Supplementary Information for: Just how good an investment is the biopharmaceutical sector?

Richard T. Thakor,[1] Nicholas Anaya,[2] Yuwei Zhang,[2] Christian Vilanilam,[2] Kien Wei Siah,[2,3] Chi Heem Wong,[2,4] and Andrew W. Lo[2-5]

[1]1Carlson School of Management, University of Minnesota, Minneapolis, Minnesota, USA.

[2]MIT Sloan School of Management and Laboratory for Financial Engineering, Cambridge, Massachusetts, USA.

[3]MIT Department of Electrical Engineering and Computer Science, Cambridge, Massachusetts, USA.

[4]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA.

[5]AlphaSimplex Group, LLC, Cambridge, Massachusetts, USA.

1 Supplementary Methods

In the main text, we classify a firm as either pharma or biotech utilizing an algorithm called "*k*-means", which incorporates financial information and company characteristics over time. In these Supplementary Methods, we also examine how our analysis differs when using other classification methods. We provide details of each classification method in this section.

K-means Algorithm

K-means is an algorithm that classifies companies into categories based on how similar they are to each other. More specifically, the algorithm starts with a small number of "seed" companies that are selected to typify a given category. We use a data-driven method in order to identify these seed companies, as we describe below. Based on the characteristics of these "seed" companies, a cluster center is calculated. For each additional company, the algorithm calculates the Euclidean distance between that firm and the cluster center of the "seed" firms. The company is then classified into the category for which the distance is minimized, assuming an equal weight to all characteristics. After this, the cluster centers are re-calculated. We run the *k*-means algorithm separately for each year of the sample starting in 1980 (the first year in which there is consistently at least one biotech seed company), which allows a dynamic classification and the possibility that a company might change industry classifications if its characteristics change enough from year to year. Prior to 1980, we classify companies as pharma companies.

The characteristics that are used for the algorithm are a wide variety of financial data and company characteristics from the CRSP/Compustat database. These characteristics include:

1. Number of Employees
2. Intangible Assets*
3. Research and Development (R&D) Expenses*
4. Value of Total Assets
5. Cash Holdings*
6. Amount of Debt*
7. Amount of Dividends Paid*
8. Property, Plant, and Equipment (PPE)*
9. Sales*
10. Company Age (Time since Initial Public Offering (IPO))
11. Capital Expenditures*
12. Advertising Expenses*

In the above, as is standard in the finance and economics literature, to control for differences in size, variables with an asterisk (*) are scaled by total assets and we use the natural logarithm of the number of employees and total assets (items 1 and 4). We focus on these characteristics because they encompass a wide range of company characteristics, and moreover the database provides time-series values for them over time, thus allowing us to dynamically run our classification method. In order to avoid using our target ex post outcome variables as selection criteria for our ex ante classification groups, we do not use earnings or stock returns as characteristics for the algorithm.

Utilizing these characteristics, the *k*-means algorithm classifies each company to either the pharma or biotech industry. For the purposes of the algorithm, finer categorizations are utilized (i.e. distinguishing between early and late biotech companies), but for our analysis we use broad pharma or biotech categories for better comparability to other classifications, which do not allow for more detailed classifications. The results for these finer categorizations are available upon request.

The pharma industry includes companies that are either “big pharma” or “smaller pharma” companies. Big pharma companies' research traditionally includes, but is not limited to, small molecules. These companies spend significant amounts of revenue on R&D, they manufacture and market products, and have numerous products that reach patients. In order to identify the seed companies each year, we translate this into the characteristics that are available to us. In particular, we consider a company to be a big pharma seed company in a particular year if the following simultaneously hold: it is in the top quartile in terms of size (total assets); it is in the top quartile in terms of number of employees; it is in the top quartile in terms of property, plant, and equipment, since this is directly related to its ability

to manufacture products; it has positive advertising expense, indicating that it markets products; it has a positive amount of intangible assets, which will include patents and goodwill; it has paid a positive amount of dividends, since more mature firms are ones the pay dividends; and it is above the median in terms of age. Examples of companies that met these thresholds are Johnson & Johnson, GlaxoSmithKline, and Merck.

Smaller pharma tend to have fewer assets than big pharma companies, and produce drugs focused on a niche disease or area and often use in-licensing (rather than R&D) to acquire drugs. These may also include orphan drugs, drug delivery, and generics companies. These companies have received FDA approval and have products on the market. We consider a company to be a smaller pharma seed company in a particular year if the following simultaneously hold: it is above-median but below the top quartile in terms of size (total assets); it is above-median but below the top quartile in terms of number of employees; it is above-median but below the top quartile in terms of property, plant, and equipment; and it has a positive amount of intangible assets. Examples of companies that met these qualifications are Forest Laboratories, Valeant Pharmaceuticals, and Telios.

The biotech industry includes companies that are either "early biotech" or "late biotech" companies. Biotech companies traditionally differ from pharma companies in their production process (using living material) but they may also produce small molecules. They may or may not have drugs on the market and often operate at a loss due to significant R&D investment despite little or no revenue. Early biotech includes smaller and younger pre-approval companies that do not have an FDA-approved product. We consider a company to be an early biotech seed company in a particular year if the following simultaneously hold: it is in the bottom quartile in terms of size (total assets); it is in the bottom quartile in terms of number of employees; it is in the bottom quartile in terms of property, plant, and equipment; and it is below the median in terms of age. Examples of early biotech seed companies are Assembly Biosciences, Catalyst Pharmaceuticals, and Evoke Pharma. We consider a company to be a late biotech seed company in a particular year if the following simultaneously hold: it is below-median but above the bottom quartile in terms of size (total assets); it is below-median but above the bottom quartile in terms of number of employees; it is below-median but above the bottom quartile in terms of property, plant, and equipment. Examples include Sarepta Therapeutics, Sangamo Therapeutics, and Repligen.

It should be noted that it is also possible to run the *k*-means algorithm "unsupervised"—in other words, without the use of seed companies. In this case, the algorithm determines groups based on the data itself, and assigns companies to those groups based on their characteristics. A potential disadvantage of running the *k*-means algorithm in this way is that the algorithm does not provide a way to place an identity on the groups. Put differently, while it may identify four groups, it does not give any guidance as to which of the groups in any given year is early biotech, which is late biotech, and so on. This disadvantage notwithstanding, for robustness we have also run our return results using the unsupervised *k*-means algorithm. To identify the four groups in a given year, we ranked the groups

according to their total assets, with the smallest being early biotech and the largest being big pharma. Our return results are qualitatively very similar when using the unsupervised *k*-means algorithm, and are available upon request.

Collaborative Filtering

We refer to our first alternative classification method as "collaborative filtering". Collaborative filtering is a machine learning method for matching similar items in order to provide recommendations. Similar to the *k*-means method above, the algorithm requires a number of example firms that are selected to typify each given category, in order to provide "training" for the machine learning. Based on those example firms, the algorithm then uses the same 12 variables described above for *k*-means algorithm, but iterates and weights characteristics based on score loadings. Companies are then assigned to each category based upon the score loadings and weights of these characteristics relative to each category.

Put differently, let $Y_{i,j}$ be the score for a company *i* that has been assigned to category (industry) *j*. The collaborative filtering algorithm assumes that *Y* takes the form of $Y = X \cdot \Theta$, where *X* is a matrix of input variables (company characteristics) and Θ is known as a "feature map", which is a matrix of weights corresponding to the input variables. Collaborative filtering iterates in order to estimate the weights Θ for every company to fit known values of *Y* (which initially come from the "training" example companies). This then produces a score *Y* for every company, which is used to classify each company into an industry.

