product substitutions are allowed ("A") or forbidden ("B"), while the second letter represents the route of administration. Due to differences in physiological absorption or performance, a suffixed digit may differentiate prod-

uct classes within the same drug substance and administration route category. Similarly, differences in product performance may lead to the FDA adjudicating that a generic cannot be substituted for the brand.

Sponsor Name: A product's sponsor is responsible for submission and maintenance of the application. Due to changes in corporate identity due to merger or restructuring, the sponsor name may change across editions of the Orange Book.

Approval Date: The approval date occurs when the FDA provides marketing authorization for a product after application review.

The discontinued product section duplicates entries of the prescription drug product list except that bioequivalence codes are unlisted. Individual product codes and strengths under an application may be withdrawn from the market while others remain marketed. The year of discontinuance is not reported and must be implied from the Orange Book edition in which the item first occurs.

The section entitled Prescription and OTC Drug Product Patent and Exclusivity List enumerates the following details:

Patent Number and Expiration Date: For each application and product number (strength), the assigned patent number and patent expiration date are listed. In rare cases, patent numbers may not apply to each strength in an application. Multiple patents may be assigned to a single drug product;

Patent Codes: Codes describe if a patent pertains to the drug substance, the drug product (dosage form), or therapeutic usage area. The usage code may be one of over 3000 four digit codes that correspond to a brief description of the clinical or treatment aspects claimed by a patent. For example: "U-2627: Topical treatment of plaque psoriasis in patients 12 years and older." Multiple usage codes may be assigned to a single patent;

Exclusivity Codes and Expiration Date: Exclusivity periods represent statutory pro-

hibitions that prevent the FDA from approving generic products for a certain time period. These arise from laws enabling new molecule development, treatment of rare diseases, clinical studies in pediatric populations, and other reasons. Exclusivity periods may run concurrently or in serial with patents. Multiple exclusivity codes may be assigned to a single patent. Certain exclusivity codes, such as those for pediatric exclusivity or orphan drug designations, may be assigned to the drug product rather than a patent and are listed as independent line items.

While the FDA maintains an electronic database of the Orange Book, the online version only captures data for the current year. Certain details are deprecated following periodic updates. As patents expire, they are removed from the online listing. As products are removed from the market, the equivalence relationship between branded and generic products is removed. To develop a comprehensive review of historical applications, it was necessary to perform data extraction from PDF renditions of the Orange Book. Historical editions were made available online via a regulatory law firm (Hyman, Phelps, McNamarra) following a Freedom of Information Act request to the FDA for electronic copies. Due to illegibility of the original scans, data could not be extracted from editions prior to the 25th edition. However, an archival project existed online that captured key elements of the patent history (Drug Database, N.D). Researchers for this database had also encountered legibility challenges and had performed manual entry. In turn, a portion of these entries were manually against the Orange Book and acceptable supplements for the extracted patent information.

A description of the parsed data is presented in Table 1. Details related to data extraction from the PDF file and associated challenges are reported in Section 4 - Methods.

2.2 Drug Product Labels

The FDA has required and collected product labeling information for all marketed drug products using a structured XML format since 2005. Complete details of the structured label

Table 1: Record Summary from the FDA Orange Book, 25th to 42nd Edition

Section	Line Items	Unique Products, Manufacturers, and Strengths	Unique Products and Manufacturers	Unique Patents
Marketed Products	285407	71321	32510	NA
Discontinued Products	222169	53339	20940	NA
Patents	182713	76366	2628	40165

content and hierarchical XML scheme are outlined in the FDA's implementation guidance (Center for Biologics Evaluation and Research, Center for Drug Evaluation and Research, and Center for Vetrinary Medicine [CBER, CDER, and CVM], 2018). All XML files are publicly available and are maintained by the National Library of Medicine National Library of Medicine (2022). Among the exhaustive information on clinical studies, safety, efficacy, and use, each label also describes the qualitative composition and descriptive details of each drug product. These data included interesting features such as product size, shape, color, and imprinting. Each label references an approved application number and can be joined with Orange Book entries to characterize products. The dataset contains 46,640 entries provided as zipped XML files. The labels reference 1,194 unique ingredients and 60 distinct packaging configurations.

In the product labels, ingredient names are paired with a unique ingredient identifier (UNII), which provides an unambgiuous alphanumeric code for the underlying molecular structure. Some variance in product names has occurred across manufacturers and years of label revision. Therefore, ingredient names were standardized based on the name associated with the first occurrence of each UNII. No standardization was performed on the packaging components.

The product label archives are limited in that they are not maintained for discontinued products. Approximately half of the distinct application numbers extracted from the Orange Book do not have associated package labels. Therefore, analysis using ingredients and

packaging may exhibit a "survivor's bias". Analysis using these features reflect a subclasses of products that have remained on the market rather than being discontinued. While features of discontinued products may contradict the analysis, I am unaware of any accessible sources for the information.

3 Statement of Security and Ethical Considerations

For this thesis, I will make security and ethical considerations concerning the availability and privacy of data used in the research as well as to the impact of conclusions realized from the analysis. Broadly, this will consider the public nature of the Orange Book information and product labels. Exploration of these data sets will attempt to establish broad market patterns and could characterize individual companies. To the extent that these relationships exist, the derivation of private practices from public data merits ethical review.

The ethical considerations of the study acknowledge that the data is derived from public sources and government records. The subject of drug approvals is corporate rather than individual, and issues of personal privacy are not material. To the extent that this report seeks to discover patterns associated with drug development and approval, the ethical considerations demand adherence to the professional code of conduct established by the American Statistical Association (Comittee on Professional Ethics of the American Statistical Association, 2018). As the outcome of this study may allude to business practices reflecting trade secrets, it does so based on a practiced reading of public record and proper application of statistical techniques.

4 Methods

4.1 Data Acquisition

Data extraction and limited cleaning was first performed using custom Python applications. Separate applications were to parse product labels, active product records, discontinued product records, and patent information. These applications implemented standard libraries including numpy, pandas, os, and re. Other packages are discussed in the descriptions of each application.

Compressed product label zip files were obtained from the National Library of Medicine's archive. Applications to traverse the zip files and parse the product labels were readily constructed, as the XML format creates a heierarchical structure of name tags and information. Both zipfile and xml.etree packages used for this are in the standard Python library. Following the schema outlined in the FDA's guidance document, inactive ingredients and packaging components were extracted from the relevant label sections.

In brief, the extraction algorithm iterated across .zip files in a directory. Each .zip file was opened, and the contained .xml file was opened for processing. Separately, the XML tree was navigated to obtain a dataframe containing product identifiers with either packaging components or a list of ingredients. During processing of each file, the results for packaging or ingredient components were maintained in a list of dataframes. Some variance was observed in the schema for ingredients, resulting in a minority of cases using the tag "inactiveIngredientSubstance" rather than "ingredient". The algorithm was modified to capture both cases. Once all files were processed, the list of dataframes was concatenated into a single data frame, and the results were written to a csv file for downstream processing. Delaying combination of dataframes until the last step improved program efficiency because of the lower computational overhead associated with working with the simpler list structure during each iteration (Pandas, 2021).

In contrast, the Orange Book files presented numerous challenges in format and content. For each section, data groups reflected a heirarchical structure based on labeled headers. Considering the Pharmaceutical Product section, each group of drug substance, route of administration, and drug product headers were followed by line item details for relevant applications grouped by sponsor. Columns were maintained using consistent white space between entries. For each different strength, sponsor and header information were not repeated. Further complicating this, sponsor names and product strengths could require multiple lines, resulting in extraneous white space. Last, the spacing of columns between Orange Book editions varied. Because the contents and size of each grouping varied, each page was effectively unique. Therefore, data extraction required tables to be parsed row-wise to collect required cells.

A custom application using tabula (Frerebeau, 2019) and (pyPDF2, 2022)] was designed. tabula provides a flexible tool for extracting PDF data from a defined table or field format. Although the variable field size prevented full use of the package, it provided a valuable tool for converting column and row based data into a dataframe. Similarly, pypdf2 provides basic tools for identifying and extracting pages. While a description of the process is provided here for extracting drug product details, the overall algorithm is applicable with modification for fields of the discontinued product and patents sections.

Column spacing for each .pdf file was manually determined and used as a predefined value for the algorithm. For each pdf file, the algorithm first identified the start and ending page for the Pharmaceutical Product section based on page header information. Within this range, individual rows containing data fields were extracted. In a second step, the extracted details were parsed. A temporary line item was created to store heading elements (drug substance, route of administration, and drug product name). Subsequent lines were parsed to collect remaining fields. If the following line was a partial entry due to long sponsor or strength entries, the contents were appended to relevant fields. If not, the entry was appended to the

temporary line, and the temporary line was added to the master data frame. If the following line contained new header details, the temporary line was further updated with new product details.

4.2 Data Cleaning and Preprocessing

The ingredient and packaging information were usable as described in the data acquisition section. Additional steps for cleaning, consistency, and consolidation were performed on data extracted from the Orange Book. Cleaning steps ensured similar punctuation used for separators and eliminated excess whitespace from fields. Also, legal entity designations "CORP", "CO", "INC", "LTD", "LLC", and "GLOBAL" were removed from sponsor names. While these terms may reflect changes in underlying operating principles or business units in a company, it was decided to minimize identity changes related to a legal status. Consistency steps addressed FDA's gradual standardization of dosage form descriptions, administrative updates, and changes in company names. Due to a wide range of administrative routes listed for injectable, intravenous, and oral products, a number of low frequency categories were reduced to more general classifications. Similar condensation of oral dosage form routes were made. Changes in entries occurring across editions are described in Table 2. Last, consolidation steps eliminated duplicate entries present following the previous standardization steps.

4.3 Topic Modeling for Exploratory Data Analysis

The primary project goal involves building a model for predicting whether generic competition occurs before the first or last brand patent expires. Key features derived from the Orange Book and product labels will be used as predictors. These include patent use codes, brand product approval dates relative to patent expiry, total patent features, ingredients, and packaging components.

Table 2: Constraints to Ensure Consistency among Orange Book Editions

Constraint	Rationale
Distinct products are defined by the application number and route of administration.	Establishes a consistent drug product identification
Duplicated entries based on application number, product number, and drug substance are limited to the first occurrence	Eliminates redundancy of the data file while recognizing status quo at time of application submission.
Where multiple strengths are issued to an application on multiple dates, only the earliest date will be considered	The thesis considers only the primary activities required for the initial drug approval. Adding a drug strength to an application often relies on significant existing research. Further, they are less likely to be impacted by patents due to elapsed time for patent expiry.
Where the sponsor's name changes due to a company action, the original name of the applicant will be retained.	The thesis considers only the corporate identity at the time of approval. Although the industry has trended towards consolidation, the ability of smaller companies to produce generic products contrasts efficiencies established at larger companies.
Only products for which the original branded product remains marketed will be considered.[1]	This ensures formulation information is available based on the removal of discontinued products from the label database.
Where multiple reference products are available, the product with the earliest approval date will be used as a reference.[1]	This ensures a consistent starting point to determine opportunities for generic entry.

[1] Constraint applied only for predictive model development

The patent use codes from the Orange Book offer possible insights towards both individual product complexity and between product similarity. Because products may receive multiple use codes for each relative patent, it was hypothesized that more complicated products may have more use codes. The number of use codes and total patents a product holds could might correlate with the occurrence of generic competition.

While the individual use codes are themselves arbitrary, their descriptive text could link different drug products. For example, the term "hypertension" appears in 65 distinct code

definitions. In turn, the related usage codes appear in 141 different products. Thus, the text of the usage codes could be used to define treatment categories across a broad range of products. These treatment categories could be assessed for differences in the rate at which generic competition occurs. Therefore, topic modeling was performed on the use codes to categorize similar treatment categories among drug products.

In topic modeling, word frequencies within a document are compared to their frequency in a library of documents (the corpus). Related terms that occur at high frequency in only a subset of documents define a topic within the collection. The number of topics found across a corpus is ultimately an arbitrary value guided by a model's end use (Liu et al., 2016). Therefore, one must estimate an appropriate number of topics that balances both discrimination of documents and identifies common features. As implemented by the stm package (Roberts et al., 2019), each resultant topic reflects a dictionary of words and their relative weight, beta, in predicting each topic. In turn, this matrix can be applied to the terms in a document to determine the probability, theta, that a document is described by a topic.

Typically the highest probability class is used to define one topic for each document. However, where the number of topics are uncertain or broadly defined relative to the total number of documents, it may be useful to retain the probabilities. The weighted membership across multiple topics may enhance classification efforts, while selection of a threshold value will eliminate topics with weak associations.

Commonly occurring words, stopwords, are a context-independent collection of articles, prepositions, helping verbs and other grammatical elements that provide minimal information about the content of a document. These words are commonly defined dictionaries accompanying the topic modeling package. Within a technical document corpus, reliance on frequency alone may be inappropriate for determining stopwords, as some technical terms may occur as frequently as common stopwords. Therefore, both domain-specific expertise

about context and numerical counts should be applied when constructing a stopword list for a specific corpus (Sarica & Luo, 2021).

Topic modeling was performed using the quanteda and stm packages. For exploratory analysis, a corpus was collected across the collection of distinct brand product applications, and a dictionary was built from the usage code definitions. The package-defined stopword list, numbers, and punctuation were removed during tokenization to single words. In addition, certain high frequency words such as "patient", "pharmaceutical", and units-of-measure were removed due to low discrimination power in this context. High frequency words like "treatment" were retained, as they could distinguish products from those designed for "prevention". The remaining individual terms were translated to a document-feature matrix with the dfm function.

The structural topic model was estimated using the stm function from package stm with 10, 60, 120, and 200 topics. Model quality was assessed holistically by comparing the number of topics above an arbitrary cutoff of 20% relevance (theta) for each document. This balances the thematic variance expected among pharmaceutical patents while minimizing the number of relevant categories. Results from this study will determine relevance of including usage codes as part of the feature sets for predictive models discussed in the following section.

4.4 Predictive Models

The specific relationship between features of the brand product and the occurrence of competition within a window will be investigated using supervised classification techniques including support vector machine (SVM), decision trees (RPART), and boosted gradient methods (XGBoost). These methods are appropriate because a mixture of discrete and continuous independent variables is used to predict binary categories of competition. The SVM algorithm attempts to establish a hyperplane separating variable values between classes. Decision trees establish a rules-based classification structure based on increasing the homogeneity of the

data set following each division (Therneau & Atkinson, 2019). However, tree structures may have poor accuracy when applied to testing sets due to sensitivity to new data, discontinuous division among values across nodes, and sensitivity to noise (Hastie et al., 2017).

The XGBoost model used in this study implements a random forest algorithm with gradient boosting. Basic random forest models form an ensemble of decision trees from random features extracted from random samples. Predictive features are determined from an average of hundreds or thousands of decision trees and their performance in classifying samples omitted from the model (the out-of-bag error estimate). Gradient boosting passes model error estimates from each tree node between iterations. In subsequent trees, decision points at each node are adjusted to minimize the tree error. The error reduction improves identification of predictive variables and selection among correlated variables (Chen & Guestrin, 2016). The gradient boosted algorithm was considered well-suited for this classification study due to the expected covariance among ingredients and dosage forms.

The data sets used for analysis were limited to only products with oral, injectable, intravenous, or topical administration (designated as "multi" in charts. From market analysis, these categories represented the majority of marketed products. Because most products are for oral administration, that route will also be investigated alone. Narrowing the number of administration routes was intended to increase similarity among sample sets and the respective companies that manufacture these products. Companies with similar products are assumed to have business practices that are more representative of each other, resulting in better model accuracy. Last, samples will be limited to product applications occurring from 1990 to 2010. The start time accounts for changes in the regulatory paradigm occurring in the early 1980s, while the end time allows for some elapsed time between product approval and potential generic entrance.

Primary data sets used for market analysis were the basis for model development. The primary data set consisted of approved products between 1990 and 2010 joined with listed

patents for approved branded products. Key independent variables from this data set were calculated with the following rationale:

Time before first (bf_first) and last (bf_last) patent expiry: Period (in days) between the branded product approval and expiry of the either the first or last patent. Both features reflect the time available for generic product development following public announcement of a product approval. Conceptually, longer development times may enable generics to come to market before patent expiry either through development efforts or through legal maneuvering

Count of patents (pat_ct) with subjects covering drug substance (DS_ct), drug product (DP_ct), pediatric exclusivity (PE_ct) or usage (use_ct): These features serve as descriptors of the patent subject matter and the overall rigor with which a company has sought legal protections.

Previously developed products (brand_prev_prod_ct): Determined from a product's rank-order relative to previous approval dates, this value represents a company's overall development experience. Conceptually, companies that have produced more products may have differing legal or development strategies for discouraging that result in less competition than companies that have produced few previous products.

For further analysis, this data set was joined with the use code topic, packaging configuration, ingredient composition, or combinations thereof. Package configuration and ingredient composition were one-hot encoded based on package type or ingredient name. The use code reflected the theta value, or proportional fit, calculated between the code and each topic. Response variables were selected as to whether generic competition occurred "before first patent expiry", before last patent expiry", or "after last patent expiry". Independent models were developed for each response variable.

Because the theta values from usage codes represent the probability of belonging to a class, the topic, they can act as a predictive feature for predicting generic competition. For models

incorporating usage codes, each training set built a term matrix based on usage codes present in the training data. Term frequency in the tests sets were then calculated relative to the training vocabulary using the stm::alignCorpus)(). Theta values for topic probability were then calculated using stm::fitNewDocuments().

Response variables were constructed based on the presence of any generic competition occurring before the first patent expiry date (has_before_first), before the last patent expiry date (has_before_last), or after the last patent expiry date (has_after_last). Competition before the first patent expiry is a subset of all competition occurring before the last patent expiry. As described in the project background, these periods may posses different legal risk and profit opportunities and may be impacted by the independent variables.

Data sets were divided into training and test sets using a 70:30 ratio. Due to both the low number of total observations and limited number of sponsors observed in certain categories, the training set was upsampled with replacement to obtain a 50:50 ratio of samples with or without observation for each category. Also, topic models were developed independently for each training set. Details of the training sets are presented in Tables 3. Details of the SVM, rpart, and boosted gradient methods are presented in Tables 4, 5, and 6.

Prevention of "data leakage" between training and test sets is a critical feature of method validation. A fair estimate of method accuracy requires that test sets are independent of the training set. Test sets can become contaminated with information from the training set when standardization or other data pretreatment is performed on the entire sample set prior to selection of training samples (Hastie et al., 2017). Therefore, it is necessary to perform scaling and categorization on training sets alone.

Training sets were scaled between 0 and 1 for each variable as

$$(x - min(X))/(max(X) - min(X))$$

Table 3: Test and Training Set Characteristics

features	Approved	Route	n, train	n, test	Prevalence, test
a	pre-first	MULTI	1088	293	0.208
	pre-last	MULTI	738	293	0.536
	post-last	MULTI	870	294	0.367
	pre-first	ORAL	756	211	0.237
	pre-last	ORAL	622	211	0.630
	post-last	ORAL	602	211	0.393
b	pre-first	MULTI	710	203	0.251
	pre-last	MULTI	602	203	0.635
	post-last	MULTI	580	202	0.391
	pre-first	ORAL	504	149	0.275
	pre-last	ORAL	510	149	0.732
	post-last	ORAL	406	148	0.419
c	pre-first	MULTI	738	209	0.249
	pre-last	MULTI	616	209	0.627
	post-last	MULTI	602	209	0.388
	pre-first	ORAL	516	152	0.276
	pre-last	ORAL	518	152	0.730
	post-last	ORAL	414	151	0.417

[a] Base Patents and Use Codes Alone
[b] Ingredient Features Included
[c] Packaging Components Included

Table 4: SVM Model Parameters

Parameter	**Type**	**Kernel**	**Gamma[a]**	**Epsilon[b]**	**tolerance[c]**	**Cross[d]**
Value	Classification	Radial	1/Dimension	0.1	0.001	5

[a] Model scalar [b] loss function coefficient [c] Termination criteria [d] Cross-validation sets

Table 5: RPart Model Parameters

Parameter	minsplit[a]	Minbucket[b]	cp[c]	maxdepth[d]	xval[e]
Value	3	1	1e-04	5	5

[a] Minimum samples in node for division
[b] Minimum samples in terminal leaf [c] complexity factor, or minimum required increase in model complexity per division
[d] maximum tree depth [e] cross validation splits

Table 6: XGBoost Parameters

Parameter	Subsample[a]	Colsample[b]	Early Stopping Rounds[c]	Max Depth[d]	Eta[e]	Gamma[f]
Value	0.5	0.5	4	6	0.33	0.001

[a] Observations chosen at random per interval [b] variables chosen at random per interval
[c] Maximum training iterations without accuracy improvement [d] maximum tree depth
[e] Learning rate parameter [f] Minimum loss reduction to partition a node

where x is the individual value, X is the vector of variable values, and min(X) and max(X) are the respective minimum and maximum values for the vector. To ensure test sets are aligned with the training set, the minimum and maximum values from the training sets are applied to the test set. Because test sets contain independent ranges, values less than 0 and greater than 1 may occur, resulting in extrapolation from the trained model.

Depending on the administration route, samples with generic competition may only have a prevalence between 20 and 40%. Random over- and under- sampling methods may be used to augment the training set. As implemented in the caret package, over-sampling adds duplicates of the minority sample until a threshold proportion is reached, while under-sampling removes majority samples from the training set. For this study, oversampling was selected to provide equal exposure to features contained in data sets with and without generic competition. However, these methods may worsen model performance due to overfitting or due to the loss of variability from the majority set (Branco et al., 2016). Conversely, ensuring class similarity between the training set and the true sample population was observed to

improve biomedical model accuracy (Lötsch et al., 2021). Therefore, the decision to resample and the choice of methods are an important feature of model design because they determine visibility of features from both the the majority and minority classes.

Because the number of "true" events in the sample sets are small, model accuracy should judge both sensitivity and detection rate in context of the event prevalence. Sensitivity is the correct identification of an event compared to its actual occurrence. In contrast, detection rate is the correct identification of an event vs. the total number of observations. This is closely related to prevalence, which is the true number of an event compared to the total number of observations. For binary systems, the larger of prevalence or non-events defines the no-information rate. This is the expected rate of choosing a correct event by chance.

For this study, the overall test set accuracy should be higher than the no-information rate, while detection rate should be similar to the event prevalence. Since observed rates of generic competition may be as low as 10 to 20%, the p-value for model accuracy being higher than the no information rate should be monitored rather than overall accuracy.

Data analysis was performed using RStudio version 2022.07.1 Build 554 on a Windows 10 Pro operating system with AMD Ryzen 5 3600x 6 core processor with 32 GB RAM.

5 Exploratory Analysis

5.1 Market Characteristics

From 1989 to 2021, the number of products approved each year has increased steadily (figure 1). This follows a period of increasing product introduction from 1982 to 1987. Reviewing additional FDA history, significant regulatory changes occurred following the introduction of the Patent Term Restoration Act in 1984. In 1987, several allegations of fraudulent generic applications and lax regulatory oversight occurred, which resulted in more stringent approval requirements by 1992 (Hornecker, 2009). Further, the agency issued policy changes beginning

in 1990 that allowed multiple product strengths, colors, and shapes to be covered under a single application (Center for Drug Evaluation and Research, 1998). This may account for the decrease in applications between 1989 and 1990. Despite changes in generic approvals, the number of branded NDAs remained relatively constant during the same period.

