Opportunities and Benefits as Determinants of the Direction of Scientific Research*

Jay Bhattacharya† Mikko Packalen‡

September 3, 2010

Abstract

Scientific research and private-sector technological innovation are different in terms of objectives, constraints, and organizational forms. For example, the for-profit objective that drives private-sector innovation is absent from much of scientific research, and individual researchers have many times more control in scientific research than in private-sector innovation. These differences and the lack of *any* obvious objective that would drive the direction of scientific research raise the possibility that the direction of scientific research is exogenous in the sense that it may not be influenced by factors such as the quality of research opportunities and the expected benefit from research that not only drive private-sector innovation but also in part determine the socially optimal allocation of research. Alternatively, some–yet largely unexplored–mechanisms drive also the direction of scientific research to respond to these factors. In this paper we test these two competing hypotheses of scientific research. In particular, we examine whether the composition of medical research responds to changes in disease prevalence and research opportunities. The extent of inventive activity is measured from the MEDLINE database on 16 million biomedical publications. We match these data with data on disease prevalence. We develop and apply a method for estimating the quality of research opportunities from structural productivity parameters. Our results show that the direction of medical research responds to changes in disease prevalence and research opportunities.

JEL Classification: O31, O33, I12, L65.
Keywords: Scientific Research, Private-Sector Innovation, Induced Innovation, Research Opportunity, Technological Opportunity, Non-Profit Incentives, Medicine.

*We thank Ian Cockburn, Amy Finkelstein, Raphael Godefrey, Darius Lakdawalla, Neeraj Sood, Scott Stern, Bruce Weinberg, an anonymous referee, and participants at the NBER Summer Institute 2008 Productivity meeting for helpful comments. Previous versions of this paper are titled "Is Medicine an Ivory Tower? Induced Innovation, Technological Opportunity, and For-Profit vs. Non-Profit Innovation."

†Stanford University School of Medicine, CHP/PCOR, 117 Encina Commons, Stanford, CA 94305-6019. Email: jay@stanford.edu. Bhattacharya thanks the National Institute on Aging for funding his work on this paper.

‡University of Waterloo, Department of Economics, 200 University Avenue West, Waterloo, ON N2L 3G1. Email: packalen@uwaterloo.ca.

1 Introduction

Scientific research and private-sector technological innovation are different in terms of objectives, constraints, and organizational forms. For example, the for-profit objective that drives private-sector innovation is absent from much of scientific research. This particular difference is important in part because other differences are likely linked to it. For example, Aghion et al. (2008) view the fact that individual researchers have much more authority in scientific research than in private-sector innovation as the defining characteristic of academia, and conjecture that this difference is a consequence of the non-profit nature of scientific research.

The lack of for-profit incentives in much of scientific research is important on its own because a key virtue of for-profit allocation is that decisions made by for-profit firms must necessarily respond to changes in the market, or else risk failure. As we discuss below, there is abundant evidence that for-profit producers innovate according to market demand. Non-profit allocation, on the other hand, imposes looser budget constraints (Lakdawalla and Philipson, 2006). In principle at least, the looser constraints could divorce production decisions from demand. For example, the choice of topics in scientific research might be mainly driven by the prospect of influencing other scientists (e.g. Dasgupta and David, 1994, and Saha and Weinberg, 2008) rather than the expected benefit to society.

These differences between scientific research and private-sector innovation and the lack of *any* obvious objective that would drive the direction of scientific research raise the possibility that the allocation of scientific research effort across research fields is exogenous in the sense that it may not be influenced by factors such as the quality of research opportunities and the expected benefit from research that drive private-sector innovation and in part determine the socially optimal allocation of research. Alternatively, some–yet largely unexplored–mechanisms drive also the direction of scientific research to respond to these factors, as has been argued by Rosenberg (1982).

In this paper we examine the determinants of the direction of medical research to test

these two competing hypotheses of scientific research. In particular, we examine whether the composition of medical research responds to changes in disease prevalence and the quality of research opportunities.

The focus on medicine is virtuous in part because, while there may be good reasons to insulate some markets from the vagaries of the market, academic medicine is not such a market. Despite the evident importance that the producers of academic medicine should respond to the market (that is, to the epidemiology of patient health), there is little extant evidence that they do and the view of academic medicine as an "ivory tower" that does not respond to changes in the needs of the population persists.