The disadvantage of the collaborative filtering algorithm is that it requires a relatively large number of data points and example (training) companies for consistent estimation. This requirement prevents us from separately running the collaborative filtering algorithm in each year. As a result, we use the same criteria to identify the training companies as we did to identify the *k*-means seed companies, except that we use the mean values of each company over the entire sample rather than in each year. After identifying the training companies in this way, the collaborative filtering algorithm allows us to match each firm-year observation to one of the four industries. Applied in this way, the collaborative filtering algorithm is dynamic in the sense that a company may change its classification over time.

Global Industry Classification Standard (GICS)

Our second alternative classification method is the Global Industry Classification Standard (GICS), which is a classification scheme is published by MSCI and frequently used by financial analysts. This standard defines each industry as follows:

- Pharmaceuticals (GICS 35202010): Companies engaged in the research, development or production of pharmaceuticals. Includes veterinary drugs.
- Biotechnology (GICS 35201010): Companies primarily engaged in the research, development, manufacturing and/or marketing of products based on genetic analysis

and genetic engineering. Includes companies specializing in protein-based therapeutics to treat human diseases.

North American Industry Classification System (NAICS)

Our third alternative classification method is the North American Industry Classification System (NAICS). This classification scheme is a widely used but relatively newer classification, and is the standard used by federal statistical agencies. The following table describes the NAICS codes corresponding to each companies in our sample, and whether it corresponds to either pharma or biotech:

NAICS Code	NAICS Description	Correspondence
325412	Pharmaceutical Preparation Manufacturing	Pharma
541710	Scientific Research and Development Services	Biotech
325414	Biological Product (except Diagnostic) Manufacturing	Biotech
325411	Medicinal and Botanical Manufacturing	Pharma
325413	In-Vitro Diagnostic Substance Manufacturing	Biotech
424210	Drugs and Druggists' Sundries Merchant Wholesalers	Pharma
541711	Research and Development in Biotechnology	Biotech

All codes in our sample apart from these are classified as “other”.

Standard Industrial Classification (SIC)

Our fourth alternative classification system is the Standard Industrial Classification (SIC) system. SIC codes are an older classification system that was established in 1937. However, it is still very widely used as the typical way to classify companies. As the classification system is older, it is often difficult to cleanly classify newer or emerging industries using it. Nonetheless, we use the following correspondences to classify each company as pharma or biotech:

SIC Code	SIC Description	Correspondence
2834	Pharmaceutical Preparations	Pharma
2836	Biological Product Except Diagnostics	Biotech
8731	Commercial Physical and Biological Research	Biotech
2833	Medicinal Chemicals and Botanical Products	Pharma
2835	In Vitro and In Vivo Diagnostic Substances	Biotech
5122	Drugs, Drug Proprietaries, and Druggists' Sundries	Pharma

All codes in our sample apart from these are classified as “other”.

Unanimity, Majority Rule, Hoberg-Philips

We also explore what our results look like when we utilize a voting method between the classifications, so that our results are not unduly influenced by companies which are ambiguous cases. More specifically, we examine companies for which all of the above classification methods agree (unanimity) in a given year and also companies for which the majority of the above classification methods agree (majority rule 3/5) in a given year.

Another more recent classification method that is used in the finance and economics literature is the Hoberg-Philips industry classification system.[1,2] The classification system uses text-based analysis of product descriptions from 10-K reports, and assigns each company to an industry number. Companies that are assigned to the same industry number are therefore in the same industry, based on how similar their products are. We use the Hoberg-Philips 500 industries, which assigns companies to 500 different industries, since it provides the most fine classification partitions. There are two shortcomings of this classification method. First, while the method assigns companies to an industry number, there is not a fixed interpretation of what industry that number corresponds to. In other words, one must identify which industry numbers correspond to pharma and which industry numbers correspond to biotech. For the purposes of our analysis, we classify industry numbers 32, 319, and 207 as pharma (based on a manual inspection of the companies included in those industries), and the other companies in our sample that are assigned codes as biotech. The second shortcoming of this classification method is that not every company in our sample is classified, partly because the industry data only go back to 1996. Thus, a significant number of companies will be omitted from the sample.

Return Calculation Methodology

In this section, we provide the methodology for our various empirical analyses.

To examine the cumulative returns of the pharma and biotech industries, we form portfolios of stocks for each industry and examine the performance of each portfolio over time. As is standard in the financial economics literature, we form a value-weighted portfolio of pharma stocks and a value-weighted portfolio of biotech stocks to proxy for an investment in an industry as a whole. In each portfolio, the stock return of each pharma or biotech firm is weighted by the firm's respective market capitalization. In other words, the return of portfolio P in month t is given by:

$$R_{P,t} = \sum_{i=1}^{N} \left(\frac{MV_{i,t-1}}{\sum_{i=1}^{N} MV_{i,t-1}} \right) \times R_{i,t}, \quad (1)$$

where R represents returns, i indexes firms in portfolio P, and $MV_{i,t-1}$ is the market capitalization of firm i (the stock price multiplied by the number of shares outstanding) as of the previous period. Our portfolios are reconstituted and rebalanced quarterly in order to match the calculation assumptions of the comparable market portfolio. This means that

while the weights $\left(MV_{i,t-1}/\sum_{i=1}^{N} MV_{i,t-1}\right)$ vary across time as the market values of companies vary, the weights for each firm are re-calculated at the beginning of each quarter in order to reflect the entry of new companies into the portfolio. Cumulative returns over a period that runs from time 1 to T are calculated via the formula:

$$R_{P,\{t=1,\ldots,T\}} = \prod_{t=1}^{T}\left(1 + R_{P,t}\right). \quad (2)$$

The annualized mean returns are calculated via the following formula for a given period:

$$Ann.\ Mean\ Return\ = R_{P,\{t=1,\ldots,T\}}^{(12/T)} - 1. \quad (3)$$

For our volatility calculations, we use (1) to calculate daily portfolio returns, and take the standard deviation of those returns in a given time period. Volatilities are annualized by multiplying by $\sqrt{252}$, given 252 trading days in a year.

Finally, Sharpe ratios are defined for a time period t as the expected return of the portfolio P in excess of the risk-free interest rate in period t, divided by the volatility of the portfolio P in period t:

$$SR_{P,t}\ = \frac{\mathbb{E}\left[R_{P,t}\right] - rf_t}{\sigma_{P,t}}. \quad (4)$$

We compute Sharpe ratios for each period using monthly data. Risk-free interest rate data are taken from Kenneth French's website.

Risk Calculation Methodology

For the risk characteristics, we first run the following regression each year for each portfolio (pharma or biotech) P:

$$\boldsymbol{R_{P,t} - rf_t = \alpha + \beta_P\left[R_{M,t} - rf_t\right] + \varepsilon_{i,t}}, \quad (5)$$

where $\boldsymbol{R_{M,t}}$ is the return on the value-weighted market portfolio; $\boldsymbol{rf_t}$ is the risk-free interest; and β_P is the beta, which captures the portfolio's co-movement with the market**.** (5) is estimated each year using the prior two years of daily data for each portfolio (we use daily instead of monthly returns to obtain more precise estimates). α in equation (5) is the estimate of the CAPM alpha, which is the deviation of the return predicted by the CAPM.