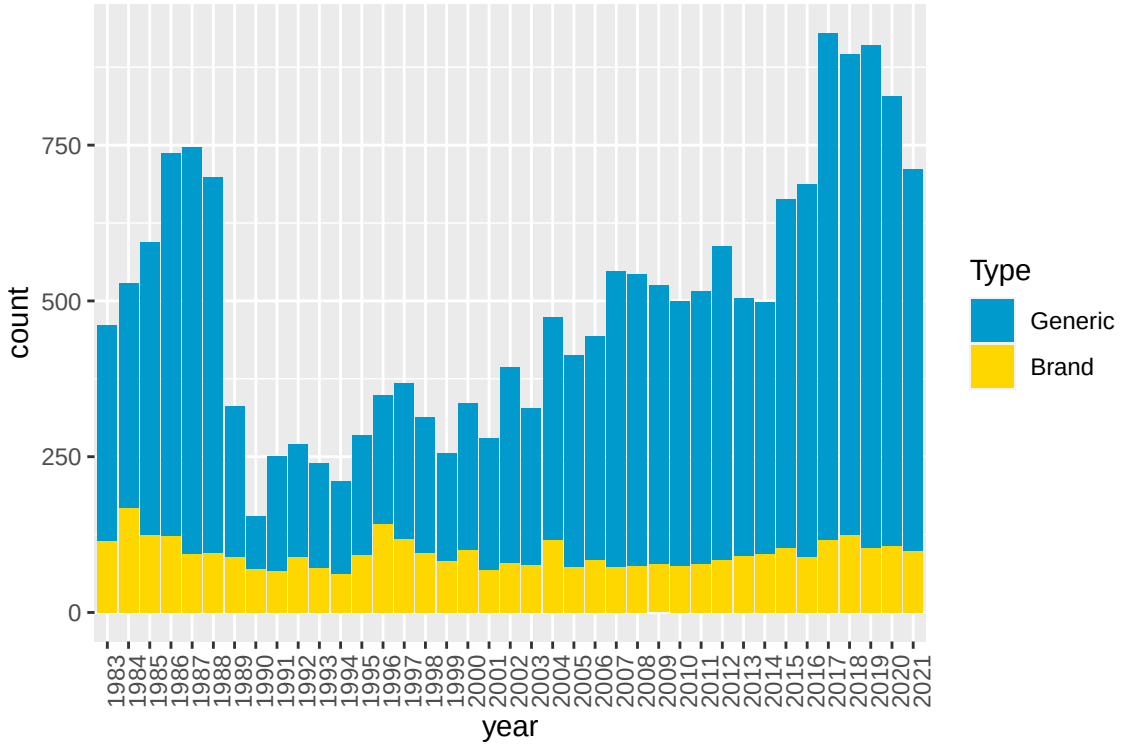


Figure 1: Total Yearly Product Applications (Brand and Generic, 1983-2021): A steady increase in generic ANDA approvals was observed from 1990 to 2021

Almost two thirds of the approved drug products were composed of orally adminstered dosage forms. Injectable, topical, intravenous, and ophthalmic drugs comprised most of the remaining drug categories (figure 2). Simple tablets and capsules designed for immediate release further composed around two thirds of the oral dosage forms, while products modified for release over time, simple liquids, or rapid dispersion in the mouth or liquids comprised the remaining third (figure 3). Less dosage form variety was observed in the injectable product subcategories (figure 4), while topical products had more gradual differences among subcategories (figure 5). Based on the market categorization, the top four routes of administration

were selected for use in downstream modeling.

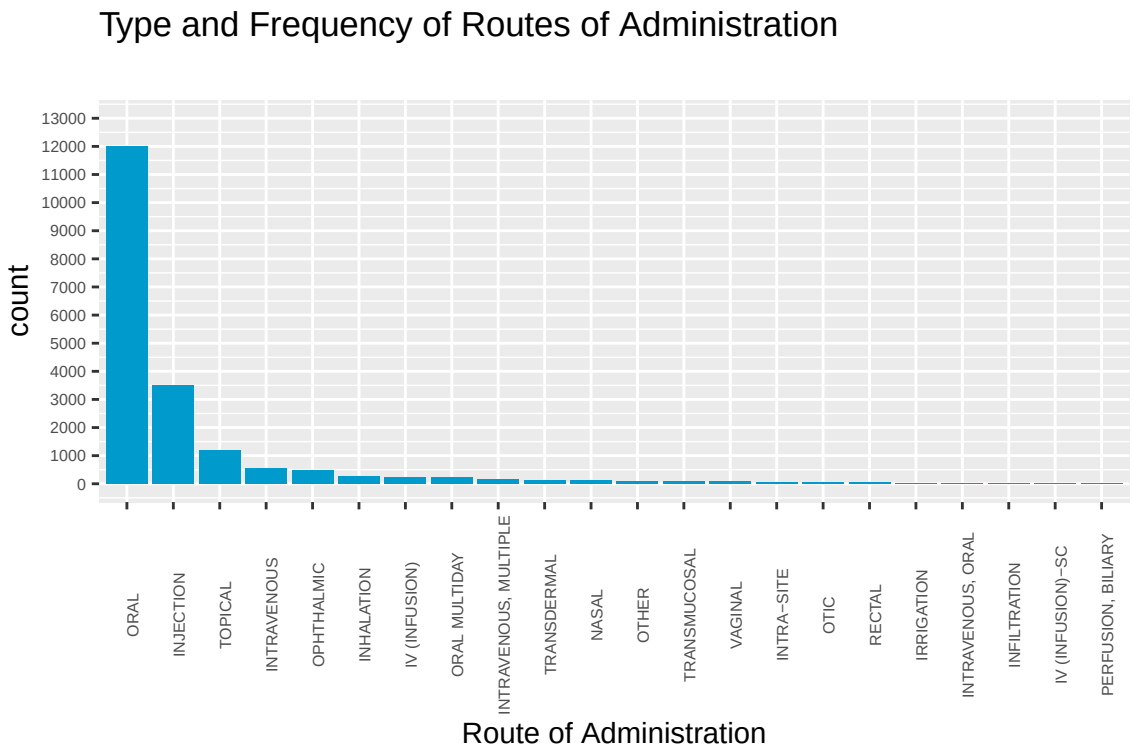


Figure 2: Type and Frequency of Routes of Administration: Products administered orally and through injections are the most common types of pharmaceuticals.

Between 2005 and 2021, almost a fifth of individual approved product strengths were discontinued (figure 6). Within this timespan, the lifetime of commercialized drug products varied from as little as one year to over 15 years (figure 7). This is significant considering patent lifespans of twenty years. Because each new drug product faces competition from multiple generic companies (figure 8), maintenance of marketing exclusivity periods through patents and other regulatory mechanisms becomes a critical element of product lifecycles. However, because generic products may challenge patent validity, the existence of multiple patents does not ensure market monopolies throughout their life. This is aligned with the findings of Carrier et al. (2020) in a review of a cohort of the top 15 litigants between 2009 and 2017. The analysis showed that 80% of cases were likely to be settled out-of-court. Similarly, generic were likely to prevail 41 to 56% of the time when cases were decided on their merits.

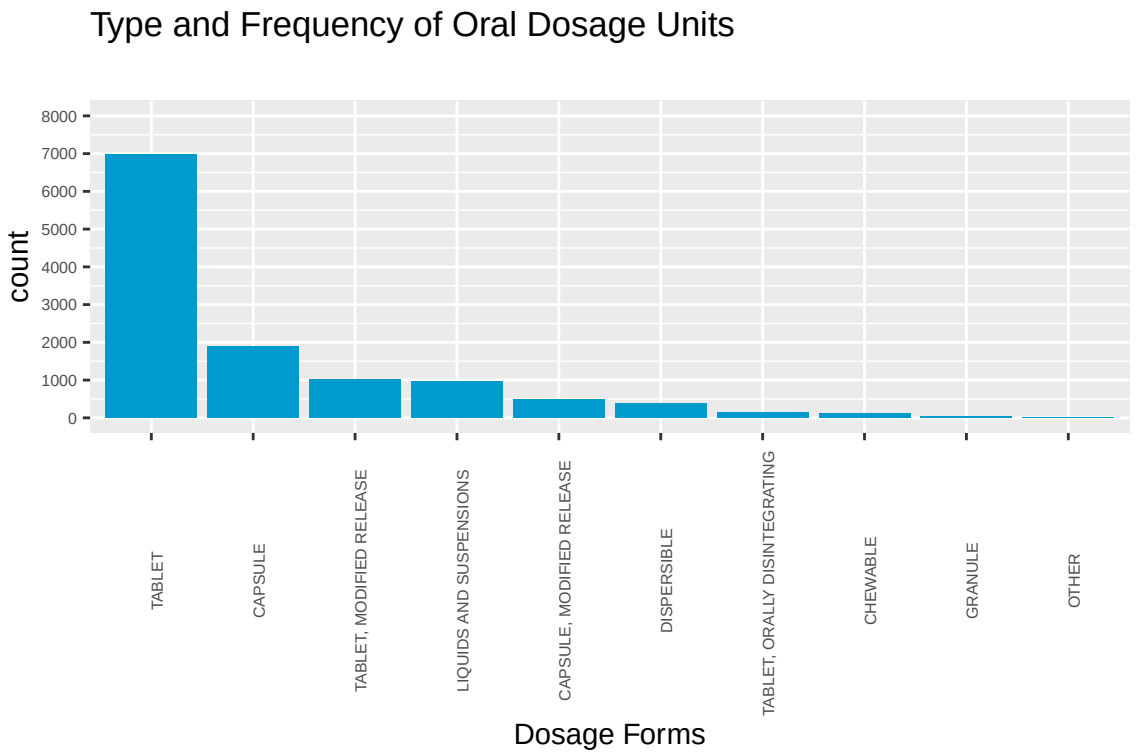


Figure 3: Type and Frequency of Oral Dosage Units: Tablets and capsules without defined drug release characteristics are the most common types of oral dosage form. More complex dosage forms that control the amount of drug delivered over time are a significant subclass.

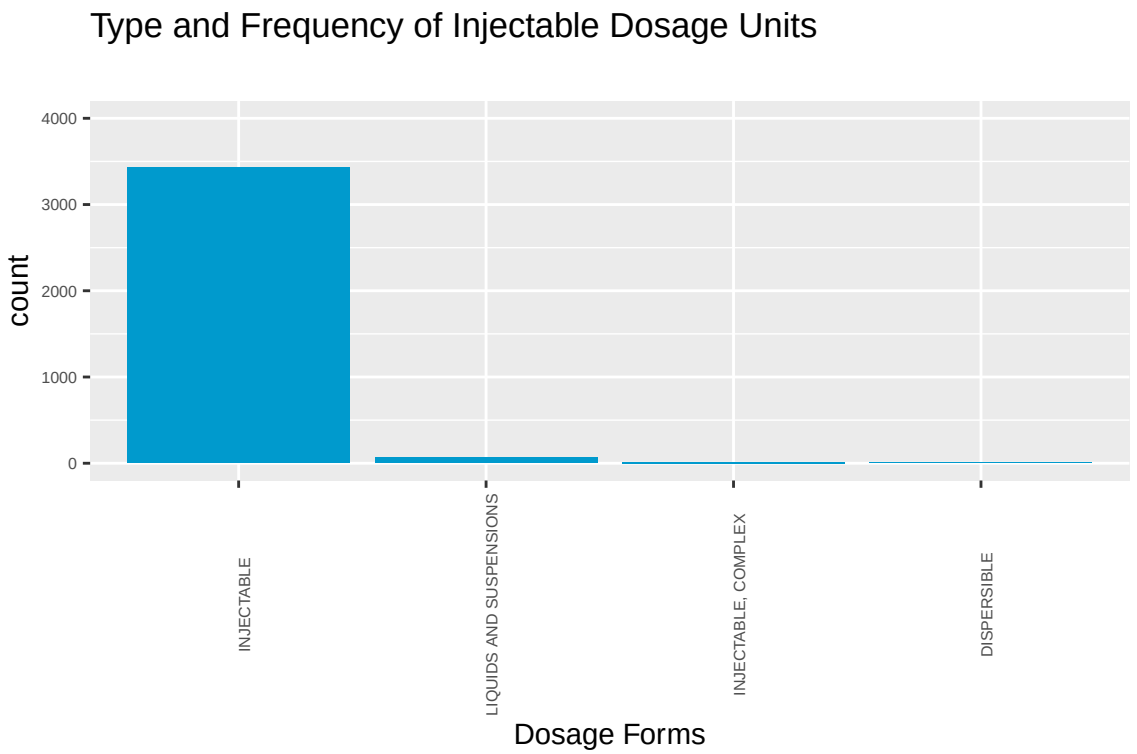


Figure 4: Type and Frequency of Injectable Dosage Units: Injectable drugs have minimal variance in dosage forms. However, more details describe location(s) for administration

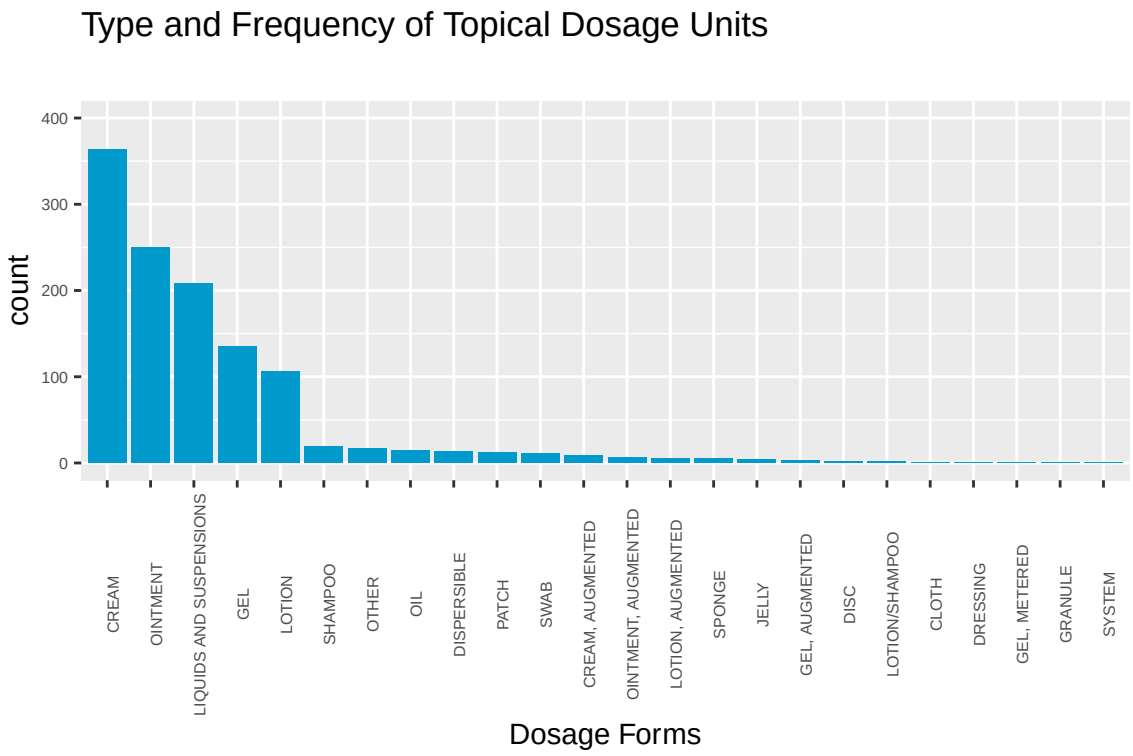


Figure 5: Type and Frequency of Topical Dosage Units: Creams, ointments, liquids, and gels comprise the majority of topical dosage forms and may require differing ingredients or manufacturing processes.

Depending on the terms of settlements, generics may come to mark ahead of patent expiry. Rulings that invalidate a patent may allow other generics to enter the market.

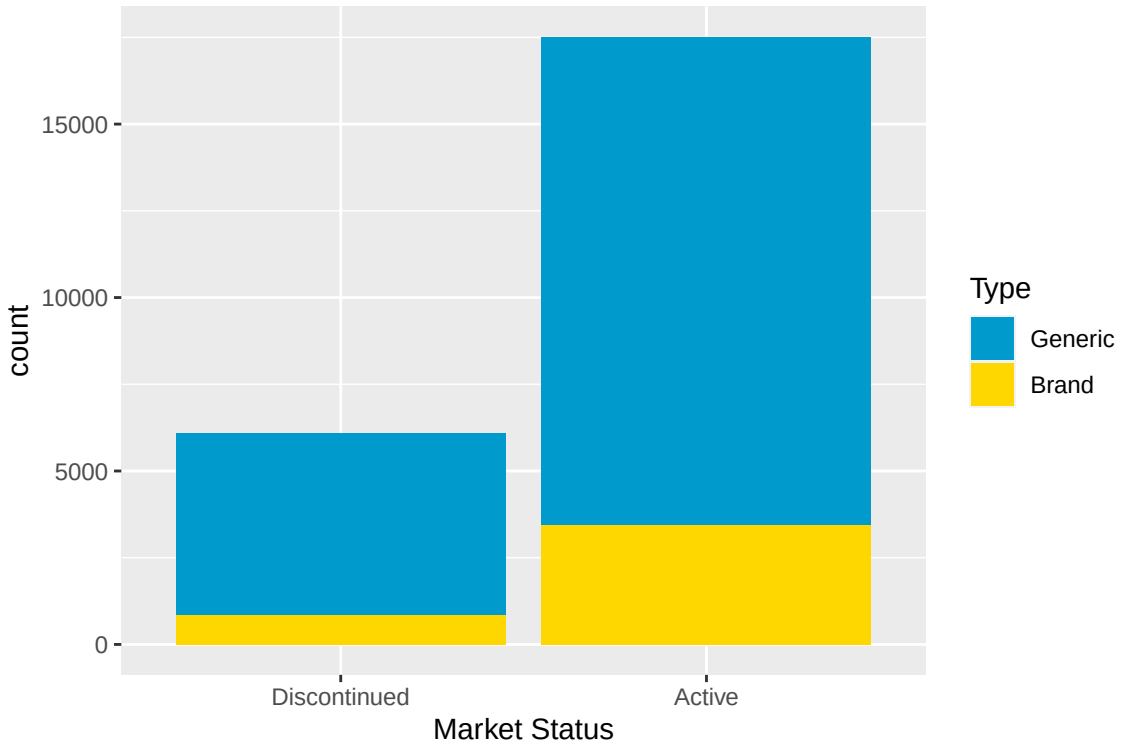


Figure 6: Discontinued vs. Active Products for Generic and Brand Products Approved between 2001-2021 : Around an eigth of all strengths across all approved products have been discontinued since approval.

Although most generic products are launched after the expiry of the first patent (figure 9) and within the first year after the last patent expiry, (figure 10), many products also enter the market before the last patent expires. This can be seen in the elapsed time between product approval and patent expiry, which is less than expected from a typical patent lifespan of 20 years (figure 11). Since 1983, the amount of time between generic and branded product approval has increased while the amount of time between branded drug approval and patent expiry has remained similar (figures 12 and 13). However, it is not apparent if this is correlated with or caused by increased patents filings per product (figure 15). These patents may cover the therapeutic drug substance, the dosage form of the drug product, or

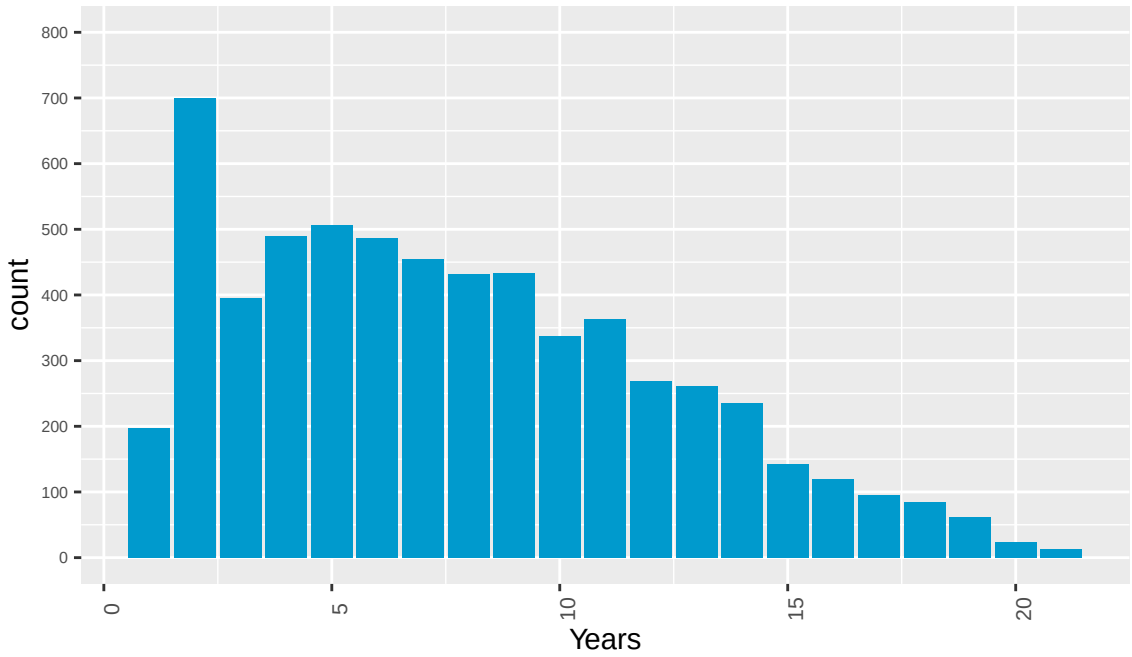


Figure 7: Years Marketed until Discontinuance for Products Approved between 2001 and 2021 : Many products were discontinued within 2 to 6 years of approval.

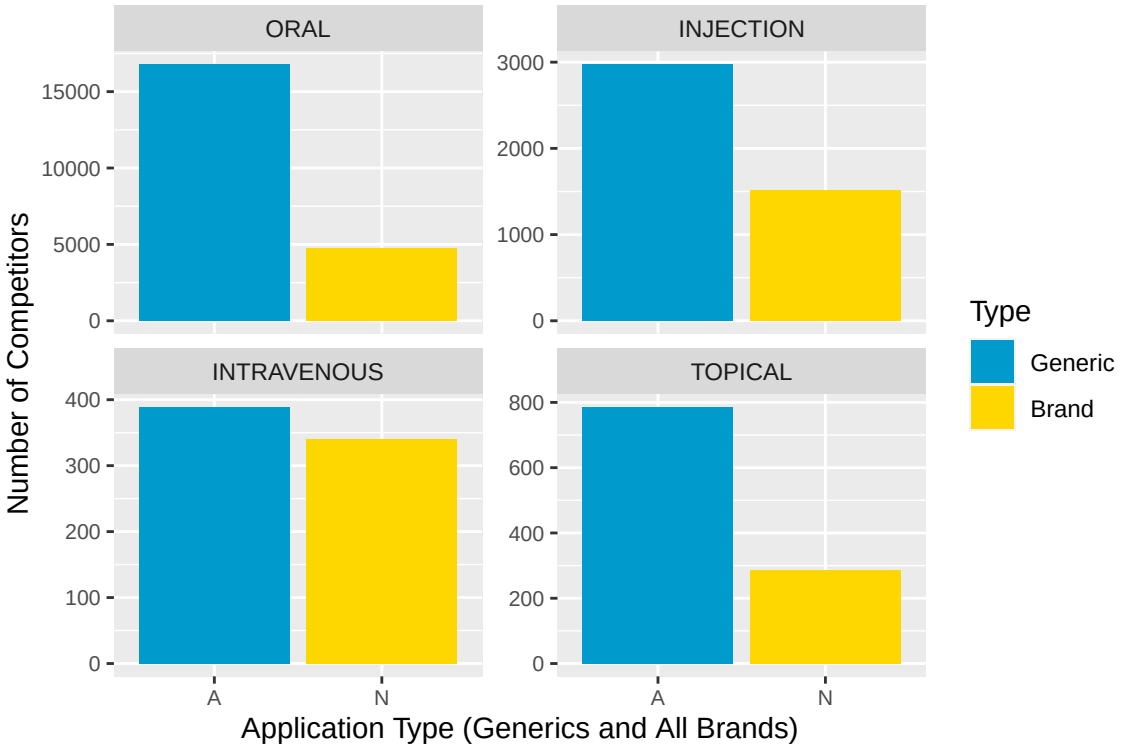


Figure 8: Generic vs. Branded Competition by Route Injectable and intravenous products have less generic competition compared to oral and topical products. This may reflect higher bariers to entry due to sterile product requirements and mathcing the delivery devices

the method of use for treating a specific disease state. The specific product use codes accompanying each patent provide a snapshot of the intended therapeutic treatment and disease state. An analysis of themes within the use codes is presented in the following section.