Our results show that the direction of medical research responds to exogenous changes in disease prevalence, which we refer to as "the induced innovation effect". For example, we find that population aging induced increases in the prevalence of a disease increase the extent of medical research on the disease. Our results also show that the direction of medical research also responds to changes in the quality of research opportunities: an increase in the quality of research opportunities in research on a disease increases the extent of research on the disease. We refer to this as the "research opportunity effect". Our econometric identification of this effect relies on our analysis of a formal model of the optimal allocation of research effort. We also develop and apply a method for measuring the research inputs associated with different research opportunities from textual information in research publications.

To our knowledge our study is the first study of scientific research that identifies the research opportunity effect, and the first study of scientific research that uses exogenous variation to identify the induced innovation effect. We review the related literature in Section 2.2. The empirical analysis is facilitated by the disease-level match between a medical vocabulary and data on disease prevalence which we have constructed for the purposes of this study. This match enables us to use the massive indexed MEDLINE database on 16 million biomedical publications to measure innovation in medicine. To our knowledge our study is

the first to take advantage of the disease-level panel nature of this database which is rich in its information content and thereby has great potential for future research on innovation.

2 Background

2.1 Incentives in Academic Medicine

A comprehensive analysis of how the incentives and constraints faced by academic medical researchers are different from those faced by researchers in the industry is beyond the scope of this paper. Any analysis of these differences is complicated by the intertwined nature of industrial R&D and academic research. Nevertheless, to take one step in this direction, in the Background Appendix we examine the many connections between industrial R&D and academic research in the biomedical sector. The discussion is relegated to an appendix to emphasize the fact that this discussion is not one of our main contributions.

In our view in each case the balance of this evidence indicates that despite the connection, pharmaceutical innovation reflects largely the functioning of for-profit incentives and biomedical publications reflect largely the functioning of non-profit incentives. Yet, we emphasize that medicine is not a perfect test case for analyzing the determinants of non-profit production of knowledge as there are important interactions between private sector innovation and academic research. Moreover, as we have indicated above, differences between industry and academia are not limited to differences in for-profit status. Hence, any results on scientific research will not necessarily extend to non-profit production of other types of knowledge.

2.2 Related Literature

Only a handful of studies have examined the determinants of scientific research and non-profit innovation in general. Rosenberg (1982) emphasizes that private-sector technological

innovation yields important inputs to scientific research and–building on prior empirical analyses on the determinants of private-sector technological innovation–conjectures that also the direction of scientific research is in part driven by the quality of research opportunities and the expected rewards from research. Lichtenberg (1999) and Lichtenberg (2006) find a positive correlation between public biomedical funding and both disease prevalence and disease severity and between cancer prevalence and the number of biomedical publications. In contrast with these two analyses, we use exogenous variation in disease prevalence to identify the induced innovation effect, and we estimate also the research opportunity effect. Finkelstein (2004) finds that the impact of vaccine policies on the number of new patent applications is small and statistically insignificant for both non-profit and for-profit entitites.

The literature on the determinants of the direction of private-sector technological innovation is more extensive. The induced innovation hypothesis originated in Hicks (1932) and Schmookler (1966). Recent empirical studies of the induced innovation hypothesis in the pharmaceutical industry include Acemoglu and Linn (2004), which we discuss below, Finkelstein (2004), Lichtenberg and Waldfogel (2003) and Yin (2008). Our research opportunity concept corresponds to the technological opportunity concept examined by Scherer (1965) and Schmookler (1966) as well as more recently by Popp (2002).