For the volatility calculations, we form the value-weighted portfolios of stocks in the pharma and biotech industries. We calculate total return volatility for each year by taking the standard deviation of that year's daily portfolio returns. These measures are annualized by multiplying by $\sqrt{252}$ (since there are typically 252 trading days within a year). We use a standard variance decomposition to split the total variance of each portfolio into systematic and idiosyncratic components:

$$Total\ Variance = Systematic\ Variance + Idiosyncratic\ Variance$$

$$\sigma_P^2 = \beta_P^2 \sigma_M^2 + \sigma_\epsilon^2, \quad (6)$$

where σ_P^2 is the total portfolio variance, β_P is the beta of the portfolio calculated via (5), and σ_M^2 is the variance of the market's returns. Thus, the systematic risk is given by $\beta_P^2 \sigma_M^2$, while the idiosyncratic risk is given by the residual, $\sigma_\epsilon^2 = \sigma_P^2 - \beta_P^2 \sigma_M^2$.

For brevity, we do not present the volatility and variance decomposition results using the alternate classification methods, as the results are very similar to the results with the *k*-means classification. The volatility for biotech is consistently higher than that for pharma when viewed through all of the different classifications, and moreover, the variance decomposition results are also very similar. These results are available upon request.

2 Supplementary Data

To construct our dataset, we focus on all companies that are broadly either in the pharmaceutical (pharma) or biotechnology (biotech) industries that are in the Wharton Research Data Services (WRDS) CRSP/Compustat Merged database. This list of companies, however, includes firms that do business in fields that are unrelated to medical development, but are still classified as either pharma or biotech firms. For example, there are a small number of firms that do agricultural biotechnology research, and are classified as biotech firms. We exclude such firms since they are different from the firms which we are primarily focused on, which do research related to medical purposes. In addition, some firms engage in substantial activities outside of biotech R&D (engaging in such research through an acquisition, for example). We exclude these firms to be conservative, and to avoid confounding our results with firms that also have considerable operations in other non-medical-related industries. *Supplementary Table 1* lists the companies that we exclude from our sample, as well as the specific reason for excluding them.

We extract stock return data for the remaining publicly traded pharma and biotech companies from 1930 to 2015. We use monthly instead of daily stock returns for most of our analysis to avoid any potential issues related to days with zero returns. However, for the calculation of our risk characteristics—volatility, alphas, and betas—we use daily stock return data for more accuracy. Our final sample consists of a total of 1,066 unique firms, for 125,277 firm-month observations (2,585,900 firm-day observations). We take the market portfolio return data from CRSP, and other factor data as well as risk-free interest rate data from Kenneth French's website.[1] We construct returns for the technology sector for comparison, which is a value-weighted portfolio of GICS sector 45 (Information Technology) stocks.

Supplementary Table 1: Excluded Firms

Sector	Permno	Company Name	Reason for Exclusion
Pharma	10347	Benda Pharmaceutical Inc.	Not directly biotech-related: laser precision technology
Pharma	10735	Squibb Corp.	Not medicine-focused: baby food, nutritional products
Pharma	10820	Vestar Inc.	Not medicine-focused: nutrional products
Pharma	11624	Alcide Corp.	Veterinary pharmaceuticals
Pharma	24110	Clinical Sciences Inc.	Not focused on drug development: Toxicology drug testing
Pharma	25786	Computer Memories Inc.	Not medicine-focused: computer hardware
Pharma	33655	Evergood Products Corp.	Not medicine-focused: nutritional products
Pharma	45604	Squibb Beech Nut Inc.	Not medicine-focused: baby food, nutritional products
Pharma	59985	Shaklee Corp.	Not medicine-focused: nutrition, weight management
Pharma	65832	A L Labs Inc.	Not medicine-focused: agriculture and environmental testing
Pharma	75228	Inland Vacuum Inds Inc.	Not medicine-focused: vacuum industry
Pharma	79305	Rexall Sundown Inc.	Not medicine-focused: nutrional products
Pharma	79469	Amrion Inc.	Not medicine-focused: nutrional products
Pharma	89552	Leiner P Nutritional Prods. Corp.	Not medicine-focused: nutrional products
Pharma	91675	Obagi Medical Products Inc.	Not medicine-focused: cosmetic products
Biotech	13038	Burcon Nutrascience Corp.	Not medicine-focused: agriculture and nutritional products
Biotech	13435	American Monitor Corp.	Not focused on drug development: decision support systems
Biotech	29728	Diagnon Corp.	Agricultural and veterinary pharmaceuticals
Biotech	34630	New Mexico & Arizona Land Co.	Not medicine-focused: global resource development
Biotech	43618	Immuno Nuclear Corp.	Not medicine-focused: device/software manufacturer
Biotech	45381	International Research & Dev Corp.	Not medicine-focused: agricultural biotech
Biotech	61210	Paraho Development Corp.	Not focused on drug development: instrument manufacturing
Biotech	65082	Teleconcepts Corp.	Not focused on drug development: telephone development
Biotech	80379	International Canine Genetic Inc.	Not medicine-focused: nutrional products
Biotech	81269	P D G Remediation Inc.	Not focused on drug development: software systems
Biotech	83527	Microcide Pharmaceuticals Inc.	Not medicine-focused: consumer wash products
Biotech	83618	Apache Medical Systems Inc.	Not focused on drug development: medical software coding
Biotech	83658	Fusion Medical Technologies Inc.	Not focused on drug development: surgical devices
Biotech	84521	Casmyn Corp	Not focused on drug development: mining operations
Biotech	85564	Medical Science Systems Inc.	Not focused on drug development: genetic testing services
Biotech	85672	Agritope Inc.	Not medicine-focused: agricultural biotech
Biotech	86278	Medcare Technologies Inc.	Not focused on drug development: medical supplies
Biotech	86767	Bioshield Technologies Inc.	Not focused on drug development: HVAC coating
Biotech	87769	Forbes Medi Tech Inc.	Not medicine-focused: nutritional products
Biotech	88380	Tioga Technologies Ltd.	Not focused on drug development: software/system solutions
Biotech	88804	Luminent Inc.	Not focused on drug development: fiber-optics communication
Biotech	88846	Biolab Inc.	Not focused on drug development: pool/spa water treatment
Biotech	92225	Synthetech Inc.	Not focused on drug development: material manufacturing
Biotech	92445	Nanosphere Inc.	Not focused on drug development: testing systems and devices
Biotech	93345	Codexis Inc.	Not focused on drug development: manufacturing services