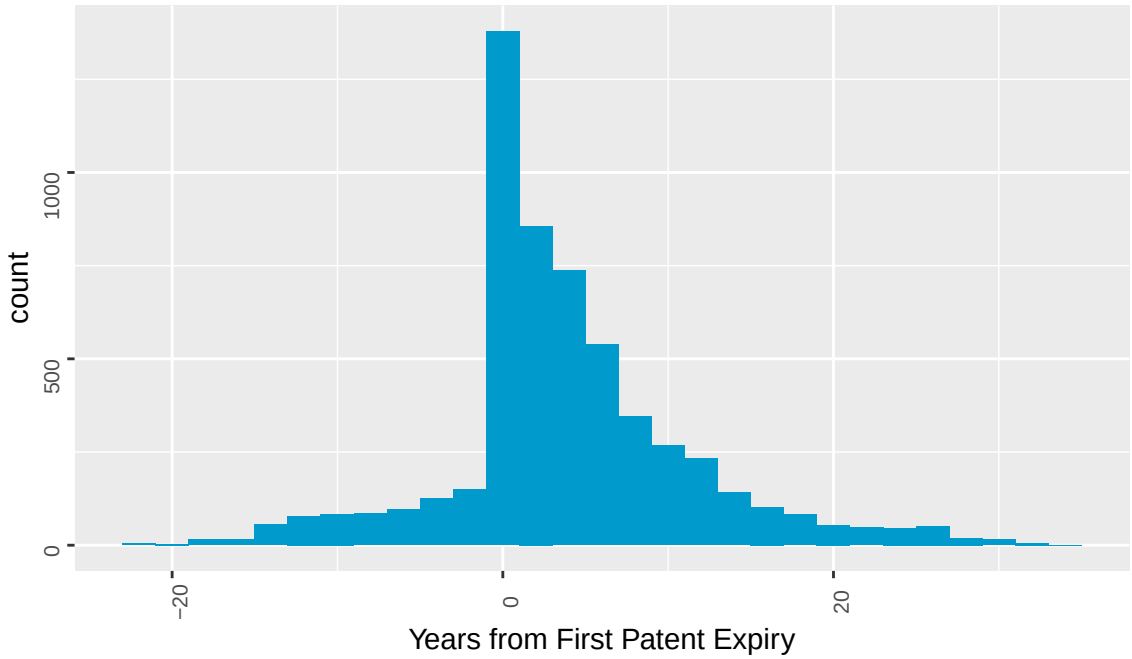


Figure 9: Generic Approval vs. First Patent Expiry, 1983-2021: A significant amount of generic products are approved within the first year of patent expiry. However, some products are approved before the first patent expires.

A cadence of 3 to 10 approvals per year distinguishes some of the top companies such as Teva, Mylan, Aurobindo, and Lupin (figure 16). Most of these products reference a brand product possessing patents (figure 17). Per the Patent Term Restoration Act, the first sponsor of a generic product to submit an application and prevail in litigation for patent infringement receives 180 days to market their product before the FDA may approve any additional generics. While most comapnies have a low percentage of products each year that receive approval before the first patent expires (figure 18), several companies frequently receive approval before the last patent expires (figure 19). Remarkably, several established

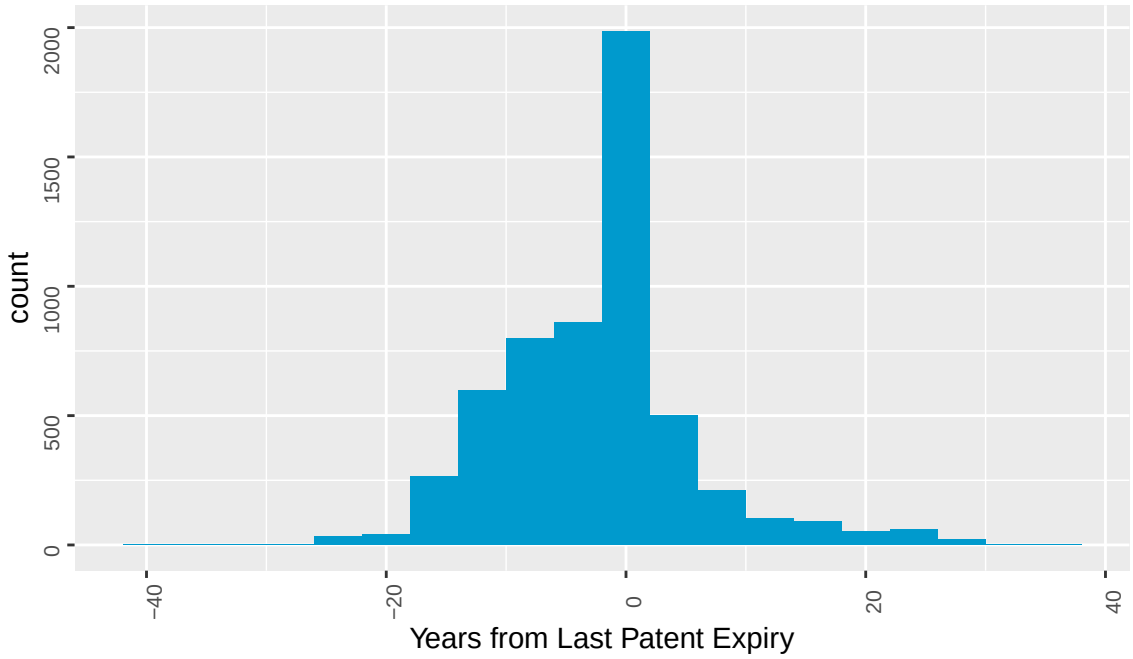


Figure 10: Generic Approval vs. Last Patent Expiry, 1983-2021 Including Products Approved before First Patent Expiry: While most products are approved after the last patent expires, a significant numebr are approved within 10 years of the last patent expiry. This may reflect the effect of infringement settlements, invalidity findings, or other agreements between brand and generic companies.

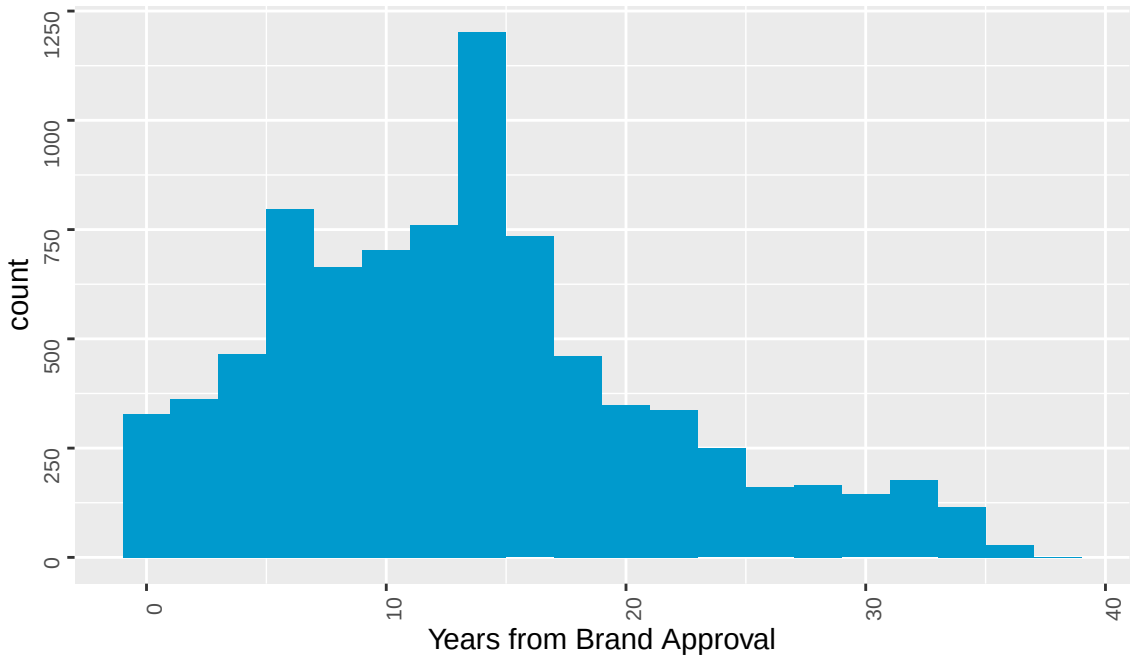


Figure 11: Generic Approval vs. Brand Approval, 1983-2021: Product approvals are somewhat earlier than expected considering the 20 year patent life. The effect of exclusivity periods is also seen by the steep increase in approvals after five years.

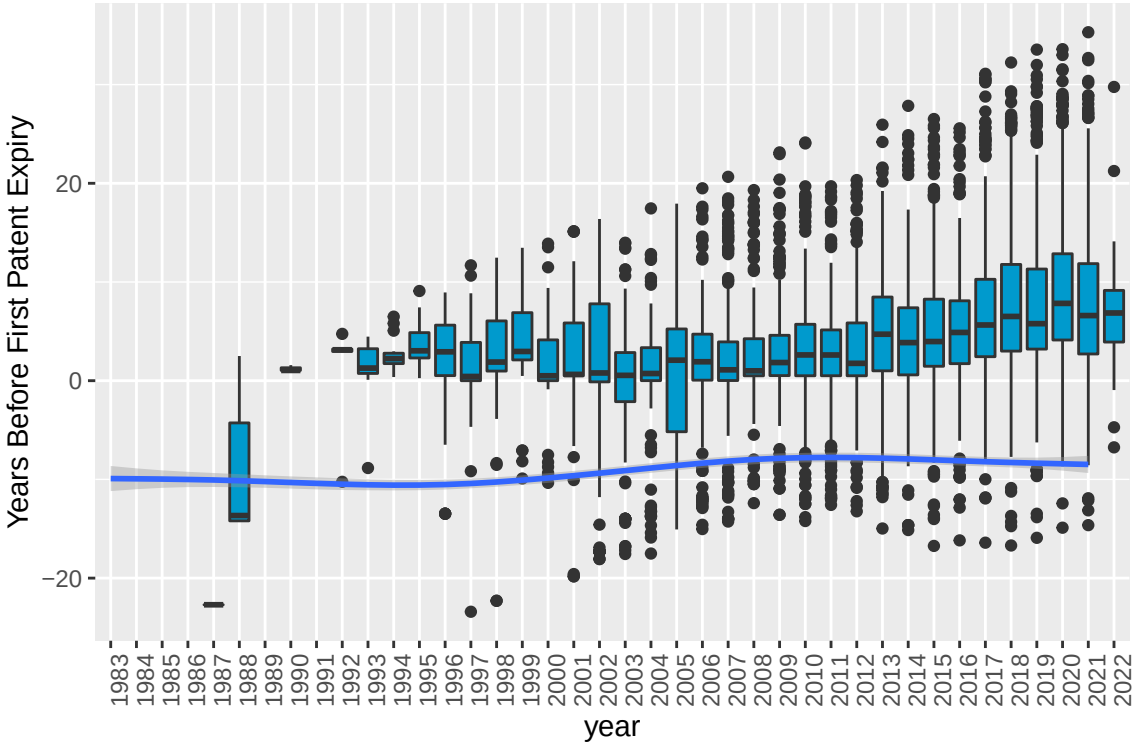


Figure 12: Generic Approval Time Prior to First Patent Expiry The median time between generic approval and first branded product patent expiry has increased and occurs after the patent has expired. In most cases, only the upper quartile of generic products are approved before patent expiry. The time of brand product approval has remained almost constant from 2005 - 2021

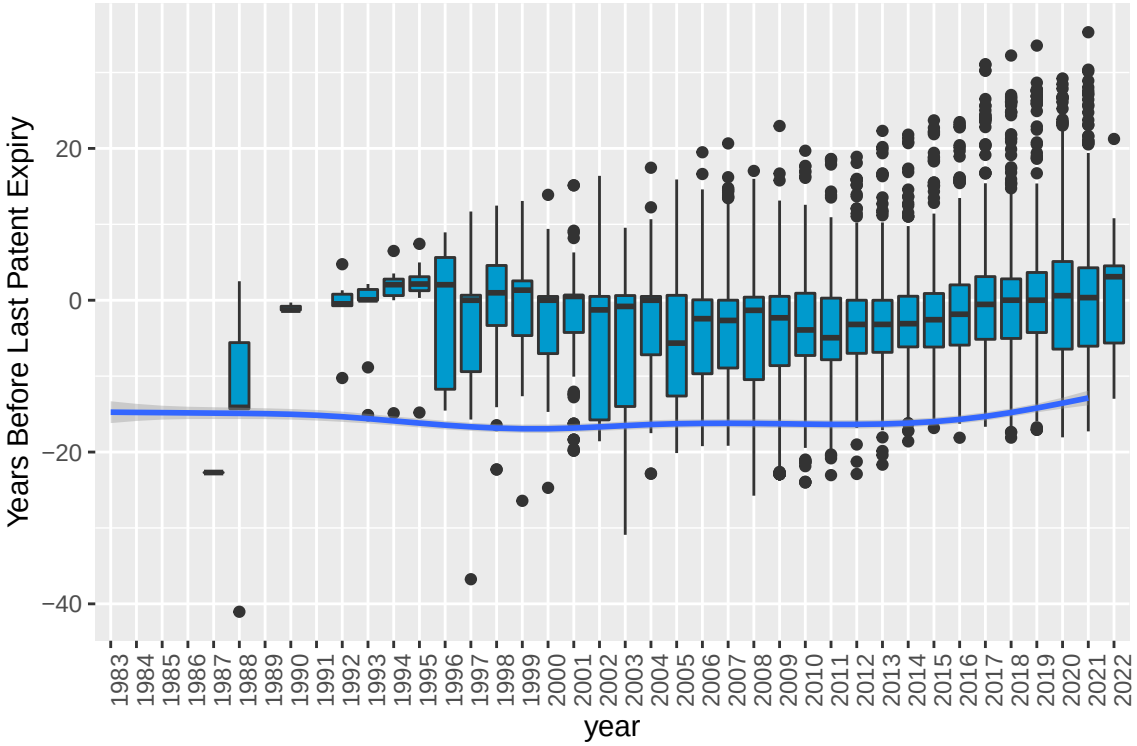


Figure 13: Generic Approval Time Prior to Last Patent Expiry The median time between generic approval and last branded product patent expiry has fluctuated but has remained within a similar range. In most years, the median approval time occurs before the patent has expired. The mean time of brand product approval has remained almost constant from 2005 - 2021

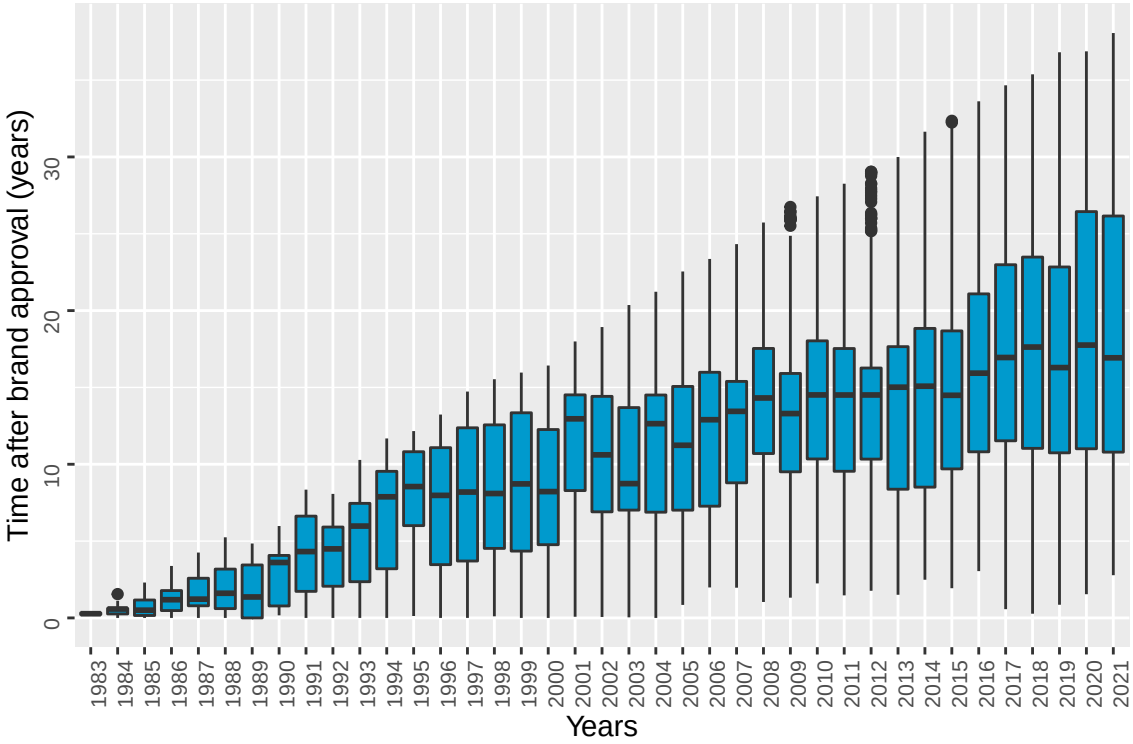


Figure 14: Generic Approval vs Brand Approval Date: Generic product approval time has ocurred increasingly longer compared to the intial brand product approval.

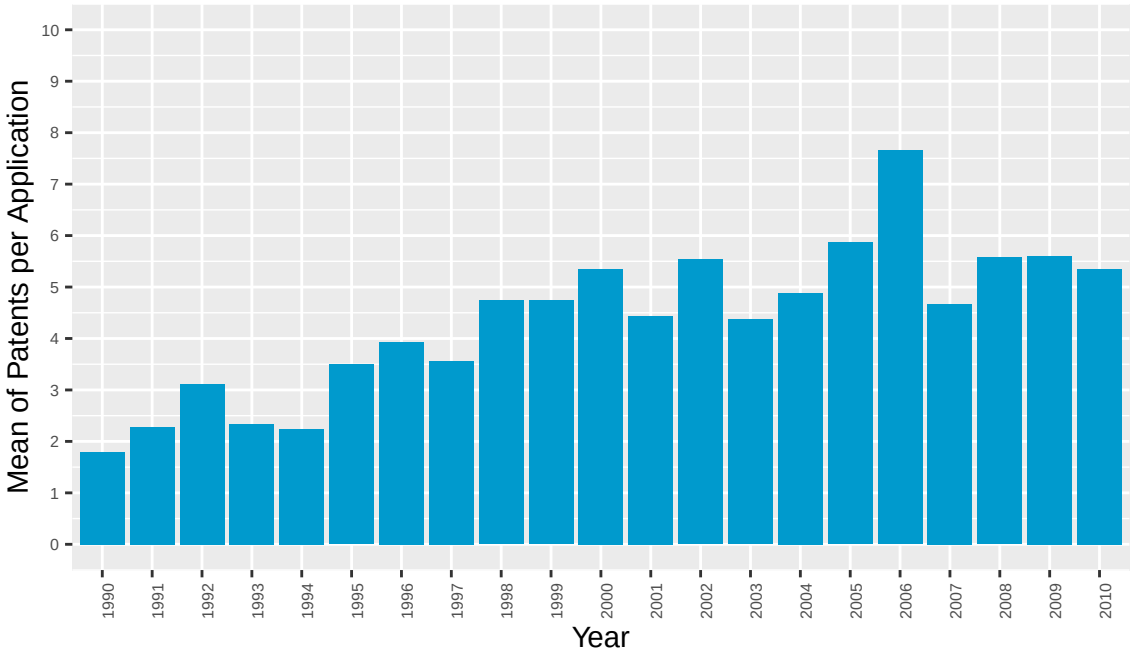


Figure 15: Mean Patents per NDA: The number of patents per application has steadily increased between 1990 and 2010.

generic companies like Mylan and Teva have achieved 50% or more products per year as first-to-file opportunities (figure 20). Due to the higher sale prices associated with the initial generic products in a market, this consistency may provide a sales price advantage. Dosage form specialization appears in the top companies. For example, the companies with the most injectable products, Hospira and Sicor, do not appear on the list of companies with the most orally administered products (figures 21 and 22).

The market analysis established several features as potential predictors and defining characteristics of competition. From a generic-product perspective, key competitive windows exist before, between, and after the expiration of the first and last patents. A question exists as to whether generic entry within any of these categories can be predicted by the number of patents, patent use category, specific dosage forms, or even specific formulation components. Because the regulatory paradigm changed significantly during the 1980s, market analysis should limit data from 1990 onwards. Consideration of brand approval vs patent expiry suggests that the data set should explore only brand products approved from 2010 or earlier. This will allow adequate time for generic competitors to enter the market ahead of patent expiry.

5.2 Topics in Usage Codes

Topic modeling was performed on patent usage codes to determine both underlying structure and feasibility of including topics in predictive models. A summary of topics compared to unique product applications is presented in figure 23. Representative topic models are presented in figures 24 and 25. As the number of topics increased, the number of products having more than one relevant term increased while the number of products with only a single matching term decreased. This indicated that higher topic counts possessed greater overlap in terms. This is undesirable because it results in less distinction between textual differences among coded documents.