The studies most closely related to ours are Popp (2002) and Acemoglu and Linn (2004). Popp (2002) uses data on energy prices and patenting activity across energy technologies over time and finds a positive relationship between innovation and both energy prices and technological opportunity. In contrast with this analysis, our main focus is on scientific research, we use data on biomedical publications to measure inventive activity, and we calculate the quality of opportunities from structural parameters. Acemoglu and Linn (2004) use changes in the age demographics of the population to identify a positive the induced innovation effect for pharmaceutical innovation.[1] In contrast with this analysis, we examine

[1] DellaVigna and Pollet (2007) exploit changes in the demographics of aging to study stock market returns. Newell, Jaffee and Stavins (1999) exploit the changes in energy prices and changes in the cost and energy

academic medical research, we use changes in both age demographics and obesity to identify the induced innovation effect, and we determine also the research opportunity effect.[2]

The methodology that we use to estimate the quality of research opportunities builds on the methodology developed in the studies on patenting by Caballero and Jaffe (1993), Jaffe and Trajtenberg (1996) and Popp (2002). While these analyses rely on a reduced-form estimation method, we derive the estimating equation from a model the benefit from medical research. We discuss the differences between the two approaches in Section 4.2.2. The main advantage of our more structured approach is that the probability that a given knowledge cohort is used in research depends not only on the quality of that knowledge cohort but also on the quality of other existing knowledge cohorts.

An additional methodological innovation in our analysis is that while in these existing analyses the opportunity variable is constructed from citations in patent data, we construct the opportunity variable from textual information in publications data. Because the proposed method does not rely on citations data, it expands the set of circumstances in which a measure of the quality of research opportunities can be constructed. Even when citation data is available, the set of research inputs captured by citations alone is limited. For example, citations in scientific publications seldom capture research inputs generated by private-sector technological innovation, which role Rosenberg (1982) emphasized. The proposed method thus also expands the set of research inputs which the constructed opportunity variable can reflect. In related existing research Azoulay et al. (2007a, 2009) determine the patentability of a scientist's research by comparing the textual content of the scientist's publications with the content of publications by scientists who have obtained patents.[3]

efficiency of air conditioners to examine the effect of energy prices on the direction of technological change.

[2]The previous of this paper (Bhattacharya and Packalen, 2008a) included estimates of the induced innovation effect in pharmaceutical innovation. For presentational clarity, we omit these analyses here. The analyses of aging and obesity induced innovation are related to the empirical studies on preference externalities by Waldfogel (2003) and George and Waldfogel (2003). In a companion paper (Bhattacharya and Packalen, 2008b) we calculate the welfare effect of the induced innovation externality of obesity. The reader is also referred to this companion paper for references to the medical and economic literatures on obesity.

[3]Our analysis also complements the graphical analysis of topic bursts by Mane and Börner (2004).

3 Theory

In this section we present a model of medical research and solve for the socially optimal outcome in the model. The analysis yields a description of how the socially optimal allocation of research across diseases is influenced by the quality of research opportunities and disease prevalence. The analysis also yields a relationship that enables us to estimate the structural productivity parameters that determine the quality of research opportunities in each disease, as well as a relationship that describes how the quality of research opportunities variable can be constructed from these productivity parameter estimates.

Our theoretical analysis reflects our focus on examining whether the direction of scientific research responds to changes in two characteristics–the quality of research opportunities and the expected benefit from research–that in part determine the socially optimal allocation. The analysis is agnostic about why scientific research would respond to changes in these two characteristics. We sidestep this question because the data that we use does not enable us to differentiate between different theories of scientific research such as altruism and prestige maximization (see e.g. Merton 1973 [1942], Glaeser, 2003, Stern 2004). These specific mechanisms are important but so is understanding the relationship between the direction of scientific research and characteristics that determine the socially optimal allocation.

3.1 A Model of the Social Benefit from Medical Research

3.1.1 Three Characteristics of Each Unit of Research

We assume that each unit of research is identified by three characteristics: the disease i which the research examines, the year t in which the research is conducted, and the cohort f of the research opportunities that are pursued in the research. The measurement and role of the opportunity cohort f in the analysis is explained shortly. These assumptions are, of course, simplifications as a research project in medicine does not necessarily examine only

one disease and a research project may rely on opportunities that do not all belong to the same opportunity cohort f. The discussion of how we address these issues in our empirical analysis is postponed until Section 5.3.

3.1.2 Determinants of the Benefit from Research

Consider next the benefit of research on disease i in year t that pursues research opportunities in the opportunity cohort f. We assume that this benefit depends on three factors: 1) the extent of research, 2) the number of people who benefit from the research, and 3) the quality of the research opportunities. We assume that the benefit from research does *not* depend on the severity of the disease because we lack of a source of exogenous variation in severity over time, because most research is incremental and thus need not bear a relationship with the harm from the disease to each inflicted individual, and because this assumption is consistent with the findings in Acemoglu and Linn (2004).