3 Supplementary Notes

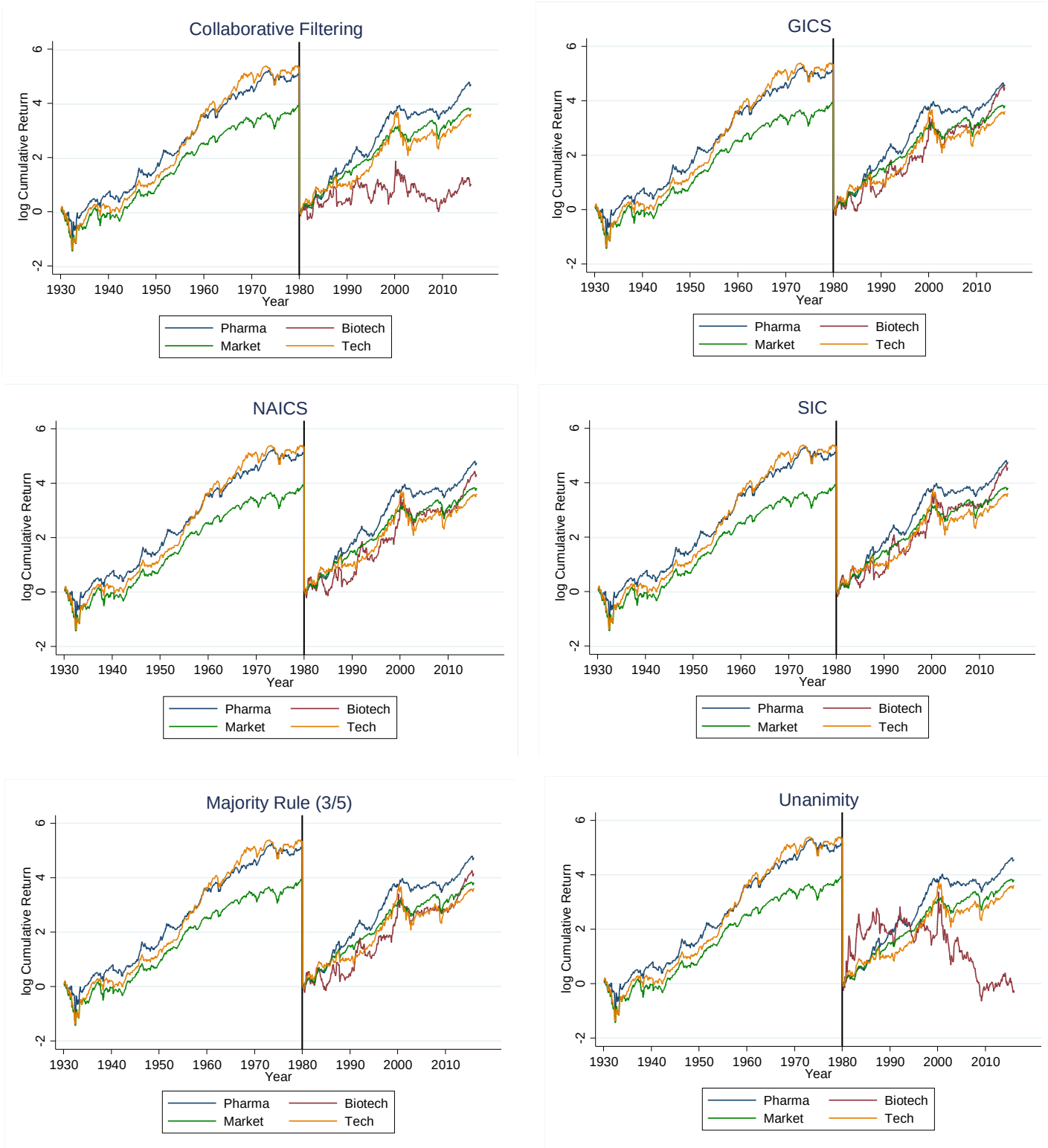
In this section, we provide additional results to supplement our main analysis.

Return and Risk Results using Alternative Classifications

Supplementary Figure 1 gives the cumulative stock returns from 1930 to 2015 for the pharma and biotech portfolios using the six alternative classification methods. For better comparability, the figure segments the sample and resets all of the cumulative returns to zero at 1980, which is when the biotech sample starts.

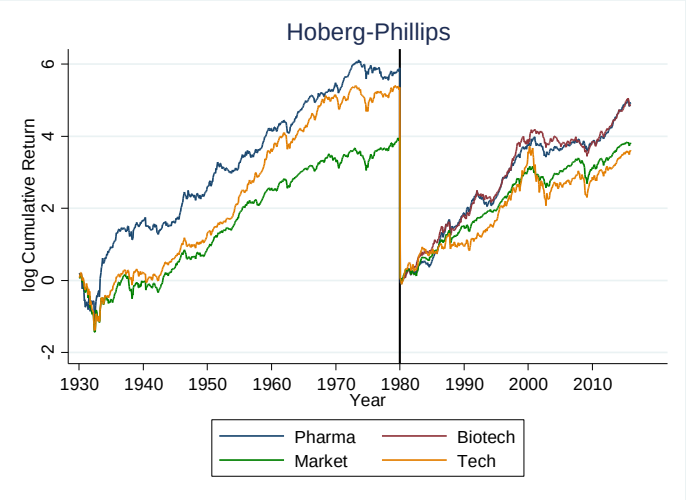
For each of the classification methods, the pharma portfolio returns are very similar. In particular, the pharma portfolio outperforms the market over the time period. However, the relative performance of the biotech portfolio varies substantially, depending on the particular classification method that is used. For example, collaborative filtering has biotech underperforming pharma, the market, and the tech sector These results are in line with those using the *k*-means classification system in the main text, except biotech's relative underperformance is even more pronounced. In contrast, with the GICS, SIC, and NAICS classifications, the returns of biotech are roughly in line (or slightly below) those of tech but underperform both pharma and the market until around 2000, after which the performance of biotech improves—by 2010 it is in line with the returns of the market, and after 2010 it overtakes the market and catches up to pharma. A majority rule classification produces returns that are close to those of the GICS, SIC, and NAICS classifications. If one only considers unanimity when classifying companies, biotech consistently underperforms all of the other indexes, to an even greater extent than the k-means or collaborative filtering classifications. used in the main text. Finally, the Hoberg-Philips classification has biotech performing very similarly to pharma. However, a shortcoming of the unanimity and Hoberg-Phillips classifications both exclude a number of firms in the sample—there are many firms for which not all of the five classification methods agree, and the sample window for Hoberg-Phillips prevent all firms from being include. The main takeaway is that the figure highlights how different the cumulative returns may be, depending on the particular classification method used.

Supplementary Figure 1: Cumulative Returns over Time, Pre- and Post-biotech



Nature Biotechnology: doi:10.1038/nbt.4023

Supplementary Figure 1: Cumulative Returns over Time, Pre- and Post-biotech



Returns for the biopharmaceutical sector. Returns are plotted over several time periods and compared to the returns of overall market and the technology sector, both taken from CRSP. Cumulative returns are plotted (on a logarithmic scale) comparing the entire pharma, tech and biotech sectors (classified according to the indicated method) to the market in two distinct time periods—from 1930 to 1980 (pre-biotech) and from 1980 to 2015 (post-biotech). The sample is segmented in this way because 1980 is the first year in which the data permit a distinction between pharma and biotech firms, thus yielding reasonable benefits from the averaging process and facilitating a fairer comparison between the groups for the pre- and post-biotech periods.

Nature Biotechnology: doi:10.1038/nbt.4023

Overall, the different classification methods paint a very different picture regarding the relative performance of the biotech industry. The main differences between the classification methods, however, can be attributed to a small number of companies with very high returns over the sample period, and the static nature of the GICS, NAICS, and SIC classifications. For example, the difference between the returns of the k-means/collaborative filtering classification and the GICS/NAICS/SIC classifications for the biotech industry can be attributed mainly to whether certain companies are classified as pharma or biotech. Two companies in particular are especially important in this respect: Gilead and Amgen. While these companies are always classified as biotech firms under the GICS/NAICS/SIC classifications, they begin as biotech companies under *k*-means but then switch to being pharma companies after they expand, and they are classified as pharma companies under collaborative filtering. If these companies were to be considered biotech companies throughout the sample under either *k*-means or collaborative filtering, the resulting pattern of returns would closely match those of the GICS/NAICS/SIC classifications. These companies affect the biotech portfolio substantially because most companies in that portfolio are smaller, with lower market capitalizations, and so adding relatively large companies to the portfolio will have a larger impact on the portfolio's returns (since they are value-weighted). Companies that gain in size and perform well will thus have an outsized impact on the biotech portfolio. In contrast, the pharma portfolio has many larger companies with higher market capitalizations that all of the classifications consider to be pharma companies, and will thus be less affected by the inclusion or exclusion of a small number of companies.

These results underscore that, while pharma seems to consistently outperform the market, the relative performance of biotech is sensitive to the particular classification method used and whether particular companies are considered to be pharma or biotech companies—biotech may thus outperform both the market and pharma if these outlier companies are included, or underperform both if these outlier companies are excluded. However, the results in the main text suggest that, barring these few companies for which there may be uncertainty regarding their industry classification, the performance of other companies in the biotech sector has been comparatively poor.