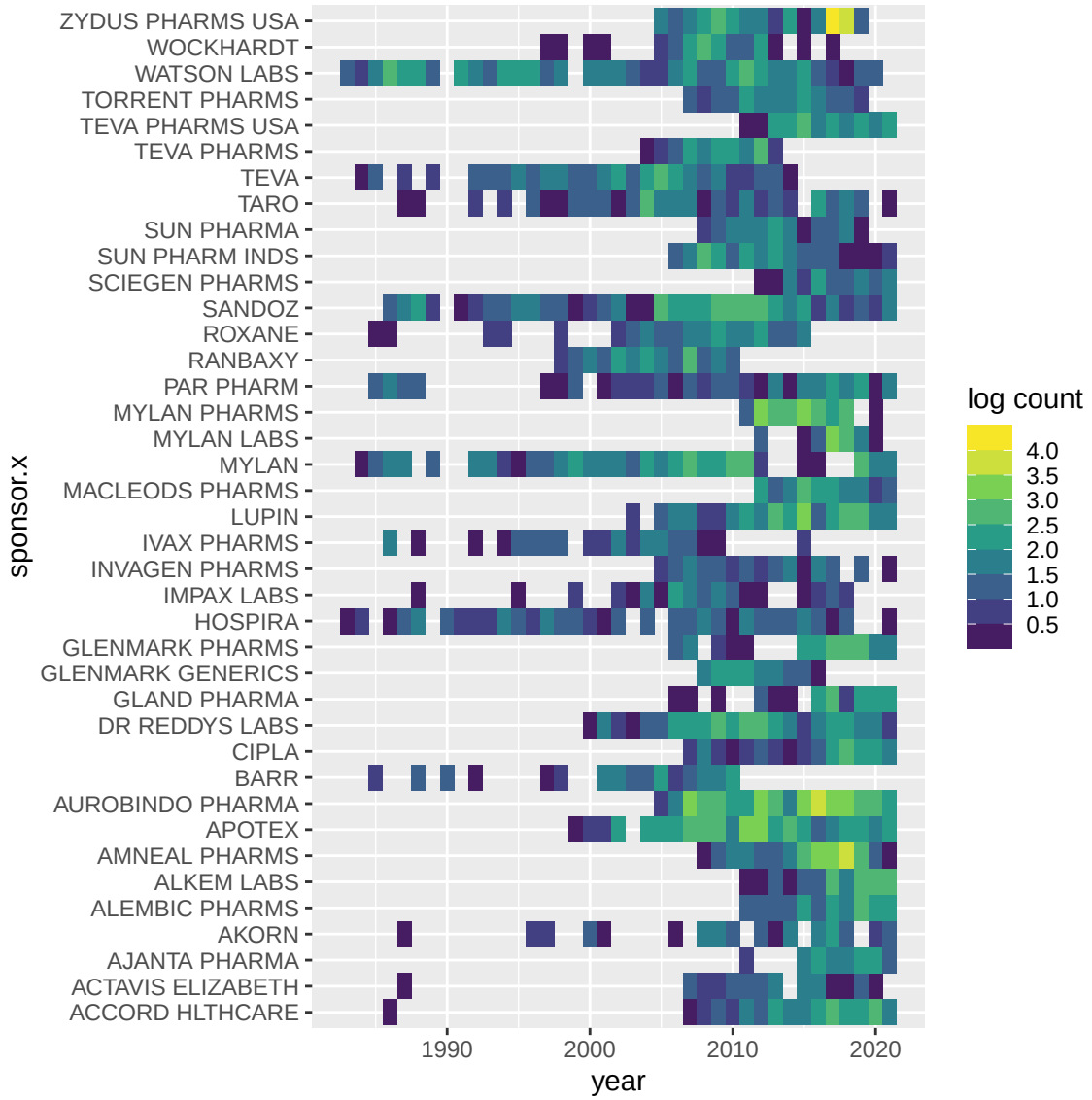


Figure 16: Number of Products per Year (Top 40 Companies by number of products): Top generic companies maintain a cadence of 3 - 10 approvals each year. While rare, some companies have achieved between 20 and 60 approvals in a year.

Figure 17: Percentage of Company's Total New Products that Referenced a Brand Product with Patents Most generic products after 1985 referenced a brand product with patent protections

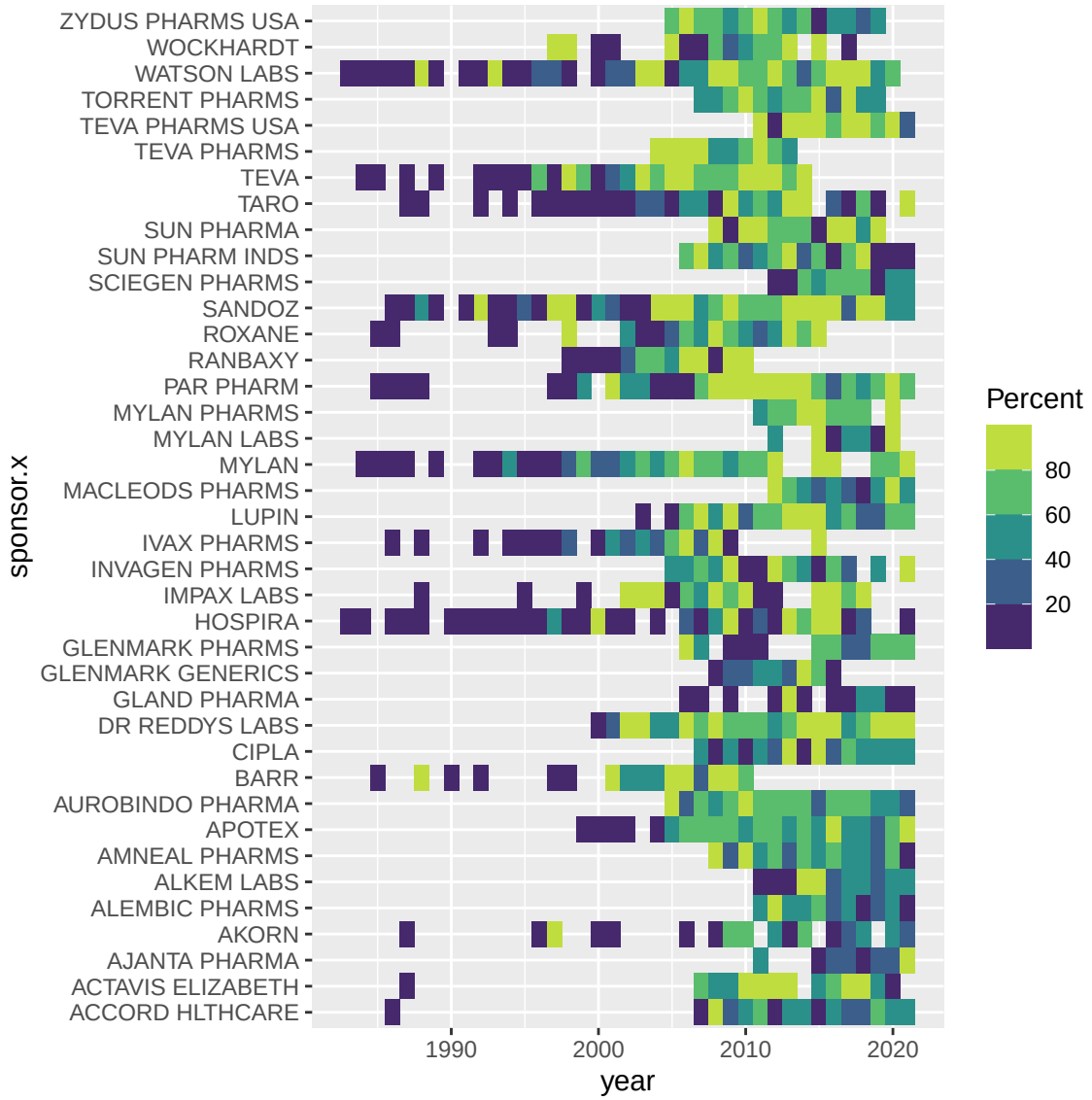


Figure 18: Generic Products Approved before Expiry of Last Brand Patent as a Percentage of Company's Products Referencing Patented Products: For many leading generic companies, more than half of the products each year are approved before the last brand patent expires.

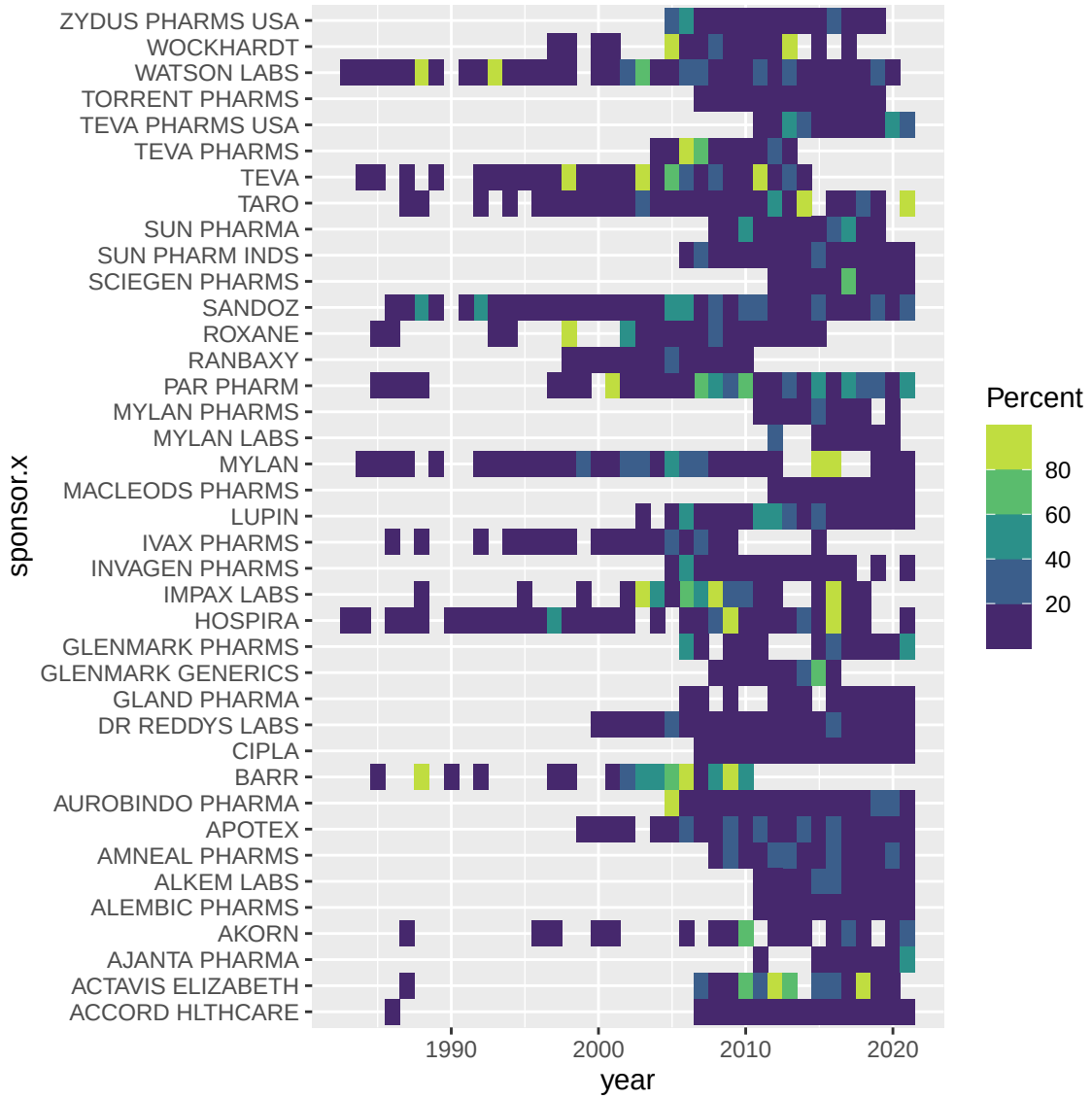


Figure 19: Generic Products Approved before Expiry of First Brand Patent as a Percentage of Company's Products Referencing Patented Products (Top 40 Companies by Product Count): Approval of products before the first patent expires does not occur frequently. A small number of companies such as Barr Par, and Actavis frequenlty had between 20 and 40 percent of products each year approved before the first brand patent expiry

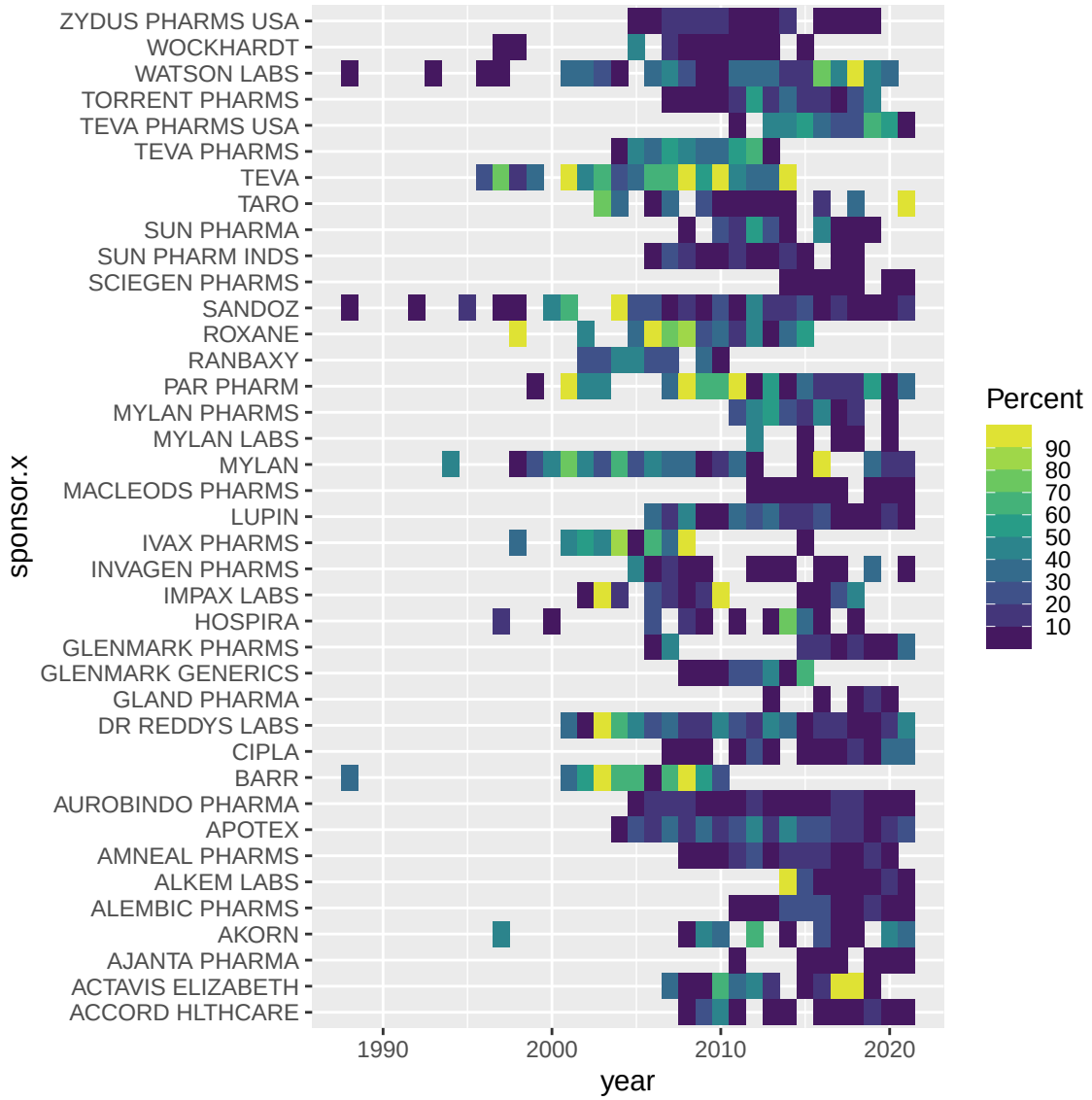


Figure 20: First Approved Generic Products when Brand has Patents Top 40 Companies by Product Count): Companies like Teva, Par, and Mylan are remarkable for the high percentage of products receiving the first approval prior to expiration of the last branded patent.

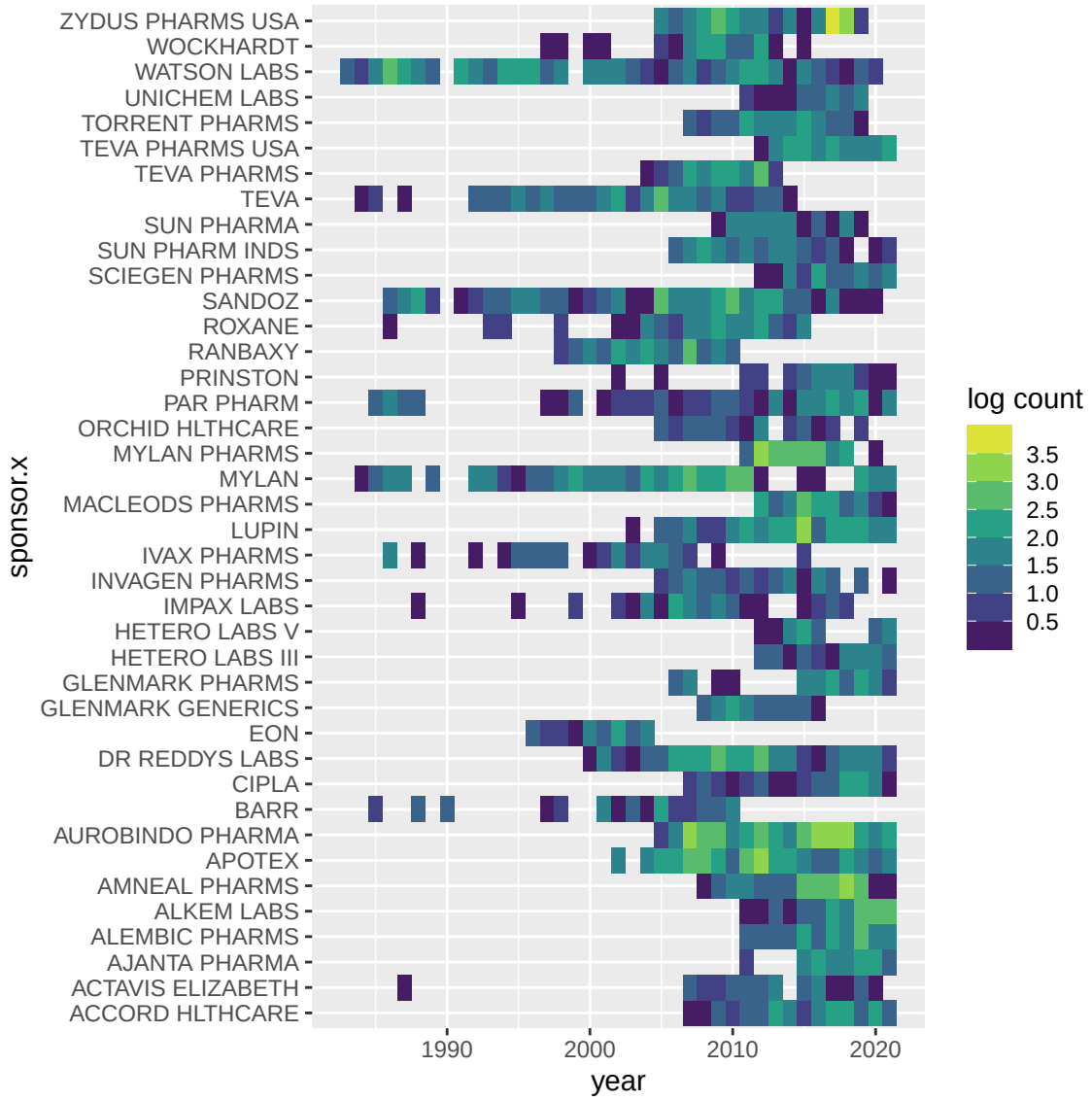


Figure 21: Number of Oral Products per Year (Top 40 Companies by number of products): Numerous companies receive approval for multiple oral products each year.

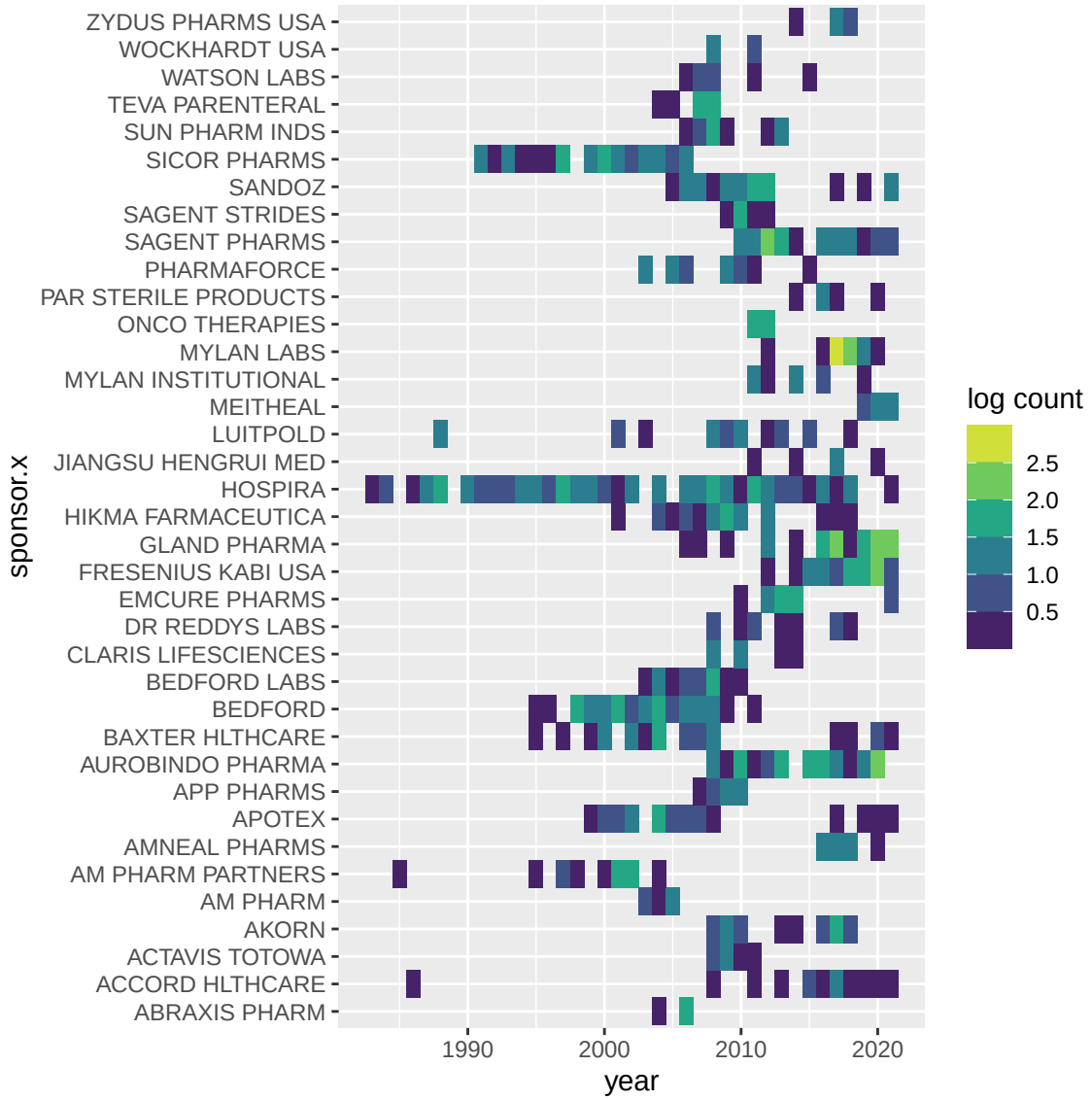
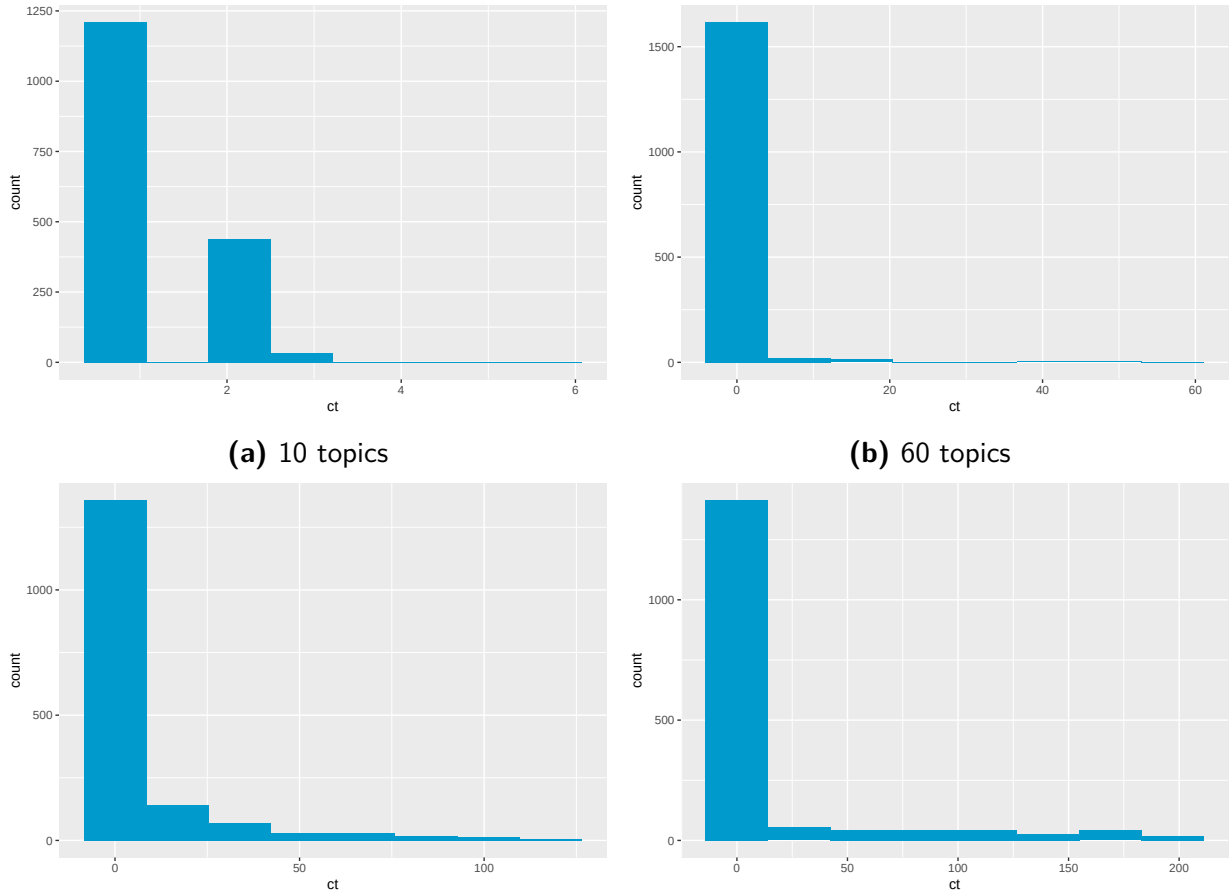