The first factor, the extent of the research effort on disease i in year t that pursues research opportunities in cohort f, is denoted by N_{itf}. The second factor, the expected number of people with the disease i in year t, is denoted by M_{it}. The third factor, the quality of research opportunities, captures the idea that the benefit from an inframarginal unit of research is higher when the inputs to the research process provide the researchers fertile applications compared to when the inputs to the research process hold only potential for average or below average applications.

3.1.3 Measurement of Research Opportunities from Research Inputs

Inputs to the research process can be tangible or intangible. Tangible research inputs in medicine include approved drugs developed by pharmaceutical companies. The properties of such drugs are examined in drug-related post-approval medical research. Ideas are an example of intangible research inputs, some of which are recorded as citations in publications.

However, from the perspective of an econometrician, the more important distinction is whether the presence of a particular set of research inputs can be measured. In what follows we refer to the set of observable characteristics of the opportunities pursued in research as "research inputs". The discussion of how we measure research inputs in our empirical analysis is postponed until Section 5.3.

3.1.4 Measurement of the Cohort of Research Opportunities

There are alternative ways of assigning a cohort to each research input. For example, when research inputs are ideas that are measured from citations, a natural choice for the cohort f of each research input is the year of publication of the cited publication. An alternative method is to set the cohort f of a research input as the year in which the research input was first applied either in medical research. The discussion of how we measure the cohort of research inputs in our empirical analysis is postponed until Section 5.3.

3.1.5 Determinants of the Quality of Research Opportunities

For a given disease i, year t, and opportunity cohort f combination, the quality of research opportunities depends on two factors: 1) the baseline productivity of research inputs in the opportunity cohort f in research on the disease i, which we denote by α_{if}, and 2) the elapsed time $t-f$ since the initial discovery of the research inputs in cohort f. The first factor reflects the fact that the productivity of research inputs in cohort f in research on a disease i will be low if the research inputs in the particular cohort f are relatively unsuitable for research on that particular disease. Such scenario is captured by a low value of the parameter α_{if}. The second factor reflects the fact that both the diffusion and exhaustion of knowledge is generally gradual: the productivity of research inputs in cohort f in research will be low both if the elapsed time $t-f$ from the year of discovery f of these research inputs is very high–so that most of the potential of that research input cohort has already been exhausted–and if

the elapsed time $t - f$ since the discovery of the ingredients is very low–so that most of the potential of that research input cohort has yet to be fully revealed to the researchers.

3.1.6 Functional Form for the Benefit from Research

We assume a specific functional form for the benefit from research on disease i in year t that relies on research inputs associated with the opportunity cohort f, namely

$$M_{it} \times \left\{ \alpha_{if} \times e^{-\beta_1 (t-f)} \times \left[1 - e^{-\beta_2 (t-f)} \right] + \varepsilon_{itf} \right\} \times \ln \left(N_{itf} \right), \tag{1}$$

where ε_{itf} denotes other factors that influence the benefit from research. We assume that the variable ε_{itf} is observable to medical researchers but is unobservable to the econometrician. The factor $\left[1 - e^{-\beta_2 (t-f)} \right]$ represents the lag between the discovery of the research inputs in cohort f and the time at which the full potential of the these research inputs in medical research is revealed. The factor $e^{-\beta_1 (t-f)}$ represents the eventual decay in the usefulness of the research inputs in cohort f in medical research as the associated research opportunities are gradually exhausted.[4]

The total benefit from research on the disease i in year t is the sum of the benefit (1) over all available research input cohorts f_0 through t:

$$M_{it} \sum_{f=f_0}^{t} \left\{ \alpha_{if} \times e^{-\beta_1 (t-f)} \times \left[1 - e^{-\beta_2 (t-f)} \right] + \varepsilon_{itf} \right\} \times \ln \left(N_{itf} \right). \tag{2}$$

The overall benefit from research in year t is the sum of the benefit (2) from research on disease i over all diseases:

$$\sum_{i} M_{it} \sum_{f=f_0}^{t} \left\{ \alpha_{if} \times e^{-\beta_1 (t-f)} \times \left[1 - e^{-\beta_2 (t-f)} \right] + \varepsilon_{itf} \right\} \times \ln \left(N_{itf} \right). \tag{3}$$

[4] We do not model explicitly the effect that the amount of research in the preceding years may have on the benefit from research in a given year. This assumption is innocuous if marginal research in each year does not influence the quality of research opportunities in future years.