Supplementary Table 2 provides the annualized mean returns for 5-year subperiods using the different classification methods. Similar to the previous section, the performance of the pharma portfolio is very similar between the different classifications, including the k-means classification method used in the main text. In the periods before 1980, the pharma portfolio generally outperforms the market portfolio, although there are three periods where it underperforms. Starting in the mid-1980s, the pharma portfolio consistently matches or outperforms the market portfolio, although the magnitude of the outperformance varies across subperiods. For the biotech portfolio, however, the relative performance differs substantially between the different classification methods. For example, with the collaborative filtering method, biotech substantially underperforms the market in

every subperiod other than the most recent subperiod from 2010 to 2015 (where it matches the market). With GICS, NAICS, and SIC codes, biotech's performance is worse than pharma prior to 1990, but is subsequently better than pharma from 1995-1999. While biotech's performance is roughly in line with pharma from 2000-2009, according to these three classification methods, it is substantially higher than that of pharma in the most recent period from 2010 to 2015. Majority rule (3/5) also paints a very similar picture, partly due to the high level of agreement between the GICS, NAICS, and SIC classifications. Finally, the performance of biotech is the worst overall with the unanimity method, with biotech posting mostly negative returns and underperforming both pharma and the market in every subperiod except for the first period from 1980-1984.

The overall takeaway is that, while pharma generally outperforms the market, this also depends on the particular subperiod—it does not always outperform the market and underperforms the market in some subperiods. Moreover, how great the outperformance is varies between subperiods. Biotech's performance relative to pharma and the market varies substantially between classification methods and subperiods, with some classification methods showing biotech outperforming pharma and the market in many periods, and other methods showing biotech underperforming. As a result, whether biotech fares well is more uncertain, and this underscores how the performance is sensitive to which firms are classified as biotech firms, as discussed previously.

Supplementary Table 2: Annualized Mean Returns using Alternate Classifications

Time Period		GICS		SIC		NAICS		Collaborative Filtering		Unanimity		Majority Rule (3/5)		Hoberg-Phillips	
	Market	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech
1930-1934	-10%	2%	-	2%	-	2%	-	2%	-	2%	-	2%	-	20%	-
1935-1939	10%	15%	-	15%	-	15%	-	15%	-	15%	-	15%	-	16%	-
1940-1944	9%	2%	-	2%	-	2%	-	2%	-	2%	-	2%	-	-2%	-
1945-1949	11%	17%	-	18%	-	17%	-	17%	-	18%	-	17%	-	22%	-
1950-1954	22%	17%	-	16%	-	17%	-	17%	-	16%	-	17%	-	15%	-
1955-1959	15%	26%	-	27%	-	26%	-	26%	-	27%	-	26%	-	20%	-
1960-1964	10%	11%	-	12%	-	11%	-	11%	-	12%	-	11%	-	14%	-
1965-1969	6%	12%	-	13%	-	12%	-	12%	-	13%	-	12%	-	14%	-
1970-1974	-4%	4%	-	4%	-	4%	-	4%	-	4%	-	4%	-	6%	-
1975-1979	18%	6%	-	4%	-	6%	-	6%	-	4%	-	6%	-	3%	-
1980-1984	14%	13%	-1%	13%	3%	13%	-3%	13%	4%	13%	43%	13%	-3%	11%	19%
1985-1989	18%	28%	16%	28%	13%	28%	14%	27%	6%	28%	3%	28%	13%	31%	20%
1990-1994	9%	9%	13%	9%	16%	9%	16%	9%	-3%	9%	-8%	9%	15%	9%	11%
1995-1999	27%	32%	39%	32%	43%	32%	43%	33%	18%	33%	19%	32%	43%	34%	33%
2000-2004	-1%	-1%	0%	-1%	-2%	-1%	-2%	-1%	-7%	-2%	-20%	-1%	-2%	-1%	0%
2005-2009	2%	2%	3%	3%	1%	3%	1%	2%	-8%	1%	-25%	3%	0%	3%	1%
2010-2015	11%	15%	26%	17%	25%	17%	25%	18%	11%	15%	-2%	17%	25%	20%	18%
1980-2015	11%	14%	13%	14%	13%	14%	13%	14%	3%	13%	-1%	14%	12%	15%	14%
1930-2015	9%	12%	-	12%	-	12%	-	12%	-	12%	-	12%	-	13%	-

This table provides annualized mean return estimates for various subperiods, using the indicated classification methods for pharma and biotech firms. The market is also included, for comparison. Returns are calculated using monthly stock return data.

Nature Biotechnology: doi:10.1038/nbt.4023

Supplementary Table 3 provides annualized volatility estimates for the 5-year subperiods for the various classification methods. The results are qualitatively very similar to the results with the k-means classification in the main text. The volatility for biotech is consistently higher than that for pharma when viewed through all of the different classifications.

Finally, *Supplementary Table 4* gives the Sharpe ratios of each portfolio and the overall market for each time period for the various classification methods. The Sharpe ratios are quite consistent between the different classification methods for the pharma portfolio. Before 2000, the pharma portfolio does not consistently have either a higher or lower Sharpe ratio than the market portfolio—it is higher than the market in roughly half of the subperiods, and lower than the market in the rest. Since 2000, all of the classification methods show pharma outperforming the market on a risk-adjusted basis.

The Sharpe ratios for the biotech portfolio give different results across the various classifications. For most of the alternate classifications, the Sharpe ratios for biotech are lower than those for both pharma and the market in roughly half of the subperiods. For example, the GICS classification, where biotech outperformed pharma in every subperiod since 2000, has biotech posting a lower Sharpe ratio than pharma in three out of the seven subperiods. In contrast, collaborative filtering has biotech posting a lower Sharpe ratio than pharma in all but one subperiod, which is along the same lines of the results for *k*-means in the main text. Thus, the risk-adjusted returns also suggest that the performance of biotech is dependent on the particular classification method used, and that the outliers may have an outsized impact on the results.

Supplementary Table 3: Annualized Volatilities using Alternate Classifications

Time Period		GICS		SIC		NAICS		Collaborative Filtering		Unanimity		Majority Rule (3/5)		Hoberg-Phillips	
	Market	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech
1930-1934	0.33	0.34	-	0.34	-	0.34	-	0.34	-	0.34	-	0.34	-	0.34	-
1935-1939	0.21	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-
1940-1944	0.12	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-
1945-1949	0.13	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-
1950-1954	0.09	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-	0.12	-
1955-1959	0.11	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-	0.15	-
1960-1964	0.10	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-
1965-1969	0.09	0.11	-	0.11	-	0.11	-	0.11	-	0.11	-	0.11	-	0.11	-
1970-1974	0.15	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-
1975-1979	0.11	0.14	-	0.16	-	0.14	-	0.14	-	0.14	-	0.14	-	0.14	-
1980-1984	0.14	0.15	0.28	0.15	0.28	0.15	0.29	0.15	0.32	0.15	0.63	0.15	0.29	0.16	0.18
1985-1989	0.16	0.21	0.30	0.21	0.27	0.21	0.28	0.21	0.26	0.21	0.33	0.21	0.27	0.22	0.21
1990-1994	0.11	0.17	0.25	0.17	0.27	0.17	0.27	0.17	0.22	0.18	0.32	0.17	0.26	0.18	0.18
1995-1999	0.15	0.21	0.24	0.21	0.27	0.21	0.27	0.21	0.23	0.22	0.24	0.21	0.26	0.21	0.24
2000-2004	0.20	0.23	0.40	0.23	0.43	0.23	0.43	0.23	0.35	0.24	0.43	0.23	0.42	0.24	0.27
2005-2009	0.24	0.19	0.22	0.18	0.24	0.18	0.24	0.19	0.26	0.19	0.30	0.19	0.24	0.19	0.22
2010-2015	0.16	0.15	0.22	0.15	0.22	0.15	0.22	0.16	0.27	0.15	0.32	0.15	0.22	0.17	0.18
1980-2015	0.17	0.19	0.28	0.19	0.29	0.19	0.29	0.19	0.28	0.19	0.39	0.19	0.29	0.20	0.21
1930-2015	0.17	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-	0.18	-

This table provides annualized volatility estimates for various subperiods, using the indicated classification methods for pharma and biotech firms. The market is also included, for comparison. Volatility is calculated using daily stock return data.