Figure 22: Number of Injectable Products per Year (Top 40 Companies by number of products and administration rotue): Few companies receive yearly approval for injectable products and other comapnies have approved injectable products as part of a larger portfolio of dosage forms



(a) 10 topics

(b) 60 topics

(c) 120 topics

(d) 200 topics

Figure 23: Number of Topics with Partial Matches (gt0.2) to Use Codes

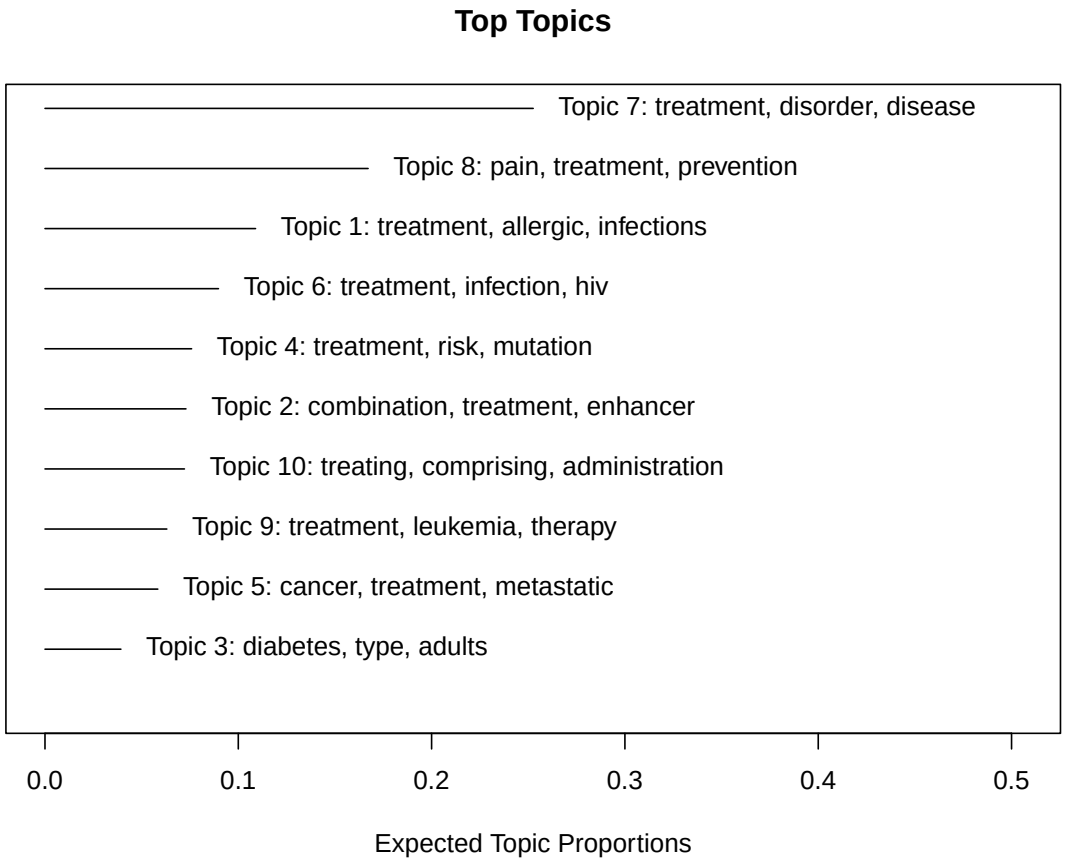


Figure 24: Use Code Topics with Terms for n = 10

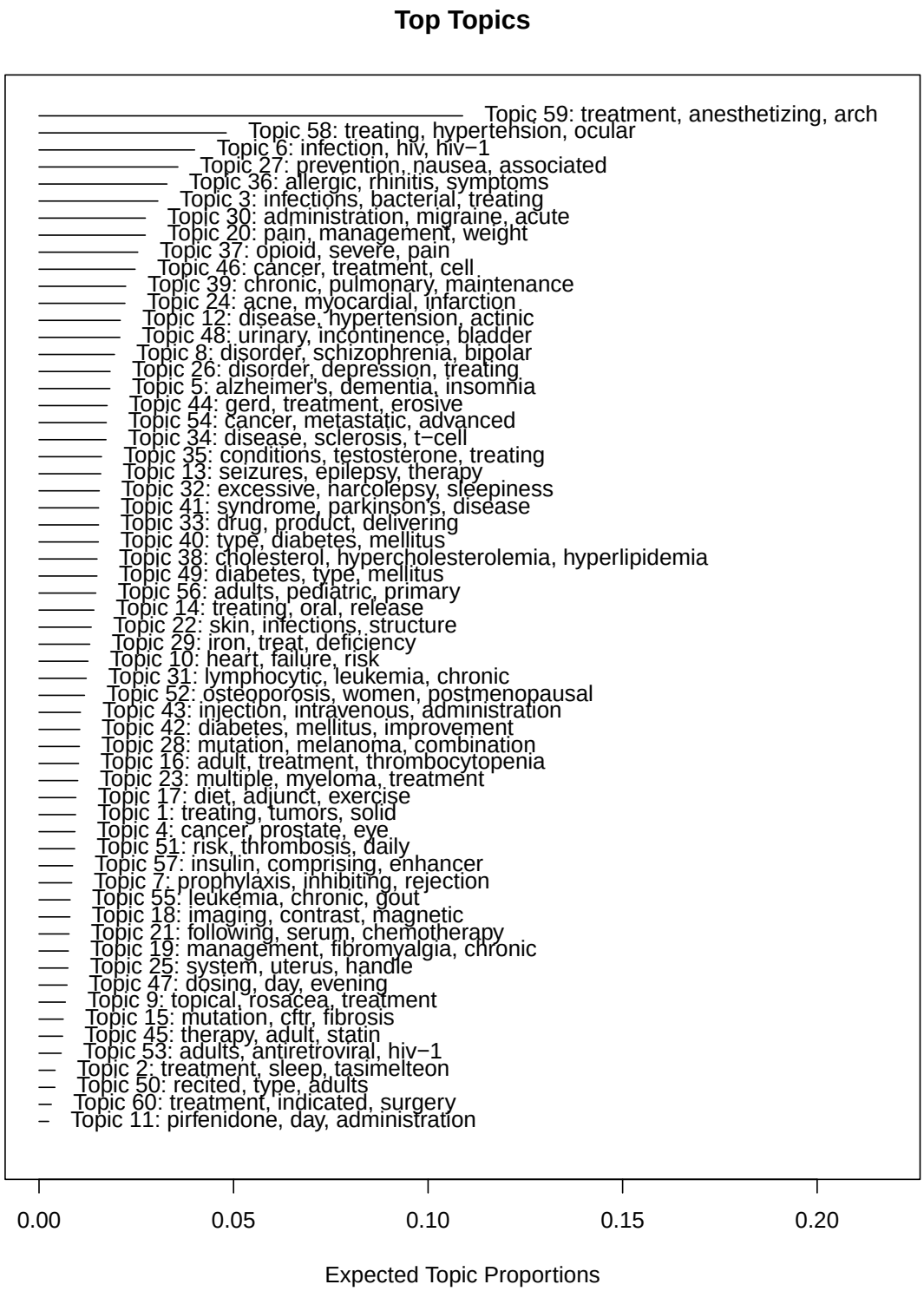


Figure 25: Use Code Topics with Terms for n = 60

Reviewing the topic models, selecting 10 topics differentiated terms among broad categories such as “chronic”, “disease”, ”allergic”, and ”pain”. As topics increased to 50 and 100, categories focus on more specific terms including the therapeutic drug substance and specific disease states. Even with 60 topics, more technical terms like “mutation”, “olazepine, and “lymphocytic” are observed frequently in the topic model, indicating that the model is beginning to fit distinct use codes. Beyond this, topics became increasingly specialized and were considered impractical for defining broad categories useful for feature identification in a predictive model.

For this study, dimensionality reduction of usage codes will supplement quantitative measures of drug product complexity. The appropriate number of topics will ultimately be defined based on whether the topics improve the predictive model accuracy. As such, the predictive model will be assessed using 10, 60, and 120 terms.

6 Results

In all studied conditions, models achieved significant accuracy ($p < 0.001$) on training sets. Overall accuracy ranged from 0.76 to 1.0 vs. a prevalence of 0.5. Although most models decreased in accuracy when analyzing the test sets, detection accuracy was significantly higher than prevalence in numerous cases.

For the multiple administration route data, the baseline patent information was sufficient to predict generic competition before last patent expiry (Approved pre-last) using all three classification algorithms (Table 7). The addition of packaging or ingredient information did not improve prediction accuracy. The inclusion of patent use categories achieved similar accuracy for predicting competition before the last patent expiry with all three classification algorithms. However, the boosted gradient method produced significant predictions for competition in all three categories with the inclusion of 60 patent use code topics (Table 8). This includes a significant accuracy of 83% for competition before the first patent expiry.

Competition before first patent expiry and after last patent expiry could be predicted with boosted gradient methods with a combined data set of patents, use codes reduced to 60 topics, and packaging information (Table 9). These models achieved 81% to 76% accuracy compared to the NIR of 75% to 61% ($p < 0.05$).

Table 7: Test Set Accuracy for Multiple Adminstration Routes with Baseline Patent Information:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.772	0.730	0.792	0.995
		pre-last	0.688	0.672	0.536	0.000
		post-last	0.710	0.707	0.633	0.004
	xgboost	pre-first	0.975	0.812	0.792	0.216
		pre-last	0.967	0.686	0.536	0.000
		post-last	0.971	0.697	0.633	0.012
	svm_model	pre-first	0.813	0.727	0.792	0.997
		pre-last	0.745	0.710	0.536	0.000
		post-last	0.775	0.663	0.633	0.152
b	rpart	pre-first	0.824	0.749	0.749	0.538
		pre-last	0.709	0.665	0.635	0.212
		post-last	0.760	0.658	0.609	0.085
	xgboost	pre-first	0.992	0.714	0.749	0.886
		pre-last	0.998	0.749	0.635	0.000
		post-last	0.991	0.693	0.609	0.008
	svm_model	pre-first	0.927	0.719	0.749	0.853
		pre-last	0.905	0.729	0.635	0.003
		post-last	0.922	0.668	0.609	0.048
c	rpart	pre-first	0.851	0.770	0.751	0.291
		pre-last	0.813	0.694	0.627	0.026
		post-last	0.761	0.660	0.612	0.088
	xgboost	pre-first	1.000	0.770	0.751	0.291
		pre-last	1.000	0.713	0.627	0.006
		post-last	0.990	0.761	0.612	0.000
	svm_model	pre-first	0.874	0.737	0.751	0.715
		pre-last	0.847	0.708	0.627	0.008
		post-last	0.839	0.732	0.612	0.000

[a] Base Patents

[b] Base and Ingredients

[c] Base and Packaging

Table 8: Test Set Accuracy for Multiple Adminstration Routes with Patent Use Codes:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.805	0.758	0.792	0.932
		pre-last	0.690	0.679	0.536	0.000
		post-last	0.751	0.707	0.633	0.004
	xgboost	pre-first	0.995	0.819	0.792	0.140
		pre-last	0.981	0.669	0.536	0.000
		post-last	0.947	0.633	0.633	0.526
	svm_model	pre-first	0.862	0.775	0.792	0.787
		pre-last	0.801	0.734	0.536	0.000
		post-last	0.841	0.629	0.633	0.574
b	rpart	pre-first	0.848	0.758	0.792	0.932
		pre-last	0.650	0.648	0.536	0.000
		post-last	0.787	0.673	0.633	0.081
	xgboost	pre-first	0.989	0.833	0.792	0.046
		pre-last	0.992	0.689	0.536	0.000
		post-last	0.970	0.694	0.633	0.016
	svm_model	pre-first	0.887	0.785	0.792	0.645
		pre-last	0.832	0.710	0.536	0.000
		post-last	0.836	0.599	0.633	0.897
c	rpart	pre-first	0.824	0.741	0.792	0.985
		pre-last	0.678	0.676	0.536	0.000
		post-last	0.776	0.687	0.633	0.029
	xgboost	pre-first	0.999	0.826	0.792	0.084
		pre-last	0.988	0.717	0.536	0.000
		post-last	0.987	0.639	0.633	0.430
	svm_model	pre-first	0.889	0.778	0.792	0.744
		pre-last	0.829	0.720	0.536	0.000
		post-last	0.826	0.544	0.633	0.999

[a] 10 Topics

[b] 60 Topics

[c] 120 Topics

Table 9: Test Set Accuracy for Multiple Adminstration Routes with 60 Patent Use Codes, Packaging, and Ingredient Features:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.882	0.766	0.751	0.349
		pre-last	0.778	0.675	0.627	0.086
		post-last	0.801	0.684	0.612	0.019
	xgboost	pre-first	1.000	0.809	0.751	0.030
		pre-last	0.979	0.636	0.627	0.417
		post-last	0.988	0.761	0.612	0.000
	svm_model	pre-first	0.944	0.718	0.751	0.884
		pre-last	0.916	0.732	0.627	0.001
		post-last	0.884	0.656	0.612	0.113
b	rpart	pre-first	0.868	0.690	0.749	0.977
		pre-last	0.841	0.704	0.635	0.023
		post-last	0.821	0.589	0.609	0.743
	xgboost	pre-first	1.000	0.798	0.749	0.060
		pre-last	0.995	0.655	0.635	0.307
		post-last	1.000	0.639	0.609	0.214
	svm_model	pre-first	0.966	0.665	0.749	0.997
		pre-last	0.949	0.734	0.635	0.002
		post-last	0.928	0.639	0.609	0.214
c	rpart	pre-first	0.868	0.690	0.749	0.977
		pre-last	0.814	0.665	0.635	0.212
		post-last	0.803	0.594	0.609	0.694
	xgboost	pre-first	0.997	0.734	0.749	0.717
		pre-last	0.982	0.650	0.635	0.360
		post-last	1.000	0.718	0.609	0.001
	svm_model	pre-first	0.961	0.695	0.749	0.967
		pre-last	0.958	0.680	0.635	0.107
		post-last	0.931	0.634	0.609	0.259

Table 9: Test Set Accuracy for Multiple Adminstration Routes with 60 Patent Use Codes, Packaging, and Ingredient Features: *(continued)*

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
d	rpart	pre-first	0.834	0.744	0.749	0.601
		pre-last	0.784	0.700	0.635	0.033
		post-last	0.721	0.683	0.609	0.017
	xgboost	pre-first	0.997	0.724	0.749	0.814
		pre-last	0.997	0.670	0.635	0.172
		post-last	0.997	0.703	0.609	0.003
	svm_model	pre-first	0.927	0.734	0.749	0.717
		pre-last	0.920	0.744	0.635	0.001
		post-last	0.929	0.693	0.609	0.008

[a] Base Patents, Packaging, and Use Topics
[b] Patents, Ingredients, Packaging, and Use Topics
[c] Patents, Ingredients, and Use Topics
[d] Patents, Ingredients, and Packaging

Using the baseline patent information to predict competition before last patent expiry, both the decision tree and boosted gradient methods assigned high weights to brand product application time prior to patent expiry, as shown in the feature importance graph (figure 27) and the decision tree (figure 26). From the decision tree, the largest classifiers based on a sum of terminal leafs are divided as:

- No competition before last patent entry (51%) when brand product was approved later than the mean time from last patent expiry;
- Competition before last patent expiry (25%) when the dosage form is a non-modified release tablet or is an injectable dosage form (injectable = 1) approved earlier than an arbitrary median point from last patent expiry (>=0.43);
- Competition before patent entry (16%) when dosage form is not a tablet (tablet = 0) approved earlier than an arbitrary median point from last patent

expiry (>=0.43)

The boosted gradient method places high feature importance on timing before first and last patent expiry (brand_pre_first_pat, brand_pre_last_pat), while total patent count (pat_ct), total drug product patents (DP_ct) , total drug substance patents (DS_ct), and previous product count (brand_prev_prod_ct) are also among key features. Looking at the SHAP plots, modified release tablets and capsules and general tablets and capsules suggest increased competition, while injectable, liquids and suspension, and dispersible dosage forms suggest a decreased likelihood of competition. Higher values for brand approval time before first and last patent expiry and previous product count decrease the likelihood of competition, while higher numbers of patents increase the likelihood of competition.

From a practical standpoint, the decision tree nodes and boosted gradient feature weights for brand approval relative to patent expiry and impact of patent count seem counter-intuitive. They suggest that branded products entering the market further ahead of patent expiry are less prone to generic competition. One would assume that this would allow longer development times for generic competitors, giving them an opportunity to find a pathway to market. In contast, considering that increased patent count decreases competition, the association of higher patent counts with less competition approximates industry behavior seen from the market analysis. More patents are associated with longer generic approval times.

The dosage forms highlighted from the SHAP plot bear further consideration. The formulations for injectable dosage forms may have varying complexity, ranging from simple solutions to products that dissolve over multiple months. Unlike most oral dosage products, delivery device engineering may need to accompany formulation development, and sterility assurance for the final product may be required (Lambert, 2010). These features may inherently decrease competition. In contrast, while modified release dosage forms have greater development and manufacturing complexity than general tablets and capsules, manufacturing and

development capabilities for solid dosage forms are relatively common. Within a model for multiple dosage forms, the polarity of these classifiers appears reasonable.

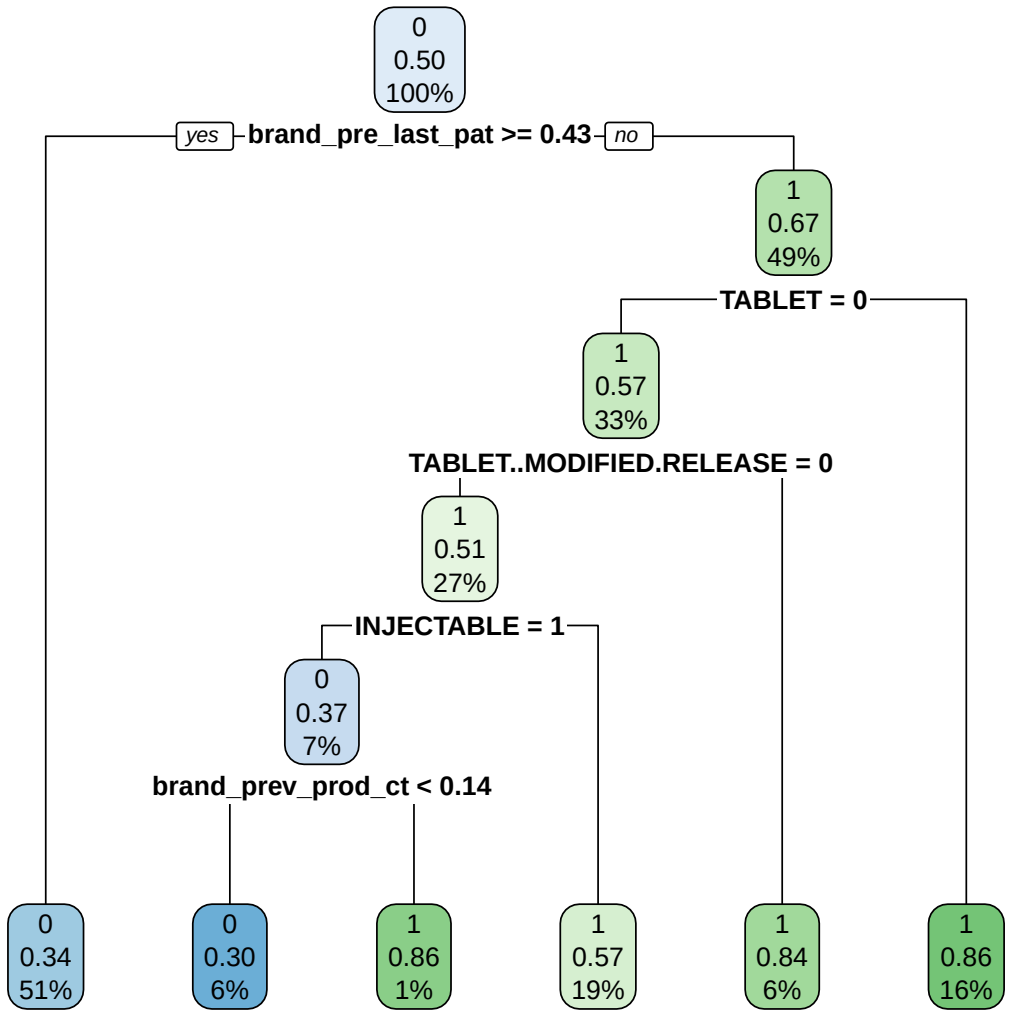


Figure 26: Baseline RPART Model, Prediction of Competition Before Last Patent Expiry, Multiple Administration Routes

Looking at competition before and after last patent expiry, the top three feature importance weights include time before first and last patent expiry and previous product count (figures 29 and 30). The predictor for competition before last patent expiry has a broader range of values associated with the features, and the drug substance, pediatric exclusivity, drug