3.2 Socially Optimal Allocation

3.2.1 Optimal Allocation across Opportunity Cohorts f within Disease i

Let N^*_{itf} denote the optimal allocation of research effort on disease i that relies on opportunity cohort f in year t. The first-order conditions for the optimum imply that

$$p^*_{itf} = \frac{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}}{\sum_{f'=f_0}^{t} \left\{\alpha_{if'} \times e^{-\beta_1(t-f')} \times \left[1 - e^{-\beta_2(t-f')}\right] + \varepsilon_{itf'}\right\}}, \tag{4}$$

where $p^*_{itf} \equiv \frac{N^*_{itf}}{\sum_{f'=f_0}^{t} N^*_{itf'}}$.[5] Equation (4) states that the share of research on disease i that relies on opportunity cohort f is equal to the ratio of the quality of the opportunity cohort f in research on the disease i and the sum of the qualities of all available research opportunity cohorts f_0 through t in research on the disease i.

3.2.2 Optimal Allocation across Diseases

When the allocation of research across opportunity cohorts f within a disease is optimal, using the previous definition $p^*_{itf} \equiv N^*_{itf} / \sum_{f'=f_0}^{t} N^*_{itf'}$ and the definition $N_{it} \equiv \sum_{f=f_0}^{t} N^*_{itf}$, the expression (3) for the overall benefit from research can be rewritten as

$$\sum_i M_{it} \sum_{f=f_0}^{t} \left\{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}\right\} \times \ln\left(N_{it} \times p^*_{itf}\right). \tag{5}$$

The first-order conditions for the optimal allocation of research effort across diseases imply that

$$N_{it} = \left(\sum_i N_{it}\right) \times \frac{M_{it} \sum_{f=f_0}^{t} \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}}{\sum_i M_{it} \sum_{f=f_0}^{t} \left\{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}\right\}} \tag{6}$$

[5] The first-order condition for the optimum is $M_{it} \times \alpha_{if} \times \frac{e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}}{N^*_{itf}} = M_{it} \times \alpha_{if'} \times \frac{e^{-\beta_1(t-f')} \times \left[1 - e^{-\beta_2(t-f')}\right] + \varepsilon_{itf'}}{N^*_{itf'}}$ for all (i,t,f,f'). Denoting $c_{itf} \equiv \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}$ this condition can be rewritten as $c_{itf} \times N^*_{itf'} = N^*_{itf} \times c_{itf'}$ for all (i,t,f,f'). Taking the sum of both sides of the equation $c_{itf} \times N^*_{itf'} = N^*_{itf} \times c_{itf'}$ over all $f' \in \{f_0, ..., t\}$ gives $c_{itf} \times \sum_{f'=f_0}^{t} N^*_{itf'} = N^*_{itf} \times \sum_{f'=f_0}^{t} c_{itf'}$ for all (i,t,f). Rearranging and using the definitions of c_{itf} and p^*_{itf} gives the relationship (4) in the text.

holds for all (t, i).

The factor $\sum_{f=f_0}^{t}\{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}\}$ in equation (6) is the sum of the qualities of available research opportunity cohorts in the disease i in year t, and we use this sum as a measure of the quality of research opportunities in research on the disease i in year t. Denoting this measure of the quality of research opportunities by K_{it}, we can rewrite equation (6) as $N_{it} = \sum_i N_{it} / \left(\sum_i M_{it} \sum_{f=f_0}^{t} K_{it}\right) \times M_{it} \times K_{it}$. Assuming that $N_{it} > 0$ and $K_{it} > 0$ for all (i, t) this can be rewritten as

$$\ln N_{it} = \ln K_{it} + \ln M_{it} + u_t, \tag{7}$$

where $u_t \equiv \ln \left[\sum_i N_{it} / \left(\sum_i M_{it} \sum_{f=f_0}^{t} K_{it}\right)\right]$.