Nature Biotechnology: doi:10.1038/nbt.4023

Supplementary Table 4: Sharpe Ratios using Alternate Classifications

Time Period		GICS		SIC		NAICS		Collaborative Filtering		Unanimity		Majority Rule (3/5)		Hoberg-Phillips	
	Market	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech
1930-1934	-0.05	0.23	-	0.23	-	0.23	-	0.23	-	0.23	-	0.23	-	0.61	-
1935-1939	0.51	0.86	-	0.86	-	0.86	-	0.86	-	0.86	-	0.86	-	0.76	-
1940-1944	0.58	0.17	-	0.17	-	0.17	-	0.17	-	0.17	-	0.17	-	0.00	-
1945-1949	0.71	0.88	-	0.88	-	0.88	-	0.88	-	0.88	-	0.88	-	0.95	-
1950-1954	1.70	0.89	-	0.88	-	0.89	-	0.89	-	0.88	-	0.89	-	0.72	-
1955-1959	1.10	1.43	-	1.48	-	1.43	-	1.43	-	1.48	-	1.43	-	1.04	-
1960-1964	0.61	0.52	-	0.54	-	0.52	-	0.52	-	0.54	-	0.52	-	0.68	-
1965-1969	0.18	0.53	-	0.60	-	0.53	-	0.53	-	0.60	-	0.53	-	0.62	-
1970-1974	-0.48	0.02	-	0.02	-	0.03	-	0.02	-	0.02	-	0.02	-	0.09	-
1975-1979	0.73	0.04	-	-0.01	-	0.04	-	0.04	-	-0.01	-	0.04	-	-0.08	-
1980-1984	0.26	0.18	-0.16	0.22	-0.07	0.18	-0.17	0.19	0.00	0.21	0.61	0.22	-0.17	0.08	0.52
1985-1989	0.69	1.02	0.42	1.02	0.33	1.01	0.35	0.97	0.14	1.03	0.12	1.01	0.33	1.13	0.67
1990-1994	0.36	0.31	0.39	0.30	0.46	0.31	0.47	0.30	-0.10	0.33	-0.13	0.30	0.44	0.28	0.42
1995-1999	1.39	1.32	1.05	1.36	1.13	1.36	1.13	1.46	0.52	1.33	0.51	1.35	1.12	1.46	1.28
2000-2004	-0.15	-0.13	0.11	-0.11	0.07	-0.11	0.07	-0.13	0.01	-0.15	-0.04	-0.11	0.06	-0.12	-0.02
2005-2009	0.02	0.05	0.13	0.11	-0.01	0.11	0.00	0.00	-0.36	-0.01	-0.93	0.10	-0.04	0.12	-0.02
2010-2015	0.87	1.26	1.36	1.32	1.35	1.33	1.33	1.35	0.54	1.26	0.08	1.31	1.32	1.39	1.19
1980-2015	0.47	0.58	0.41	0.62	0.42	0.61	0.39	0.61	0.11	0.57	0.13	0.61	0.38	0.62	0.58
1930-2015	0.40	0.50	-	0.51	-	0.51	-	0.51	-	0.50	-	0.51	-	0.53	-

This table provides annualized Sharpe ratio estimates for various subperiods, using the indicated classification methods for pharma and biotech firms. The market is also included, for comparison. Sharpe ratios are calculated using monthly stock return data.

Nature Biotechnology: doi:10.1038/nbt.4023

4 Supplementary Results

Additional Beta Results

While the CAPM alpha provides an estimate of returns above those associated with the market, there are other aggregate risk factors that investors demand to be compensated for in the form of higher returns. To account for these factors, we also estimate alphas relative to the Fama-French risk factors—size and value (market-to-book)—using the following regression:[3]

$$\boldsymbol{R_{P,t}} - \boldsymbol{rf_t} = \boldsymbol{\alpha} + \boldsymbol{\beta_{P,t,mkt}}\left[\boldsymbol{R_{M,t}} - \boldsymbol{rf_t}\right] + \boldsymbol{\beta_{P,t,SMB}}\boldsymbol{R_{SMB,t}} + \boldsymbol{\beta_{P,t,HML}}\boldsymbol{R_{HML,t}} + \epsilon_{\boldsymbol{i,t}}, \quad (7)$$

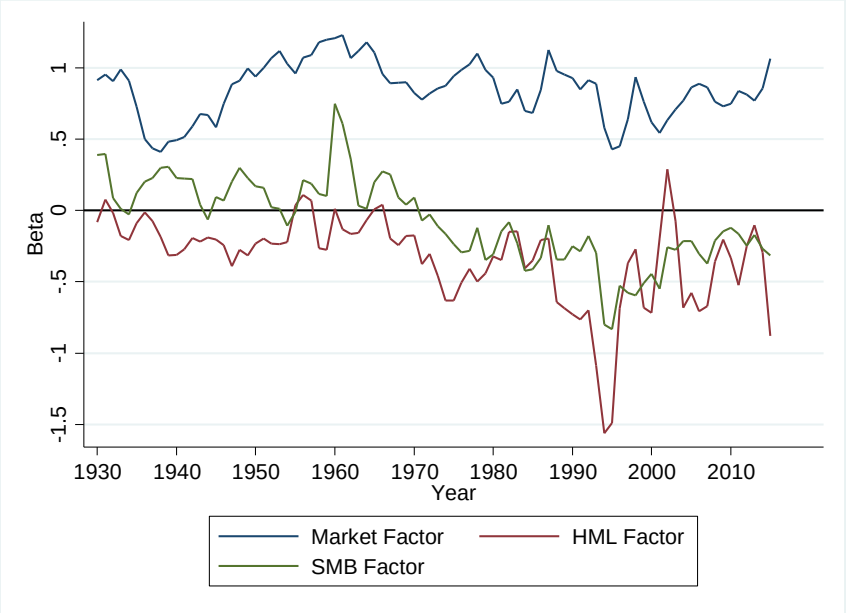
Where $\boldsymbol{R_{SMB,t}}$ is the return of the SMB (small minus big, i.e. size) portfolio and $\boldsymbol{R_{HML,t}}$ is the return of the HML (high minus low, i.e. value) portfolio. Estimates of these returns are taken from Kenneth French's website.[4] As before, (7) is estimated each year using the prior two years of daily data for each portfolio.

These beta estimates are provided in *Supplementary Figure 2* for the *k*-means classification method. These results are consistent with the previous results using only market betas. In particular, even when controlling for the value (HML) and size (SMB) systematic factors, each of the portfolios has a substantial market beta, and the market beta of the biotech portfolio is still higher than that of the pharma portfolio.