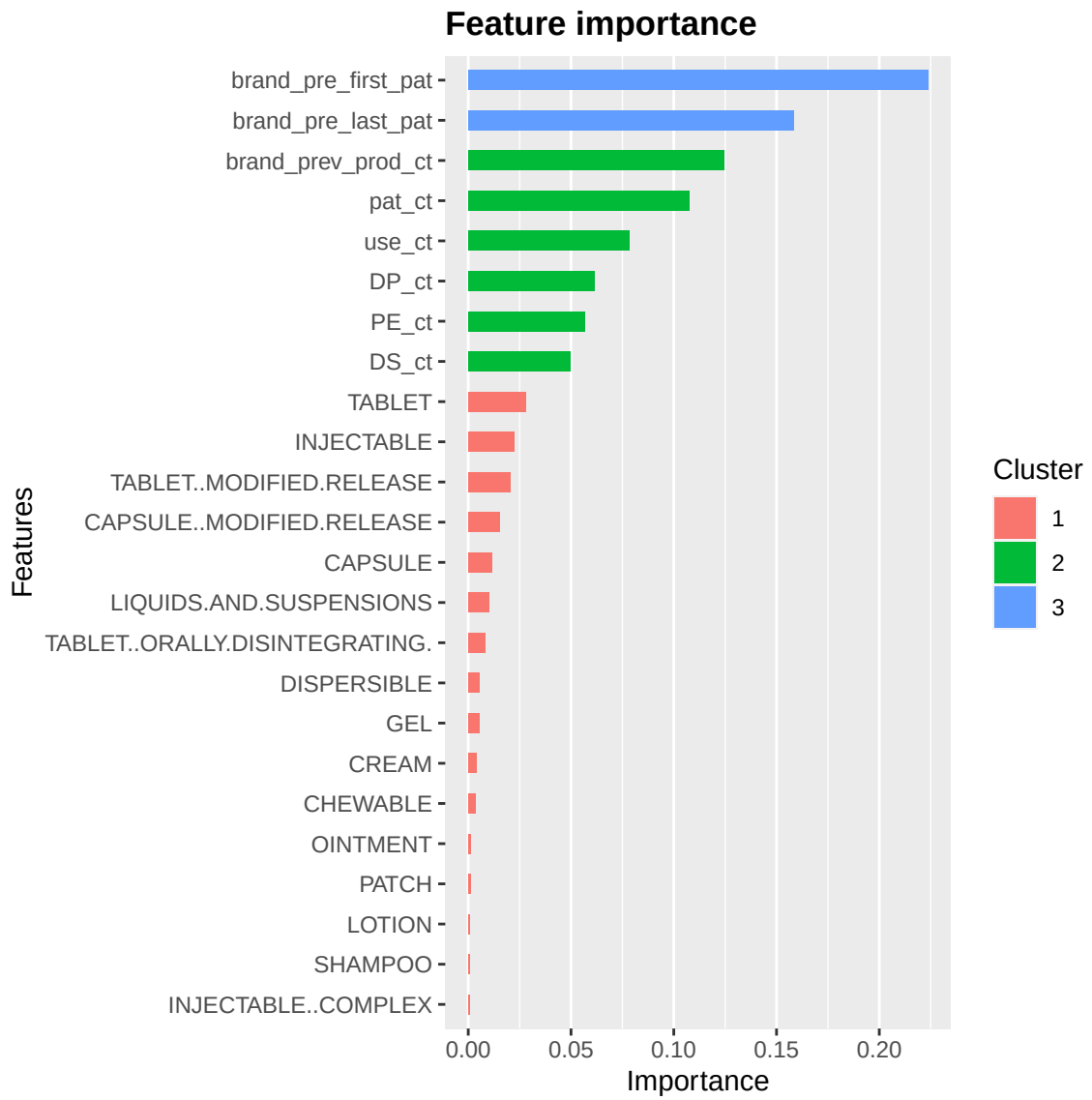


Figure 27: Baseline Feature Importance for XGBoost Model, Prediction of Competition before Last Patent Expiry, Multiple Administration Routes

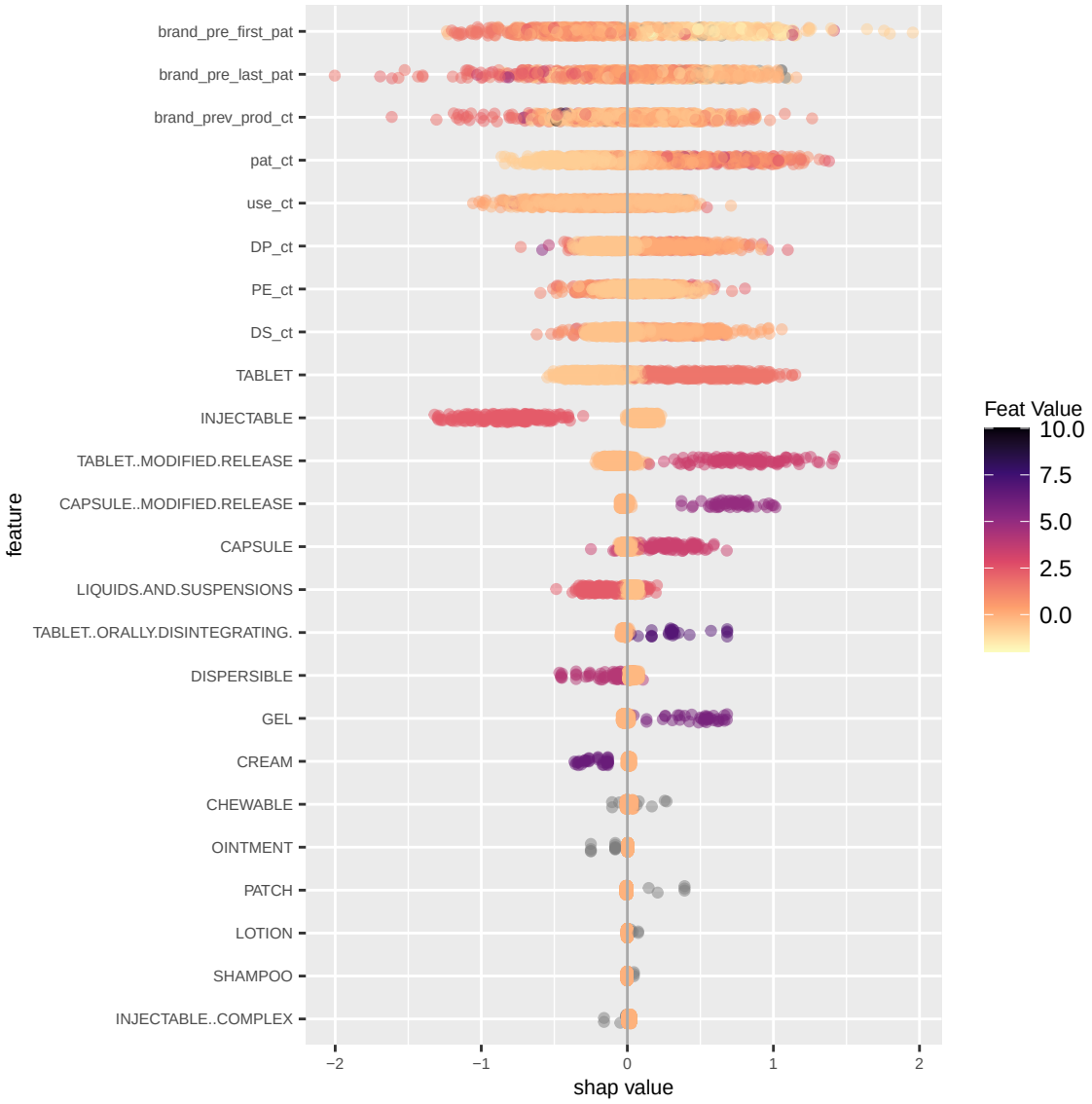


Figure 28: SHAP Plot for XGBoost Model, Prediction of Competition before Last Patent Expiry, Multiple Administration Routes

product, and total patent count are significant regressors. Both sets include different sets of topic models. This suggests that different therapeutic areas could have different tendencies towards generic competition.

In the case of expected competition before first patent expiry, the results are more plausible. Figure 31 presents the feature importance plots for the boosted gradient model. The extracted topic definitions are presented in Figure 32. Time before patent expiry and previous product count were also features used in predicting competition before last patent expiry. In the top 4 topic models (T14, T37, T12, and T33), several prevalent disease states are noted (with no specific distinction)with reference to to the topic definitions: Alzheimers, dementia, HIV, depressive, and insomnia. In context of brand dosage forms, HIV describes 16% of different brand products, while the other terms describe 2 to 3 % of products (Table 10. Thus, the usage codes may discriminate edge cases more so than functioning as a general rule. However, these details may improve overall model accuracy when present.

Table 10: Key Topic Term Frequency in Usage Codes

topic_term	ct	percent
HIV	155	15.7%
DEMENTIA	14	1.4%
ALZHEIMER	15	1.5%
DEPRESSIVE	26	2.6%
INSOMNIA	13	1.3%

When packaging and usage information are added to the boosted gradient model, additional weight is given to “injectable” and “modified release” dosage forms (figure 33). Selected terms from the usage codes are also informative (figure 34) and are different from those chosen withot the packaging components. In the top 4 topic models (T14, T2, T6, and T16), a mixture of common terms and disease states are noted with reference to to the topic definitions: intercourse, keratoses, adjunctive, therapy, and bacteria. In context of

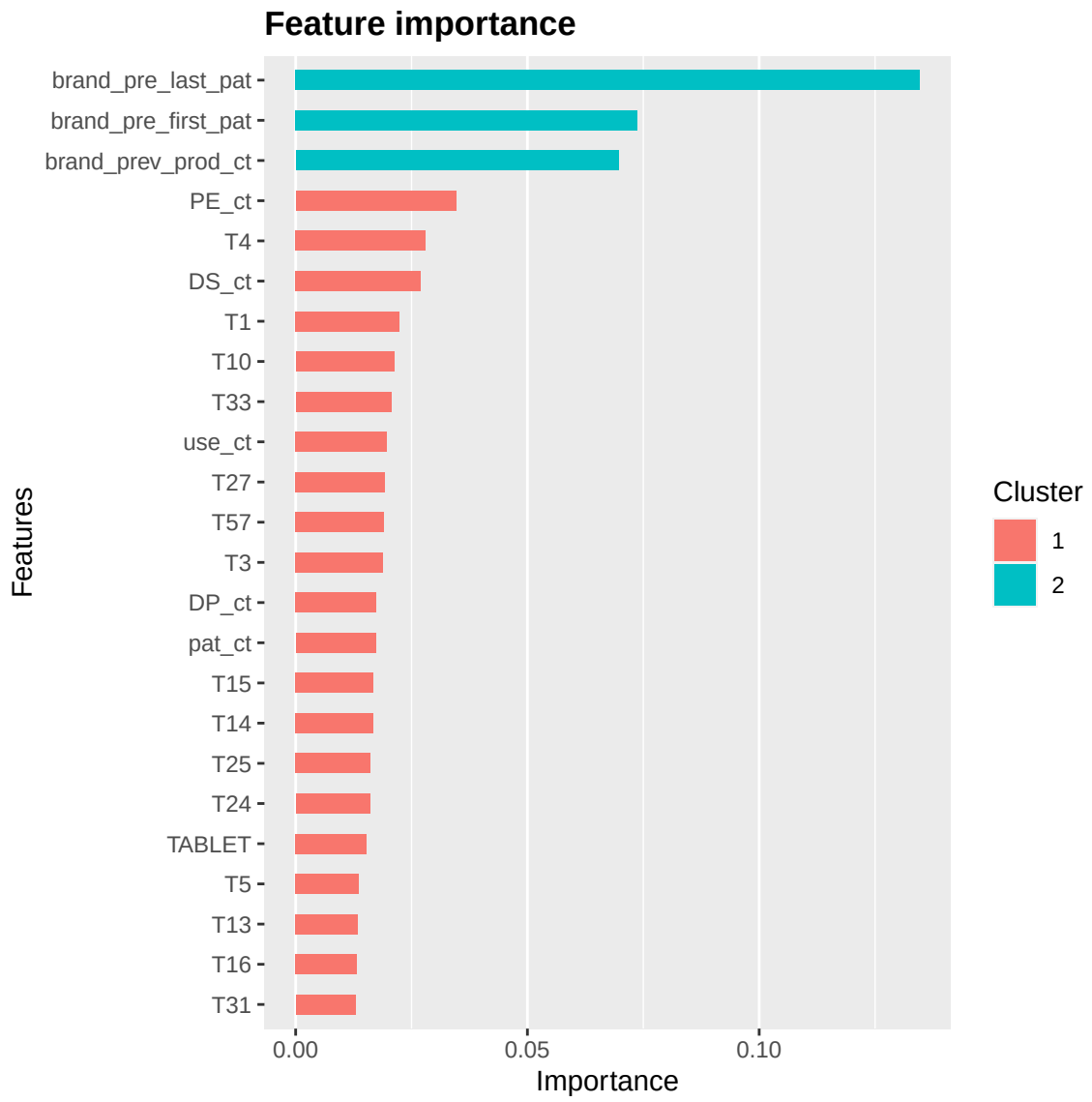


Figure 29: Feature Importance for XGBoost Model with Baseline Patents and 60 Use Code Topics: Prediction of Competition after Last Patent Expiry, Multiple Administration Routes

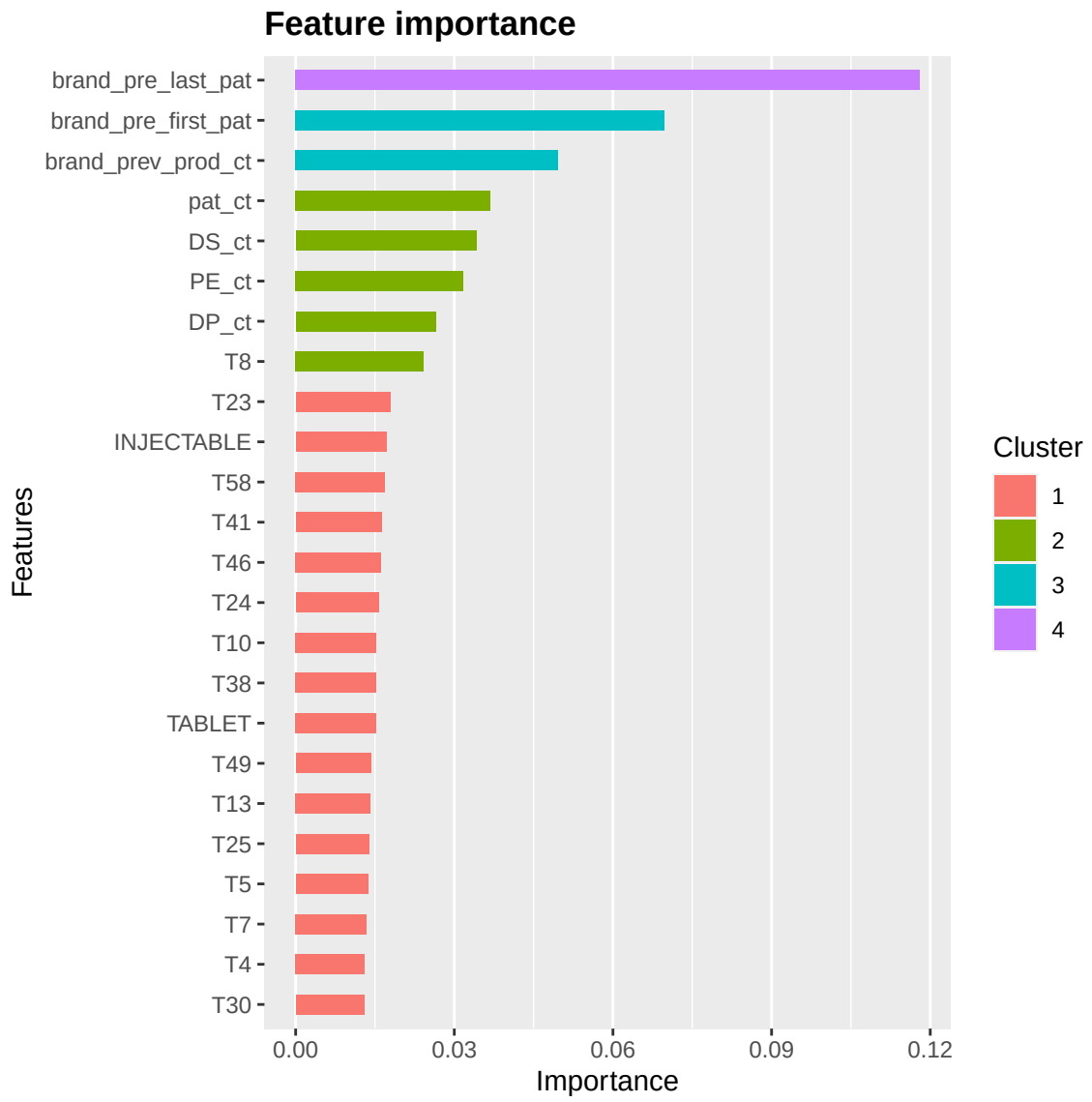


Figure 30: Feature Importance for XGBoost Model with Baseline Patents and 60 Use Code Topics: Prediction of Competition before Last Patent Expiry, Multiple Administration Routes

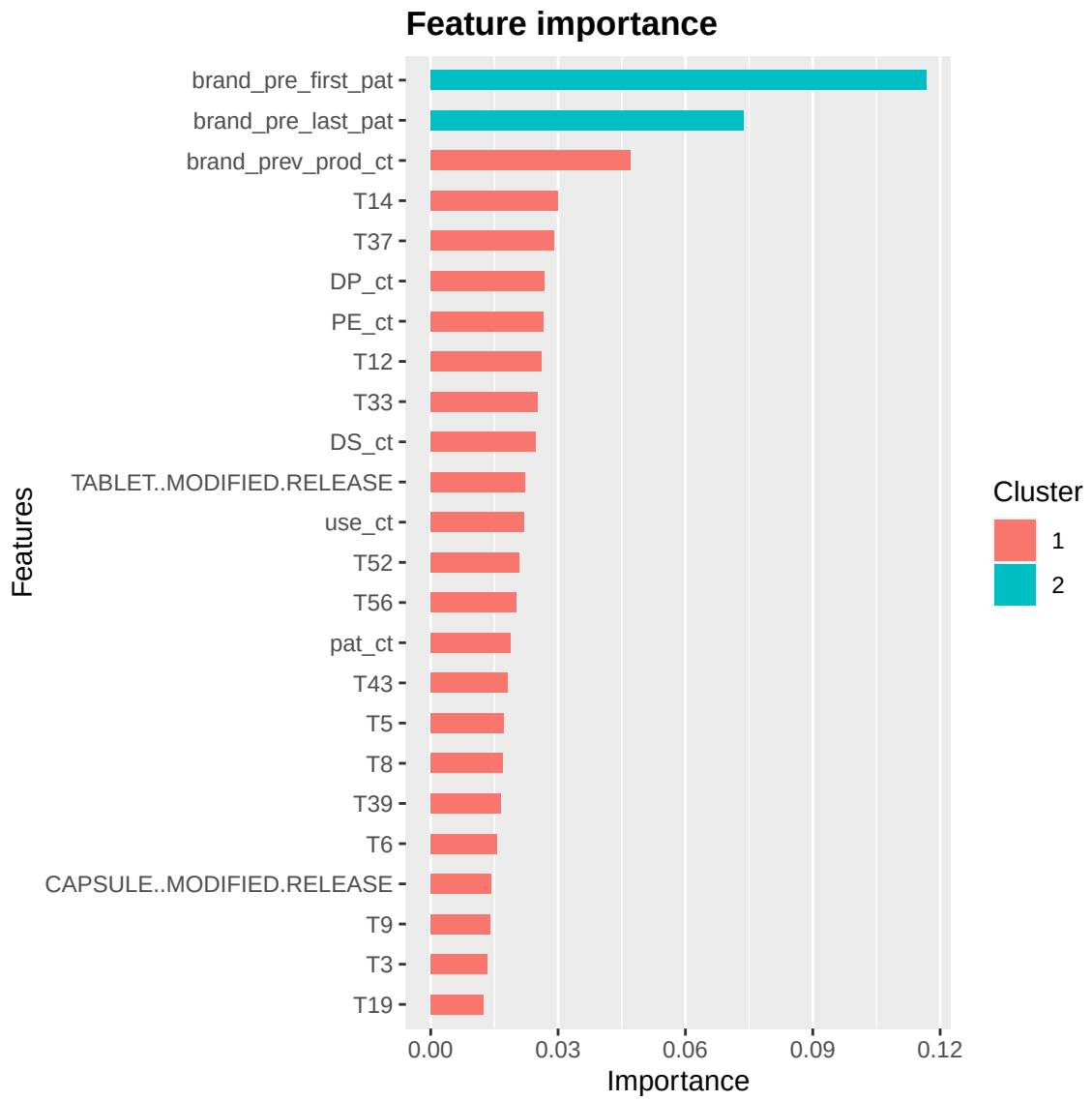


Figure 31: Feature Importance for XGBoost Model with Baseline Patents and 60 Use Code Topics: Prediction of Competition before First Patent Expiry, Multiple Administration Routes

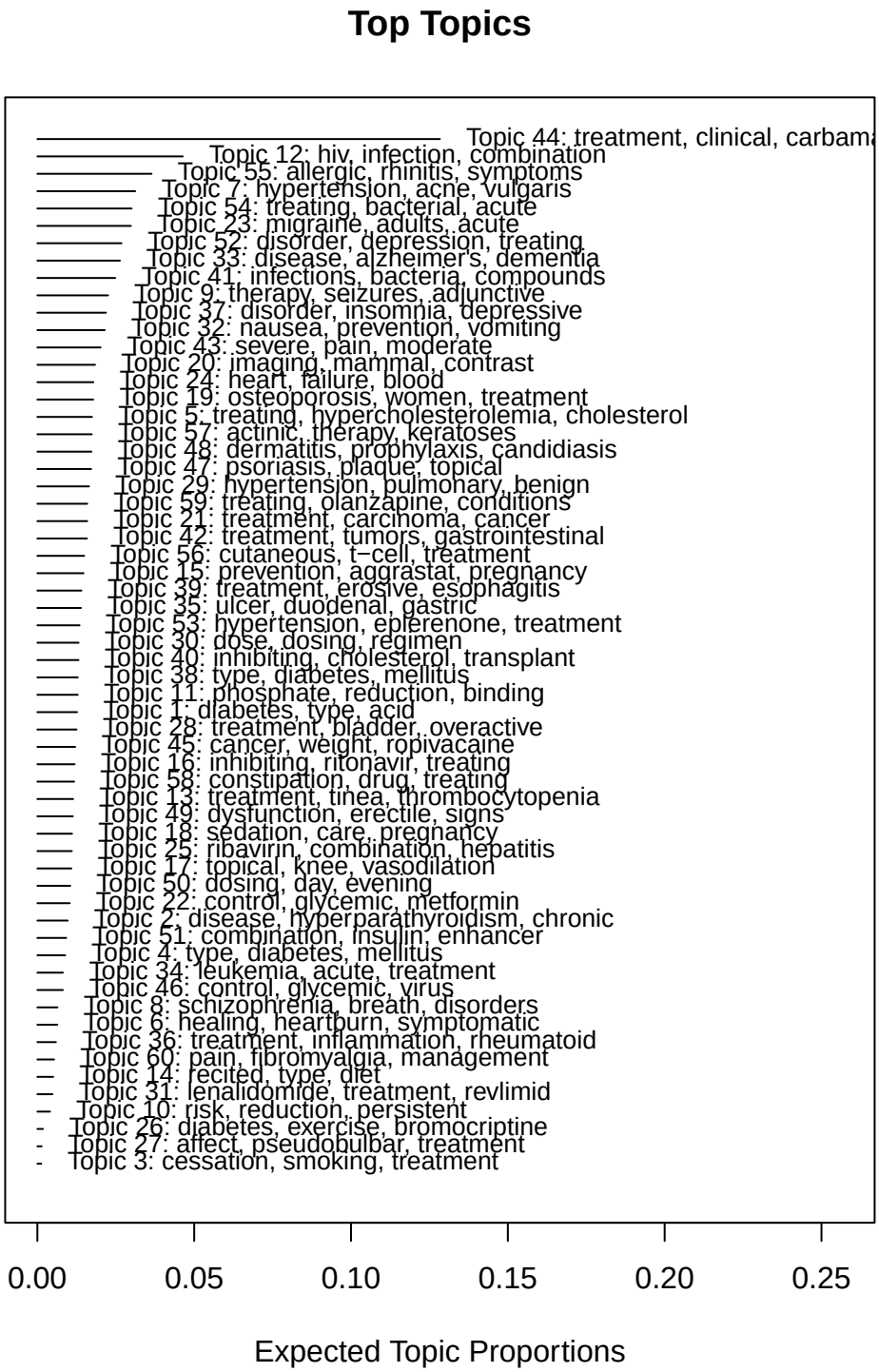


Figure 32: Topic Descriptions from XGBoost Model with Baseline Patents and 60 Use Code Topics to Predict Competition before First Patent Expiry, Multiple Administration Routes: Specific treatements and broad thereapeutic areas represent a reasonable set of classifiers

brand dosage forms, bacteria and therapy describe 10 to 31% of different brand products, while the other terms describe 1 to 5% of products (Table 10). Again, the usage codes may discriminate edge cases more so than functioning as a general rule.