3.3 Implications for Empirical Analysis

3.3.1 Two Determinants of the Direction of Medical Research

Equation (7) describes a proportional relationship between research effort in a disease and the quality of research opportunities, and a proportional relationship between the research effort in a disease and disease prevalence. With a different functional form for the overall benefit from research both relationships would still be positive but non-proportional. We allow for this possibility in our empirical framework (see Section 6).

3.3.2 Estimation of the Quality of Research Opportunities

Equation (4) describes a relationship between the parameters α_{if}, β_1 and β_2 that govern quality of research opportunities variable K_{it} and the probability p^*_{itf} that research on disease i in year t relies on research inputs in the research input cohort f. In Section 4 we explain how we use this predicted relationship to estimate the parameters α_{if}, β_1 and β_2 and how the measure of the quality of research opportunities is calculated from these estimates.

3.3.3 Implications when the Quality of Research Opportunities is not Estimated

As we explain in Section 5.3, we limit the scope of the analysis by constructing a measure of the quality of research opportunities only for drug-related medical research. When the role of research opportunities is omitted from the model–similar to Acemoglu and Linn (2004)–the theoretical analysis describes a proportional relationship between the extent of research effort on each disease and disease prevalence. Of course, different assumptions about the preferences would again imply that the relationship is positive but non-proportional. We allow for this possibility in the empirical analysis.

3.3.4 Changes in the Composition of Research across Types of Research

What this model of medical research and the discussion so far has not considered is that a factor that changes disease prevalence may also change the type of medical research on a disease. For example, while population aging increases the prevalence of many diseases it may also shift research effort away from a more general type of drug-related medical research on those diseases and toward research that is more focused on the physiology of those diseases in the old-age population. Similarly, while an increase in obesity increases the prevalence of many diseases it may also shift research effort away from a more general type of drug-related research on those diseases and toward research that is focused on the physiology of those diseases in the obese. As a result, changes in disease prevalence may not influence the amount of drug-related research as much as is implied by the above model. The estimated disease prevalence effect on drug-related research may even negative if the change in disease prevalence affects the composition of research within the disease but does not affect the measure of total amount of research on the disease.[6] We allow for these possibilities in the empirical analysis.

[6]Even when the total research effort in the disease changes, our measure of total research effort in a disease (publications) may not change if one type of research (say, research that examines the physiology of the disease in a subpopulation) requires more inputs than another type of research (say, drug-related research on the disease).

4 Estimation of the Quality of Research Opportunities

4.1 Calculation of the Quality of Research Opportunities

As was explained in the discussion preceding equation (7), a measure of the quality of research opportunities in research on the disease i in year t is given by the expression

$$K_{it} \equiv \sum_{f=f_0}^{t} \left\{ \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf} \right\}, \tag{8}$$

where the parameters α_{if} specify the baseline productivity of the research opportunity cohort f in research on disease i, the parameter β_1 governs the eventual decay in the research potential of the research input in any given cohort, the parameter β_2 governs the rate at which the full potential of research inputs in any given cohort is revealed to researchers, and ε_{itf} denotes other factors that influence the productivity of research that relies on research inputs in the cohort f. We assume that $E[\varepsilon_{itf}] = 0$ and that ε_{itf} is independent and identically distributed.

Provided we can obtain estimates $\hat{\alpha}_{if}$ and $\hat{\beta}_1, \hat{\beta}_2$ of the parameters α_{if}, β_1 and β_2, we can thus calculate an estimate of the quality of research opportunities in disease i in year t using the formula

$$\hat{K_{it}} \equiv E[K_{it}|\hat{\alpha}_{if}, \hat{\beta}_1, \hat{\beta}_2] \approx \sum_{f=f_0}^{t} \hat{\alpha}_{if} \times e^{-\hat{\beta}_1(t-f)} \times \left[1 - e^{-\hat{\beta}_2(t-f)}\right]. \tag{9}$$

The econometric challenge is to estimate the parameters α_{if}, β_1 and β_2.

We emphasize that our calculation of the measure of the quality of research opportunities from the estimates of these parameters is standard. Popp (2002), Caballero and Jaffe (1993) and Jaffe and Trajtenberg (1996) each use either the equation (9) or a close equivalent. The difference between our approach and these existing approaches lies in how these parameters are estimated.