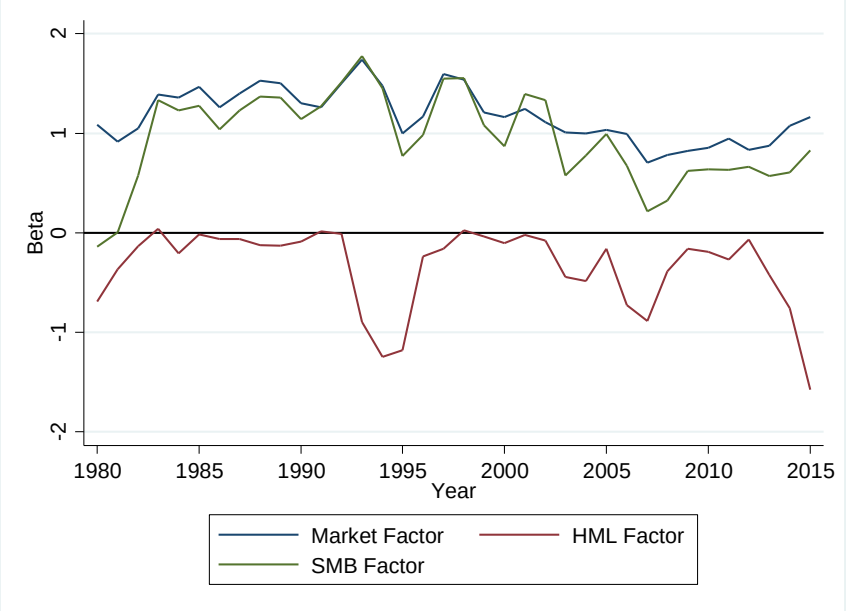
A potential concern with the interpretation of these betas is that they may be due to financial leverage—a higher amount of debt makes the equity of a company more risky, leading to a higher beta. *Supplementary Figure 3* compares the mean and median leverage ratio of pharma and biotech firms over the sample period. Apart from a one-year spike in biotech mean leverage during the financial crisis of 2008, biotech firms have very consistently had *lower* leverage ratios than pharma firms. This is true when looking at both mean and median leverage ratios. Given the lower leverage ratios of biotech firms, this suggests that the higher betas and systematic risk of these firms compared to pharma firms is not simply due to a financial leverage effect.

Supplementary Figure 2: Fama-French Beta Estimates

Panel A: Pharma Value-weighted Portfolio Beta Estimates



Panel B: Biotech Value-weighted Portfolio Beta Estimates

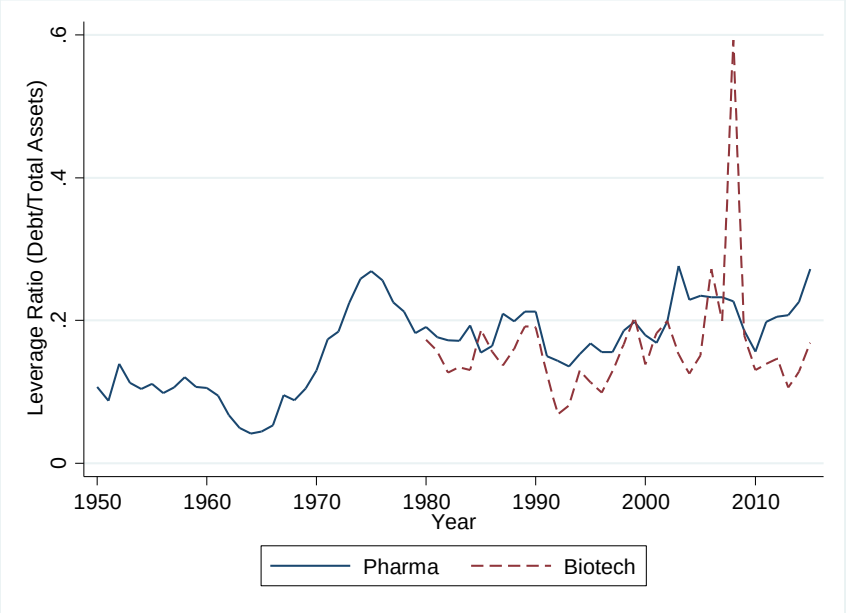


This figure shows the Fama-French three-factor beta estimates of the pharma and biotech value-weighted portfolios. Panel A gives the results for the pharma portfolio, while Panel B gives the results for the biotech portfolio. In each panel, the blue line is the market factor, the green line is the SMB (size) factor, and the maroon line is the HML (value) factor. The pharma and biotech portfolios are classified using the k-means algorithm.

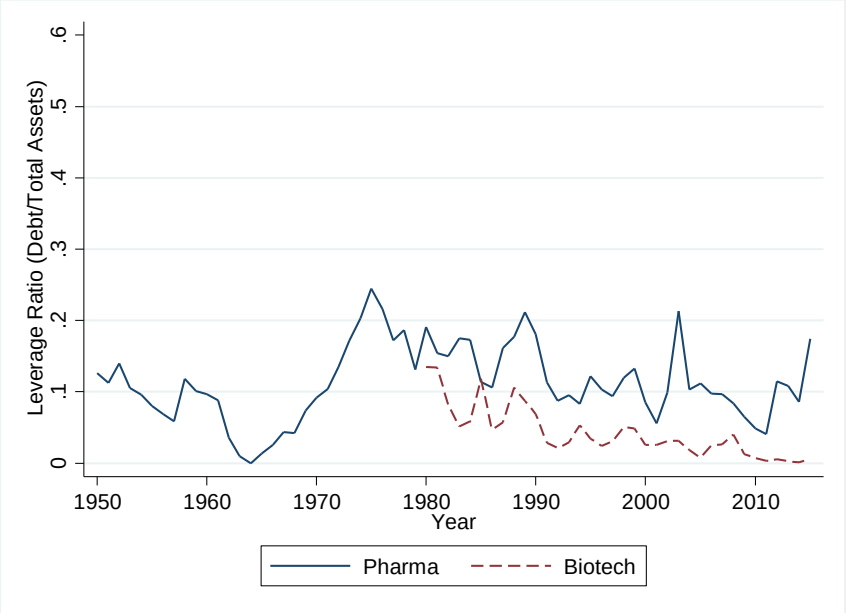
Nature Biotechnology: doi:10.1038/nbt.4023

Supplementary Figure 3: Leverage Ratios

Panel A: Mean Leverage Ratios



Panel B: Median Leverage Ratios



This figure depicts mean and median leverage ratios for pharma and biotech firms. Panel A gives mean leverage ratios, while Panel B gives median leverage ratios. In each panel, the solid blue line represents pharma firms, while the dashed red line represents biotech firms.

Nature Biotechnology: doi:10.1038/nbt.4023

CAPM and Fama-French Alphas using Alternative Classifications

We now provide CAPM alpha estimates for the pharma and biotech industries using the six alternative classification methods. The alpha estimates are calculated via equation (5). *Supplementary Table 5* gives alpha estimates for 5-year subperiods using each of the methods. The results for pharma are very similar to those for the *k*-means method. However, for the SIC, GICS, NAICS, and majority rule methods, biotech posts a positive and significant alpha from 2010-2015, but an insignificant alpha overall from 1980 to 2015. This is in contrast to the *k*-means method, where biotech posted negative alphas for a number of subperiod as well as overall since 1980. Similarly, collaborative filtering posted a negative and significant alpha from 1980 to 2015. This reinforces the notion that the performance of biotech, even when adjusted for risk, is contingent on the particular classification method used and that including the few high-performing outlier companies previously discussed can change some of the conclusions.