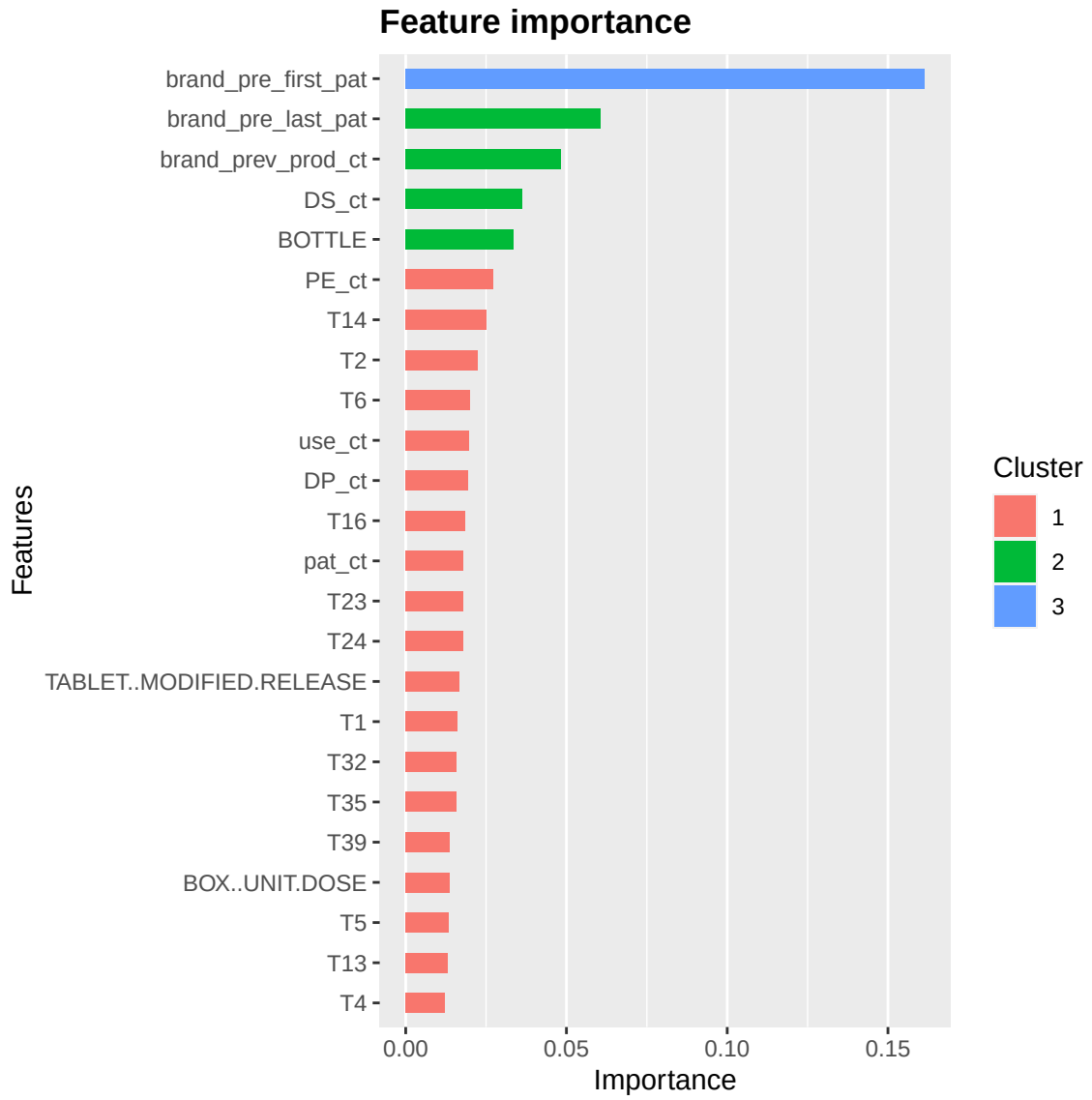


Figure 33: Feaature Importance Plot for XGBoost Model with Baseline Patents, 60 Usage Codes and Packaging for Prediction of Competition before Last Patent Expiry, Multiple Administration Routes

Because products for different routes of administration have different ingredient compositions, limiting the data set to only orally administered products was hypothesized to produce

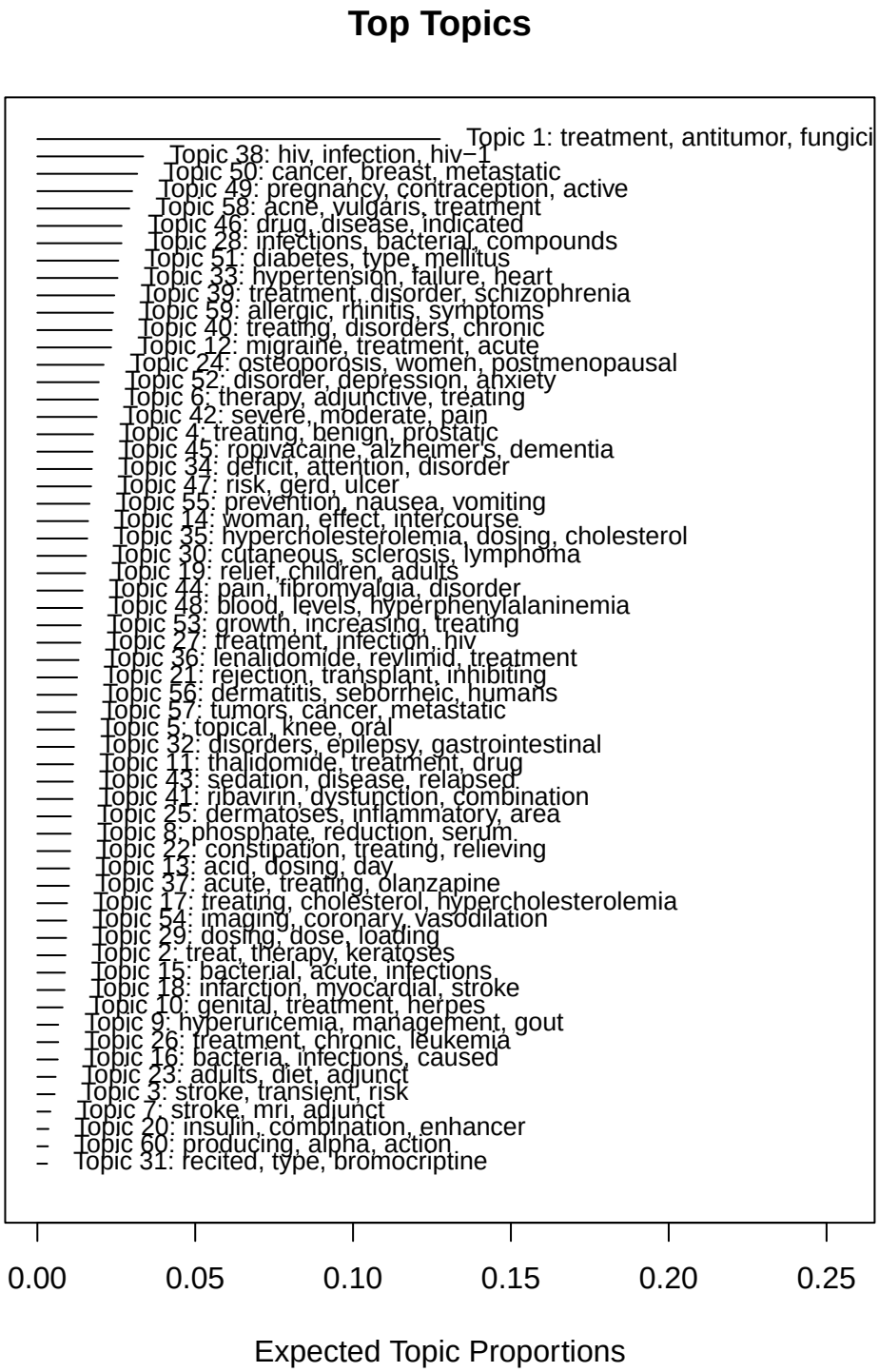


Figure 34: Topics Descriptions from XGBoost Model with Baseline Patents, Packaging Components, and 60 Use Code Topics to Predict Competition before First Patent Expiry, Multiple Administration Routes:

more accurate models. However, model accuracy appeared to be similar to the multiple administration route date set in most cases (Tables 11, 12, and 13. Using the baseline patent information alone, generic competition before the last patent expiry was predicted with significant accuracy using all three methods. For competition before first patent expiry, the base features with packaging information was useful for the SVM model to achieve significant accuracy. Competition before the first patent expiry as also predicted using the boosted gradient model with base patent features and ingredient information. Inclusion of patent use codes decreased model accuracy for svm and decision tree algorithms. However, these features did not affect performance for the boosted gradient algorithm. Combinations of base, usage codes, ingredients, and packaging components produced accurate models only for prediction of competition after last patent expiry.

Table 11: Test Set Accuracy for Oral Adminstration Routes with Baseline Patent Information:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.841	0.682	0.763	0.997
		pre-last	0.770	0.720	0.630	0.004
		post-last	0.718	0.635	0.607	0.220
	xgboost	pre-first	0.984	0.768	0.763	0.473
		pre-last	0.926	0.678	0.630	0.087
		post-last	0.978	0.682	0.607	0.014
	svm_model	pre-first	0.828	0.716	0.763	0.953
		pre-last	0.759	0.735	0.630	0.001
		post-last	0.782	0.664	0.607	0.052
b	rpart	pre-first	0.796	0.698	0.725	0.797
		pre-last	0.769	0.591	0.732	1.000
		post-last	0.771	0.655	0.581	0.039
	xgboost	pre-first	0.986	0.805	0.725	0.015
		pre-last	0.992	0.678	0.732	0.940
		post-last	0.995	0.635	0.581	0.105
	svm_model	pre-first	0.954	0.765	0.725	0.156
		pre-last	0.918	0.604	0.732	1.000
		post-last	0.931	0.736	0.581	0.000
c	rpart	pre-first	0.833	0.730	0.724	0.469
		pre-last	0.817	0.625	0.730	0.998
		post-last	0.679	0.675	0.583	0.012
	xgboost	pre-first	0.988	0.750	0.724	0.265
		pre-last	0.983	0.711	0.730	0.741
		post-last	1.000	0.656	0.583	0.041
	svm_model	pre-first	0.859	0.803	0.724	0.016
		pre-last	0.863	0.651	0.730	0.987
		post-last	0.867	0.682	0.583	0.008

[a] Base Patents

[b] Base and Ingredients

[c] Base and Packaging

Table 12: Test Set Accuracy for Oral Adminstration Routes with Base Patents and Patent Use Codes:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.851	0.711	0.763	0.966
		pre-last	0.772	0.682	0.630	0.066
		post-last	0.814	0.621	0.607	0.364
	xgboost	pre-first	1.000	0.706	0.763	0.976
		pre-last	0.990	0.692	0.630	0.036
		post-last	0.978	0.659	0.607	0.069
	svm_model	pre-first	0.901	0.645	0.763	1.000
		pre-last	0.799	0.654	0.630	0.262
		post-last	0.839	0.659	0.607	0.069
b	rpart	pre-first	0.833	0.716	0.763	0.953
		pre-last	0.743	0.630	0.630	0.531
		post-last	0.794	0.630	0.607	0.264
	xgboost	pre-first	0.993	0.687	0.763	0.995
		pre-last	0.994	0.692	0.630	0.036
		post-last	0.992	0.720	0.607	0.000
	svm_model	pre-first	0.911	0.588	0.763	1.000
		pre-last	0.820	0.640	0.630	0.418
		post-last	0.837	0.592	0.607	0.690
c	rpart	pre-first	0.864	0.711	0.763	0.966
		pre-last	0.801	0.682	0.630	0.066
		post-last	0.821	0.640	0.607	0.180
	xgboost	pre-first	0.999	0.758	0.763	0.601
		pre-last	0.994	0.659	0.630	0.217
		post-last	0.988	0.697	0.607	0.004
	svm_model	pre-first	0.886	0.545	0.763	1.000
		pre-last	0.823	0.611	0.630	0.741
		post-last	0.834	0.592	0.607	0.690

[a] 10 Topics

[b] 60 Topics

[c] 120 Topics

Table 13: Test Set Accuracy for Oral Adminstration Routes with 60 Patent Use Codes, Packaging, and Ingredient Features

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.876	0.691	0.724	0.841
		pre-last	0.842	0.572	0.730	1.000
		post-last	0.787	0.636	0.583	0.107
	xgboost	pre-first	0.996	0.664	0.724	0.955
		pre-last	0.977	0.671	0.730	0.957
		post-last	0.995	0.709	0.583	0.001
	svm_model	pre-first	0.921	0.553	0.724	1.000
		pre-last	0.921	0.592	0.730	1.000
		post-last	0.896	0.623	0.583	0.182
b	rpart	pre-first	0.865	0.685	0.725	0.882
		pre-last	0.806	0.725	0.732	0.614
		post-last	0.818	0.635	0.581	0.105
	xgboost	pre-first	0.996	0.698	0.725	0.797
		pre-last	1.000	0.698	0.732	0.845
		post-last	0.995	0.730	0.581	0.000
	svm_model	pre-first	0.964	0.664	0.725	0.957
		pre-last	0.965	0.584	0.732	1.000
		post-last	0.956	0.703	0.581	0.002
c	rpart	pre-first	0.855	0.698	0.725	0.797
		pre-last	0.806	0.725	0.732	0.614
		post-last	0.788	0.682	0.581	0.007
	xgboost	pre-first	0.986	0.691	0.725	0.844
		pre-last	0.996	0.691	0.732	0.884
		post-last	0.993	0.676	0.581	0.011
	svm_model	pre-first	0.950	0.638	0.725	0.992
		pre-last	0.953	0.584	0.732	1.000
		post-last	0.946	0.696	0.581	0.003

Table 13: Test Set Accuracy for Oral Adminstration Routes with 60 Patent Use Codes, Packaging, and Ingredient Features *(continued)*

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
d	rpart	pre-first	0.798	0.685	0.725	0.882
		pre-last	0.757	0.597	0.732	1.000
		post-last	0.825	0.689	0.581	0.004
	xgboost	pre-first	1.000	0.758	0.725	0.206
		pre-last	0.990	0.718	0.732	0.682
		post-last	0.990	0.716	0.581	0.000
	svm_model	pre-first	0.956	0.745	0.725	0.327
		pre-last	0.935	0.584	0.732	1.000
		post-last	0.931	0.723	0.581	0.000

[a] Base Patents, Packaging, and Use Topics
[b] Base Patents, Ingredients, Packaging, and Use Topics
[c] Base Patents, Ingredients, and Use Topics
[d] Base Patents, Ingredients, and Packaging

Table 14: Test Set Accuracy for Oral Adminstration Routes with 10 Patent Use Codes, Packaging, and Ingredient Features:

features	model	Approved	Acc. Train	Acc. Test	NIR	Pval
a	rpart	pre-first	0.860	0.704	0.724	0.740
		pre-last	0.830	0.546	0.730	1.000
		post-last	0.807	0.675	0.583	0.012
	xgboost	pre-first	0.996	0.724	0.724	0.541
		pre-last	0.990	0.664	0.730	0.970
		post-last	1.000	0.682	0.583	0.008
	svm_model	pre-first	0.882	0.730	0.724	0.469
		pre-last	0.873	0.618	0.730	0.999
		post-last	0.870	0.662	0.583	0.028
b	rpart	pre-first	0.827	0.550	0.725	1.000
		pre-last	0.810	0.624	0.732	0.998
		post-last	0.810	0.649	0.581	0.056
	xgboost	pre-first	0.996	0.752	0.725	0.263
		pre-last	0.998	0.745	0.732	0.396
		post-last	0.993	0.662	0.581	0.027
	svm_model	pre-first	0.950	0.698	0.725	0.797
		pre-last	0.955	0.591	0.732	1.000
		post-last	0.948	0.716	0.581	0.000
c	rpart	pre-first	0.823	0.550	0.725	1.000
		pre-last	0.810	0.624	0.732	0.998
		post-last	0.786	0.669	0.581	0.018
	xgboost	pre-first	0.998	0.631	0.725	0.995
		pre-last	1.000	0.732	0.732	0.542
		post-last	1.000	0.682	0.581	0.007
	svm_model	pre-first	0.950	0.691	0.725	0.844
		pre-last	0.931	0.591	0.732	1.000
		post-last	0.936	0.730	0.581	0.000

[a] Base Patents, Packaging, and Use Topics

[b] Base Patents, Ingredients, Packaging, and Use Topics

[c] Base Patents, Ingredients, and Use Topics

Unlike the set of multiple administration routes, the oral administration route alone did not have a single combination of predictors with significant accuracy for all three competition categories. Considering the prediction for competition before last patent expiry in the oral route of administration, the rpart decision tree (Figure 35) has a complicated mixture of rules assigned to the timing of product approval before first and last patent expiry. Exploring the decision tree, the rules seem to create numerous fine cutoff points for similar categories. In one large category, drug products approved much earlier than an arbitrary median with relatively few general patents that do not apply to drug substances explains 27% of the categories without expected generic competition. On the opposite side of the decision tree, a product with few patents approved closer to the first patent expiration date corresponds to the likely entrance of generic competition. However, the feature importance seems similar to both the boosted gradient tree (Figure 36) and the decision tree for the multiple-administration routes.

For the oral administration products with competition before the first patent expiry, the boosted gradient model used features from the base set and ingredients to produce a model with significant accuracy. The key features selected for competition after last patent expiry appear similar, albeit weighted differently than, those selected for competition before last patent expiry (Figures 37, 38, and 39).

7 Discussion

Overall, the model weighting on product approval prior to patent expiry merits a question as to whether generic competition is a matter of timing. Given enough notice, some generic competition seems likely to occur regardless of the complexity of the dosage form or patent portfolio. However, the models also singled out “injectable”, “liquids”, and “modified release” dosage forms as predictive features. Each of these reflect differing levels of complexity, which

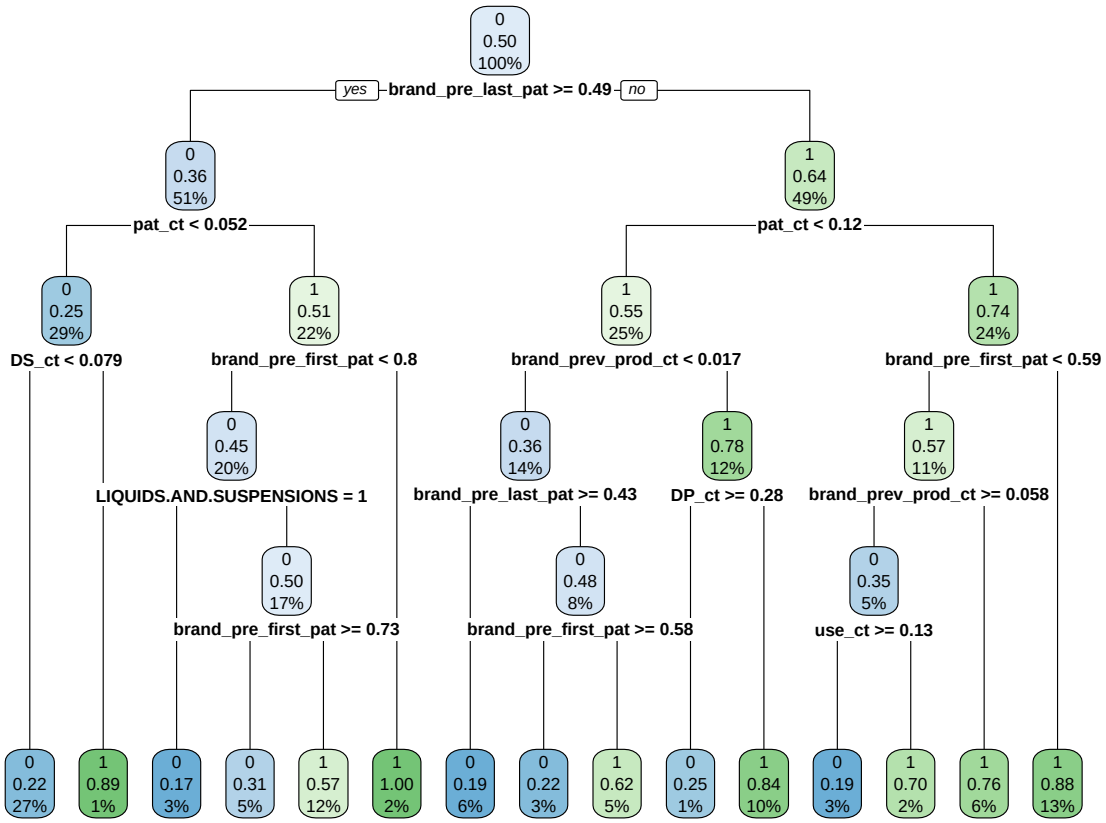


Figure 35: Baseline RPART Model, Prediction of Competition Before Last Patent Expiry, Oral Administration Routes