We also provide alpha estimates for the Fama-French model estimated via equation (7). These estimates are included in *Supplementary Table 6*. Similar to the regressions with the CAPM market factor, over the entire sample period and for all the classification methods, the pharma industry has a positive and significant alpha. This is again consistent with the findings of previous studies, and suggests that the pharma industry has maintained excess returns over a long horizon above the market and other factors.[5] However, once again the results show that the pharma industry does not consistently maintain these abnormal alphas over every time period. Similar to the results using the CAPM, for the *k*-means algorithm the biotech industry once again has a negative though insignificant alpha from 1980 to 2015. Collaborative filtering again posts a negative and significant alpha for biotech. However, in contrast to the CAPM results, biotech posts positive and significant alphas via the GICS, SIC, NAICS, and Majority Rule classifications. This again underscores that, even on a risk-adjusted basis, the relative performance of biotech is dependent on the particular classification method used.

Supplementary Table 5: CAPM Alpha Estimates using Alternate Classifications

Time Period	GICS		SIC		NAICS		Collaborative Filtering		Unanimity		Majority Rule (3/5)		Hoberg-Phillips	
	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech
1930-1934	0.104		0.104		0.104		0.104		0.104		0.104		0.104	
1935-1939	0.091*		0.091*		0.091*		0.091*		0.091*		0.091*		0.091*	
1940-1944	-0.023		-0.023		-0.023		-0.023		-0.023		-0.023		-0.023	
1945-1949	0.069*		0.071*		0.069*		0.069*		0.069*		0.069*		0.069*	
1950-1954	-0.034		-0.026		-0.034		-0.034		-0.034		-0.034		-0.034	
1955-1959	0.086**		0.096**		0.086**		0.086**		0.086**		0.086**		0.086**	
1960-1964	0.013		0.021		0.013		0.013		0.013		0.013		0.013	
1965-1969	0.059*		0.071**		0.059*		0.059*		0.059*		0.059*		0.059*	
1970-1974	0.104***		0.106**		0.105***		0.103***		0.103***		0.103***		0.103***	
1975-1979	-0.109***		-0.124***		-0.109***		-0.109***		-0.109***		-0.109***		-0.101***	
1980-1984	-0.002	-0.116	0	-0.073	-0.002	-0.121	-0.005	-0.064	0.000	0.49	0	-0.122	-0.017	0.056
1985-1989	0.066*	-0.02	0.066*	-0.042	0.063*	-0.033	0.058*	-0.093	0.069*	-0.104	0.063*	-0.042	0.087**	0.013
1990-1994	0.008	0.044	0.005	0.075	0.007	0.08	0.004	-0.098	0.01	-0.116	0.004	0.067	0.002	0.031
1995-1999	0.039	0.078	0.04	0.108	0.04	0.109	0.045	-0.048	0.042	-0.02	0.039	0.108	0.05	0.047
2000-2004	0.005	0.075	0.009	0.067	0.009	0.068	0.005	-0.016	-0.005	-0.128	0.009	0.065	0.005	0.031
2005-2009	0.004	0.026	0.016	-0.007	0.016	-0.007	-0.004	-0.087	-0.006	-0.247***	0.011	-0.014	0.017	-0.006
2010-2015	0.063*	0.146**	0.074**	0.141**	0.075**	0.136**	0.081**	-0.005	0.062*	-0.123	0.073**	0.134**	0.099**	0.083*
1980-2015	0.042*	0.039	0.046**	0.042	0.046**	0.036	0.041**	-0.057*	0.041*	-0.057	0.045**	0.031	0.049**	0.052*
1930-2015	0.041***		0.042***		0.042***		0.04***		0.04***		0.041***		0.044***	

This table shows annualized alpha estimates of the pharma and biotech value-weighted portfolios using the various alternative classification methods. Alpha estimates are calculated via the CAPM using daily returns from the indicated subperiods. *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

Nature Biotechnology: doi:10.1038/nbt.4023

Supplementary Table 6: Fama-French Alpha Estimates using Alternate Classifications

Time Period	k-means		GICS		SIC		NAICS		Collaborative Filtering		Unanimity		Majority Rule (3/5)		Hoberg-Phillips	
	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech	Pharma	Biotech
1930-1934	0.104		0.104		0.104		0.104		0.104		0.104		0.104		0.104	
1935-1939	0.084*		0.084*		0.084*		0.084*		0.084*		0.084*		0.084*		0.084*	
1940-1944	-0.01		-0.01		-0.01		-0.01		-0.01		-0.01		-0.01		-0.01	
1945-1949	0.077*		0.077*		0.079*		0.077*		0.077*		0.077*		0.077*		0.077*	
1950-1954	-0.037		-0.037		-0.028		-0.037		-0.037		-0.037		-0.037		-0.037	
1955-1959	0.077*		0.077*		0.085*		0.077*		0.077*		0.077*		0.077*		0.077*	
1960-1964	0.029		0.029		0.04		0.029		0.029		0.029		0.029		0.029	
1965-1969	0.05		0.05		0.059*		0.05		0.05		0.05		0.05		0.05	
1970-1974	0.114***		0.114***		0.112***		0.114***		0.114***		0.114***		0.114***		0.114***	
1975-1979	-0.049*		-0.035		-0.041		-0.035		-0.035		-0.036		-0.035		-0.027	
1980-1984	0.034	-0.081	0.031	-0.163*	0.034	-0.142	0.031	-0.179*	0.03	-0.137	0.039	0.417	0.035	-0.18**	0.026	0.086
1985-1989	0.081**	-0.125**	0.083**	0.023	0.085**	0.006	0.081**	0.017	0.078**	-0.075	0.087***	-0.093	0.081**	0.006	0.113***	0.022
1990-1994	0.065	-0.062	0.061	0.081	0.058	0.123	0.059	0.129	0.058	-0.1	0.066	-0.129	0.057	0.113	0.06	0.078
1995-1999	0.064	0.002	0.055	0.099	0.056	0.132	0.056	0.133	0.062	-0.055	0.059	-0.03	0.055	0.131	0.065	0.069
2000-2004	0.009	0.015	0.002	0.115	0.006	0.122	0.006	0.124	0.012	-0.104	-0.003	-0.25*	0.007	0.118	0.014	0.027
2005-2009	0.009	-0.012	0.019	0.042	0.03	0.011	0.03	0.012	0.011	-0.083	0.009	-0.249***	0.026	0.004	0.034	0.007
2010-2015	0.044	-0.017	0.033	0.093*	0.04	0.09	0.042	0.085	0.042	-0.033	0.032	-0.144*	0.039	0.083	0.056	0.05
1980-2015	0.058***	-0.016	0.055***	0.067**	0.06***	0.071**	0.059***	0.065*	0.057***	-0.054*	0.053**	-0.06	0.058***	0.059*	0.067**	0.062**
1930-2015	0.052***		0.053***		0.055***		0.054***		0.053***		0.052***		0.054***		0.057***	

This table shows annualized alpha estimates of the pharma and biotech value-weighted portfolios using the various alternative classification methods. Alpha estimates are calculated via the CAPM using daily returns from the indicated subperiods. *** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

Nature Biotechnology: doi:10.1038/nbt.4023

References

1 Hoberg, Gerard, and Gordon Phillips. "Product market synergies and competition in mergers and acquisitions: A text-based analysis." *Review of Financial Studies* 23, no. 10 (2010): 3773-3811.

2 Hoberg, Gerard, and Gordon Phillips. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124, no. 5 (2016): 1423-1465.

3 Fama, Eugene F., and Kenneth R. French. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33, no. 1 (1993): 3-56.

4 French, Kenneth R. Data Library: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

5 Koijen, Ralph SJ, Tomas J. Philipson, and Harald Uhlig. "Financial health economics." *Econometrica* 84, no. 1 (2016): 195-242.