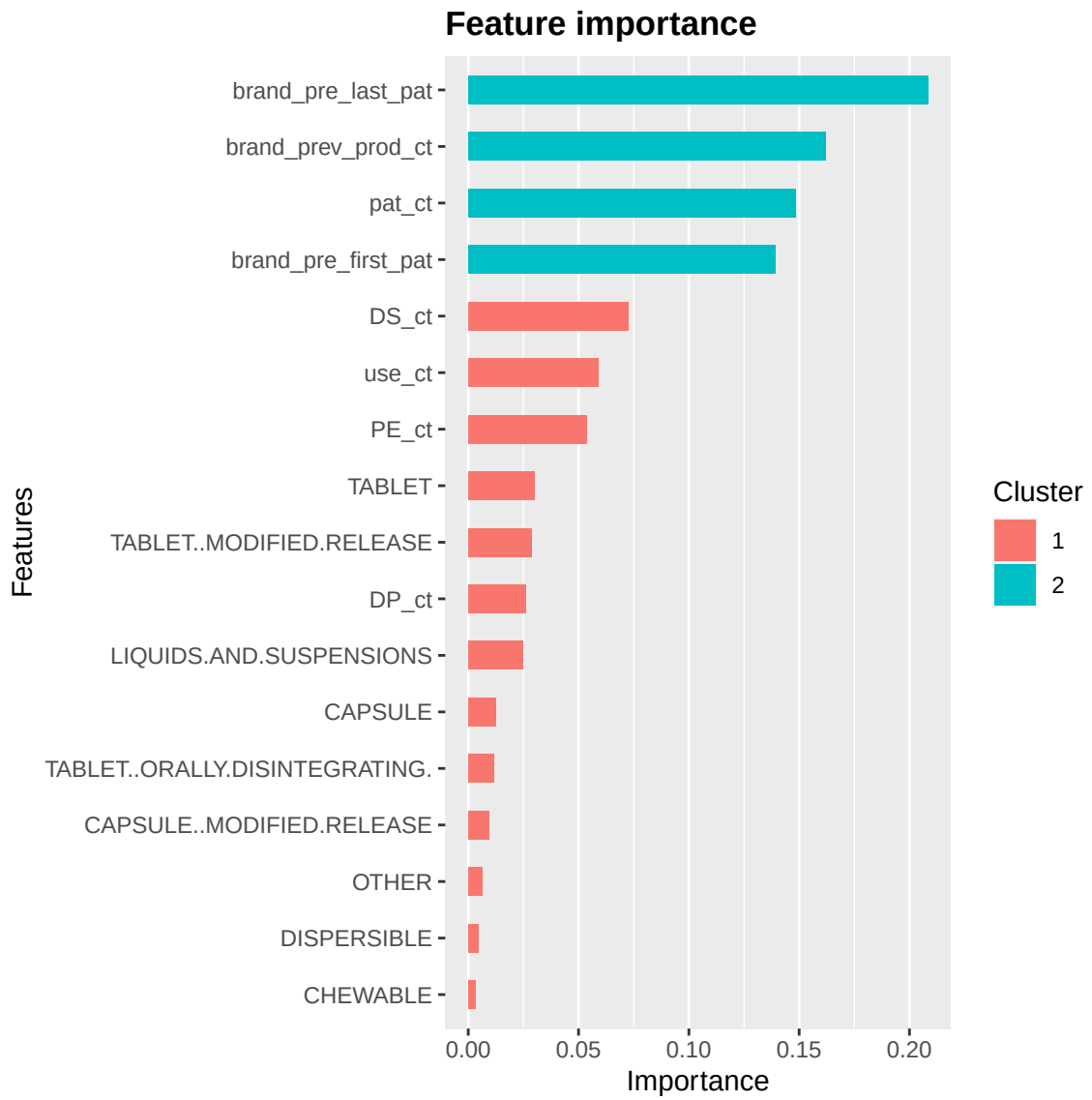


Figure 36: Baseline Feature Importance for XGBoost Model, Prediction of Competition before Last Patent Expiry, Oral Administration Routes

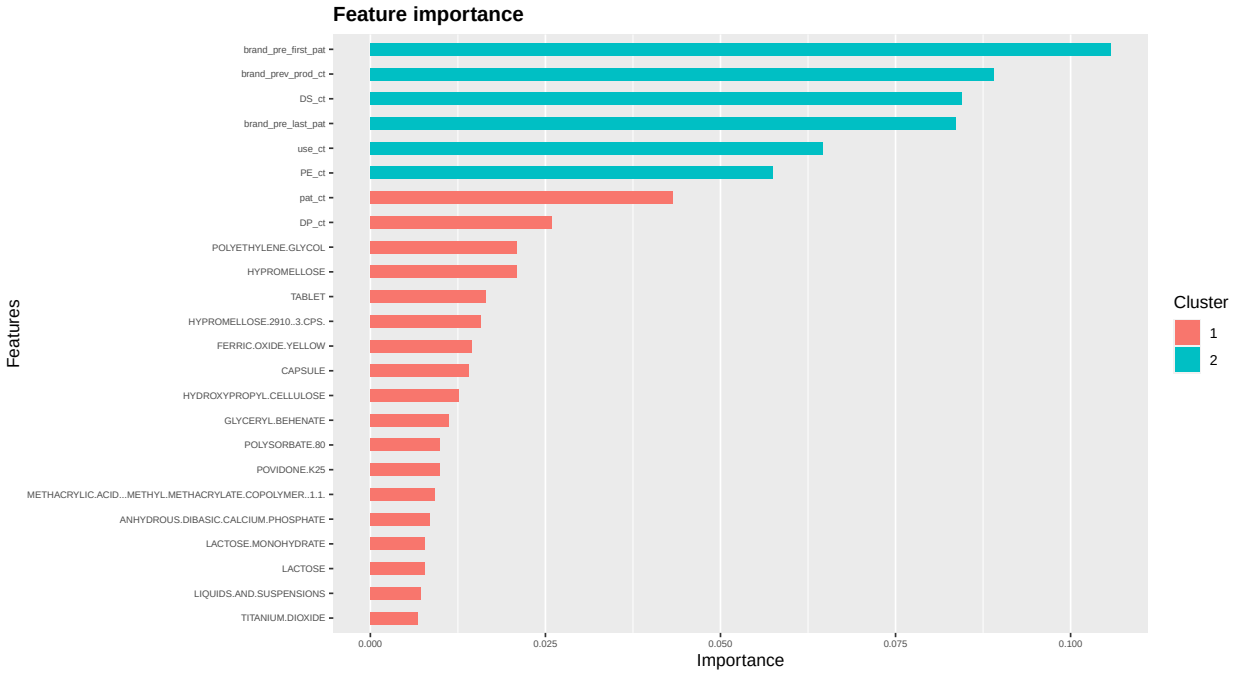


Figure 37: Feature Importance for XGBoost Model using Patent and Ingredient Features for Prediction of Competition before First Patent Expiry, Oral Administration Routes

suggests that product design may matter.

While boosted gradient models accurately predicted generic competition with multiple data sets, the 15 – 20% loss in accuracy between test and training sets suggests the data was overfit during training. With other algorithms, very few models achieved greater than 90% training set accuracy, and few models saw more than 10% decline in accuracy between the training and test set analysis. By oversampling the training set, the event distributions were artificially changed from 15-30% to 50%. While this was intended to improve model discrimination by providing equal visibility to sample features, it may have allowed the boosted gradient methods to overweight test set characteristics. Conservative learning rates and aggressive random sampling and variable replacement parameters may have tempered training, leading to acceptable accuracy for the test sets.

Across all models, the capability for predicting competition after last patent expiry belies some complexity with filtering the basis set data analysis. Using too many exclusion crite-

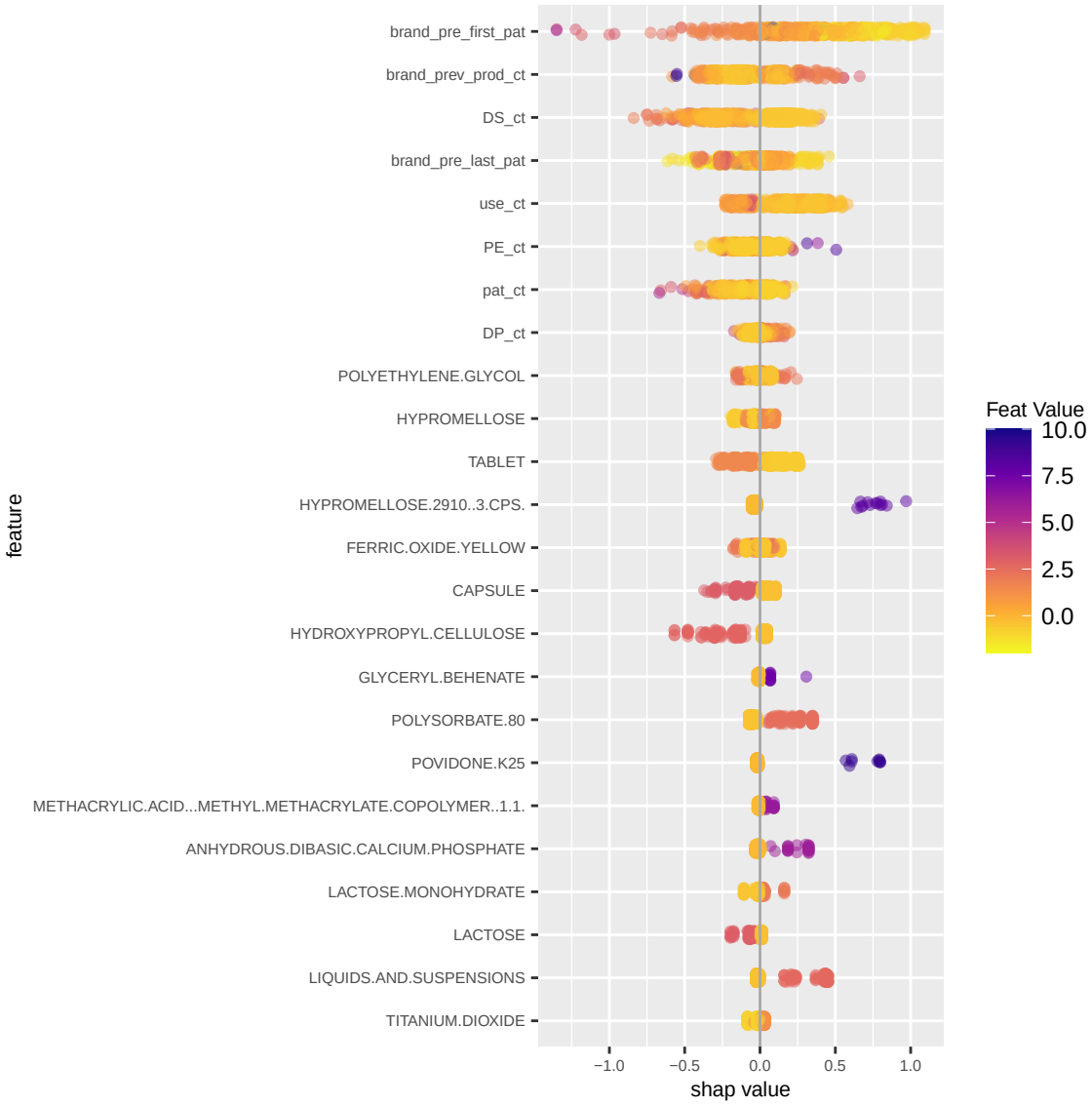


Figure 38: SHAP Plot for XGBoost Model with Baseline Patents and Ingredients, Prediction of Competition before first Patent Expiry, Oral Administration Routes

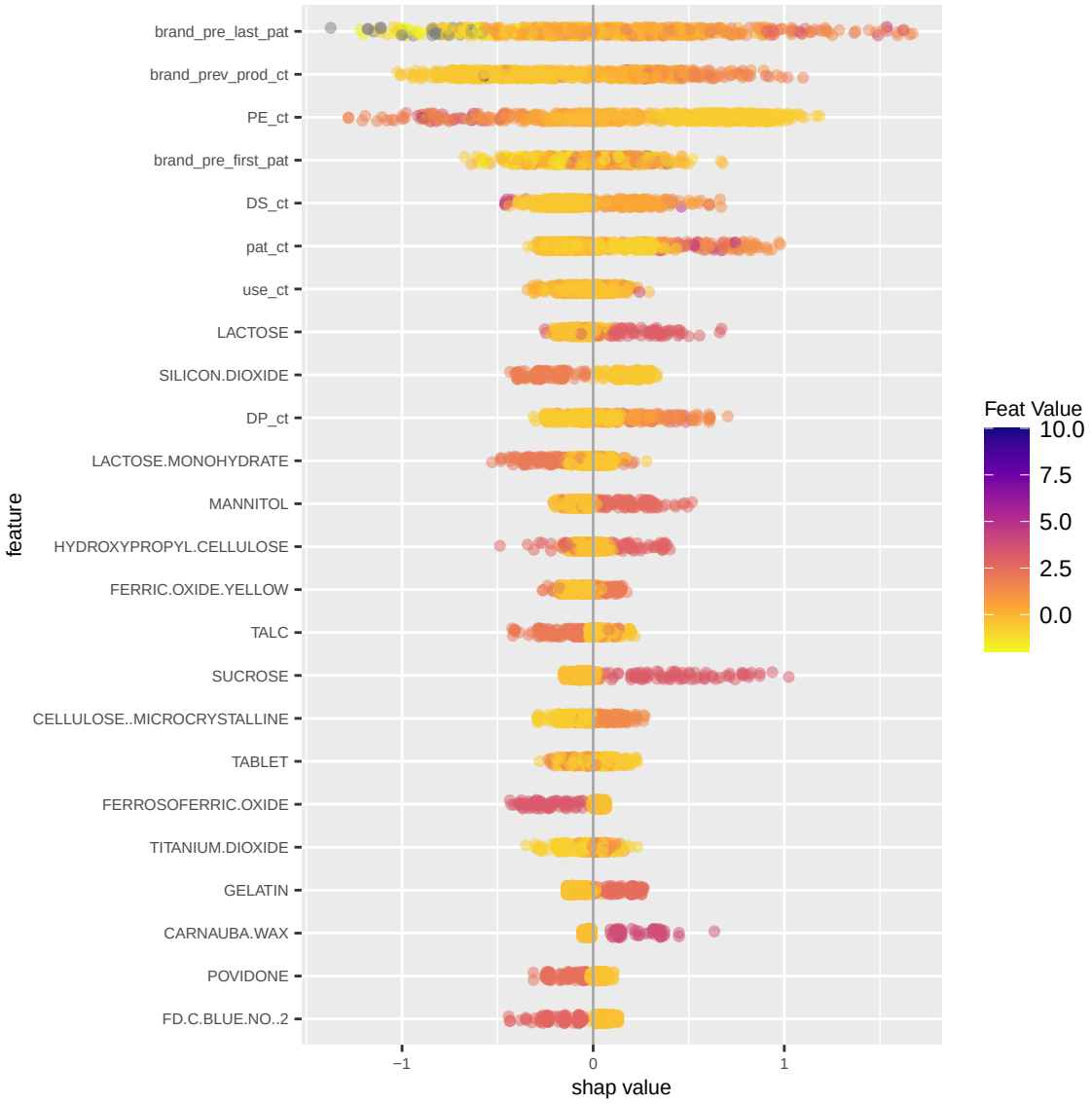


Figure 39: SHAP Plot for XGBoost Model with Baseline Patents and Ingredients, Prediction of Competition before first Patent Expiry, Oral Administration Routes

ria can produce insufficient cases for analysis or a rank-deficient matrix. Branded products used in this study were approved no later than 2010. Thus, patent expiry may be well after the final Orange Book data point in 2022. This creates a situation where the analysis may be biased towards a simple has / does not have patent discrimination. Likewise, limiting the analysis to only the 4 most prominent routes of administration produced a tradeoff between additional datapoints and complexity in the routes of administration. Although these windows highlight a priori choices about the data, setting broader windows or a systemic evaluation may have provided greater insights.

From this study, the market analysis may represent the largest achievement. For many industry insiders, there is often an intuitive feel for competitive aspects without having a quantitative picture of the competitive space. This study provides a succinct, longitudinal summary of the interplay between brand and generic companies. More important, the compiled data provides a basis for future studies into how band and generic companies differ. For example, differences between generic and branded formulations can be compared, while market segmentation may be performed on companies based on product selection, time-to-market, and product composition. Recognizing that competition reflects financial viability, the data set could readily be joined with sales cost and sale volume data.

8 Conclusion

FDA approved products, brand product patents, and generic competition between 1982 and 2021 show increasing patent protection on products in parallel with increasing numbers of abbreviated product applications. Although patents themselves provide the largest barrier to generic entry, predictive models demonstrate that compositional elements or specific areas of therapeutic treatment may present limited barriers to generic competition. For brand and generic competitors alike, it appears that maximizing product lifetime is a race against market saturation with other products.

References

Beall, R., Darrow, J., & Kesselheim, A. (2018). A method for approximating future entry of generic drugs. *Value in Health*, *21*(12), P1382–1389. https://doi.org/https://doi.org/10.1016/j.jval.2018.04.1827

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, *49*(2), 1–50.

Buckley, L. A., Bebenek, I., Cornwell, P. D., et al. (2020). Drug development 101: A primer. *International Journal of Toxicology*, *39*(5), 379–396. https://doi.org/10.1177/1091581820939083

Carrier, M., Lemley, M., & Miller, S. (2020). Playing both sides? branded sales, generic drugs, and antitrust policy. *Hastings Law Journal*, *71*, 2. https://repository.uchastings.edu/hastings_law_journal/vol71/iss2/3

Center for Biologics Evaluation and Research, Center for Drug Evaluation and Research, and Center for Vetrinary Medicine. (2018). *Structured product labeling (spl) implementation guide with validation procedures* (tech. rep.). U.S. Department of Health, Human Services, Food, and Drug Administration. https://www.fda.gov/media/84201/download

Center for Drug Evaluation and Research. (1998). *Guidance for industry: Variations in a drug product that may be included in a single anda*, U.S. Department of Health, Human Services, Food, and Drug Administration. https://www.fda.gov/files/drugs/published/Variations-in-Drug-Products-that-May-Be-Included-in-a-Single-ANDA.pdf

Center for Drug Evaluation and Research (CDER). (2017). *What is the approval process for generic drugs*, U.S. Department of Health, Human Services, Food, and Drug Administration. https://www.fda.gov/drugs/generic-drugs/what-approval-process-generic-drugs

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Comittee on Professional Ethics of the American Statistical Association. (2018). *Ethical guidelines for statistical practice*. https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Congressional Budget Office. (2021). *Research and development in the pharmaceutical industry* (tech. rep.). https://www.cbo.gov/publication/57025

Congressional Research Service. (2021). *Drug prices: The role of patents and regulatory exclusivities* (tech. rep.). https://crsreports.congress.gov/product/pdf/R/R46679

Conrad, R., & Lutter, R. (2019). *Generic competition and drug prices: New evidence linking greater generic competition and lower generic drug prices* (tech. rep.) (Accessed 11/7/2022). U.S. Department of Health, Human Services, Food, and Drug Administration. https://www.fda.gov/media/133509/download

Contents and term of patents, provisional rights (tech. rep.). (2013). 35 U.S.C § 154.

Dave, C., Pawar, A., Fox, E., Brill, G., & Kesselheim, A. (2018). Predictors of drug shortages and association with generic drug prices: A retrospective cohort study. *Value in Health*, *21*(11), 1286–1290. https://doi.org/10.1016/j.jval.2018.04.1826

Deloitte Centre for Health Solutions. (2021). *Nurturing growth: Measuring the return from pharmaceutical innovation* (tech. rep.). https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/Measuring-the-return-of-pharmaceutical-innovation-2021-Deloitte.pdf

Drake, K. M., He, R., McGuire, T. G., & Ndikumana, A. K. (2021). *No free launch: At-risk entry by generic drug firms*. National Bureau of Economic Research. https://doi.org/10.3386/w29131

Drug price competition and patent term restoration act of 1984, Pub. L. No. 98-417, 98 Stat. 1585 (1984). https://www.govinfo.gov/content/pkg/STATUTE-98/pdf/STATUTE-98-Pg1585.pdf

Frerebeau, N. (2019). tabula: An R Package for Analysis, Seriation and Visualization of Archaeological Count Data. *Journal of Open Source Software*, *4*(44), 1821. https://doi.org/10.21105/joss.01821

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning. 2nd edition*. Springer.

Hornecker, J. (2009). Generic drugs: History, approval process, and current challenges. *U.S. Pharm*, *34*(6), (Generic Drug Review Supplement) 26–30.

IQVIA Institute for Human Data Science. (2022). *The use of medicines in the u.s. 2022: Usage and spending trends and outlook to 2026* (tech. rep.). https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/the-use-of-medicines-in-the-us-2022/iqvia-institute-the-use-of-medicines-in-the-us-2022.pdf

Lambert, W. (2010). Considerations in developing a target product profile for parenteral pharmaceutical products. *AAPS PharmSciTech*, *11*(3), 1476–1481. https://doi.org/10.1208/s12249-010-9521-x

Lee, C., Chen, X., Romanelli, R., & Segal, J. (2016). Forces influencing generic drug development in the united states: A narrative review. *Journal of Pharmaceutical Policy and Practice*, *26*(9), 1. https://doi.org/https://doi.org/10.1186/s40545-016-0079-1

Lichtenberg, F. (2016). How many life-years have new drugs saved? a three-way fixed-effects analysis of 66 diseases in 27 countries, 2000-2013. *International Health*, *11*(5), 403–416. https://doi.org/https://doi.org/10.1093/inthealth/ihz003

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1–22.

Lötsch, J., Malkusch, S., & Ultsch, A. (2021). Optimal distribution-preserving downsampling of large biomedical data sets (opdisdownsampling). *PLoS ONE*, *16*(8), e0255838. https://doi.org/10.1371/journal.pone.0255838

Midha, K., & McKay, G. (2009). Bioequivalence; its history, practice, and future. *The AAPS journal*, *11*(4), 664–670. https://doi.org/https://doi.org/10.1208/s12248-009-9142-z

National Library of Medicine. (2022). *Spl resources* [DailyMed]. https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-drug-labels.cfm

Olson, L., & Wendling, B. (2013). *The effect of generic drug competition on generic drug prices during the hatch-waxman 180-day exclusivity period* (tech. rep.). Federal Trade Commission. https://www.ftc.gov/reports/estimating-effect-entry-generic-drug-prices-using-hatch-waxman-exclusivity

Pandas. (2021). *Api reference*. https://pandas.pydata.org/pandas-docs/stable/reference/index.html

pyPDF2. (2022). *Welcome to pypdf2*. https://pypdf2.readthedocs.io/en/latest/index.html

Reiffen, D., & Ward, M. (2005). Generic drug industry dynamics. *The Review of Economics and Statistics*, *87*(1), 37–49. https://doi.org/https://doi.org/10.1162/0034653053327694

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Rome, B. N., Lee, C. C., Gagne, J. J., & Kesselheim, A. S. (2021). Factors associated with generic drug uptake in the united states, 2012 to 2017. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *24*(6), 804–811. https://doi.org/https://doi.org/10.1016/j.jval.2020.12.020

Russo, A. A., & Johnson, J. (2014). Research use exemptions to patent infringement for drug discovery and development in the united states. *Cold Spring Harbor perspectives in medicine*, *5*(2), a020933. https://doi.org/10.1101/cshperspect.a020933

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE*, *16*(8), e0254937. https://doi.org/https://doi.org/10.1371/journal.pone.0254937

Sun, D. G., Hu, W., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it. *Acta Pharmaceutica Sinica B*, *12*(7), 3049–3062. https://doi.org/https://doi.org/10.1016/j.apsb.2022.02.002

Tamimi, N., & Ellis, P. (2009). Drug development: From concept to marketing! *Nephron Clin Pract*, *113*, c125–c131. https://doi.org/10.1159/000232592

Therneau, T., & Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees* [R package version 4.1-15]. https://CRAN.R-project.org/package=rpart

United States Food and Drug Administration. (2005 to 2022). *Approved drug products with therapeutic equivalence evaluations. 25th to 42nd editions* [Retrieved from The FDA Law Blog, 9/2/2022]. https://www.thefdalawblog.com/orange-book-archives/