4.2 Estimation of the Structural Productivity Parameters

4.2.1 The Estimating Equation

The theoretical analysis in Section 3.1 predicts that when the allocation of research effort across opportunity cohorts f within disease i is socially optimal, the relationship (4) holds between the probability p^*_{itf} that research on disease i in year t relies on the research opportunity cohort f and the unknown productivity parameters α_{if}, β_1 and β_2. Existing analyses calculate the probabilities $p^*_{itf} \equiv \frac{N^*_{itf}}{\sum_{f'=f_0}^{f} N^*_{itf'}}$ from citations in patent data whereas we calculate these probabilities from textual information in publications data (see Section 5.3).

With observations on the probabilities p^*_{itf}, the relationship (4) can be used as the basis for obtaining estimates of the parameters α_{if}, β_1 and β_2. Denoting $\alpha_{it} \equiv 1/[\sum_{f=f_0}^{t}\{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}\}]$ the relationship (4) may be rewritten as

$$p^*_{itf}/\alpha_{it} = \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}. \tag{10}$$

When $t - f_0$ is large, we have that $\sum_{f=f_0}^{t} \varepsilon_{itf} \approx 0$. Applying this simplification modifies the definition of α_{it} as follows:

$$\alpha_{it} \equiv 1/\left[\sum_{f=f_0}^{t} \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right]\right] \tag{11}$$

and also modifies the relationship (4) as follows:

$$p^*_{itf} = \frac{\alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}}{\sum_{f=f_0}^{t} \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right]}. \tag{12}$$

The above equation forms our estimating equation for the parameters α_{if} and the parameters β_1 and β_2 that govern the quality of opportunities.

4.2.2 Comparison with the Reduced-Form Estimation Approach

Using patent citation data on technological innovation, Popp (2002), Caballero and Jaffe (1993) and Jaffe and Trajtenberg (1996) estimate the parameters α_{if}, β_1 and β_2 that govern the the quality of opportunities using the reduced-form empirical model

$$p^*_{itf} = \alpha_{if} \times e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}. \tag{13}$$

Hence, the difference between this existing approach and our structural approach to estimating the parameters α_{if}, β_1 and β_2 is that in our estimation approach the probability that research relies on a given opportunity cohort f depends not only on the quality of the opportunity cohort f but also on the quality of other available opportunity cohorts–as is indicated by the presence of the denominator in our estimating equation (12)–whereas in these prior estimation approaches the probability that research relies on an opportunity cohort f only depends on the quality of the opportunity cohort f and not on the quality of other available opportunity cohorts. This feature is the main advantage of our estimation approach relative to prior approaches to estimating the quality of research opportunities.

An added advantage of our more structured approach is that potential limitations in empirical work become more transparent. For example, the analysis reveals that if the expression (2) is not a good representation of the true benefit from research, the measure of the quality of research opportunities variable K_{it} would likely be influenced by the extent of research N_{it} on the disease (see Section 7.1.2). It is therefore important to examine whether such potential reverse causality influences estimates of the "research opportunity effect". We are not aware of this issue having been considered in existing work. Also other identifying assumptions become more transparent. The quality of research opportunities variable is calculated under the assumption that allocation across cohorts within diseases is optimal, whereas the research opportunity effect itself refers to allocation across diseases.

4.2.3 An Iterative Estimation Procedure

In our empirical analysis the number of parameters a_{if} is over 5000 because we have 127 separate diseases and f ranges from 1960 to 2002. Estimating the parameters a_{if} and the parameters β_1 and β_2 using non-linear least squares and the equation (12) is therefore computationally quite demanding. Instead, we first estimate the parameters β_1 and β_2 using non-linear least squares applied to the equation (10) while assuming fixed values for the parameters α_{if} and α_{it}.[7] We then estimate the parameters α_{if} using the following iterative procedure:

1. We start by calculating initial estimates of α_{it} by plugging in the estimates of the parameters β_1 and β_2 as well as arbitrary (starting) values of α_{if} into the expression (11).[8]

2. Using the estimates of the parameters α_{it} and the estimates of the parameters β_1 and β_2, we estimate the parameters α_{if} by least squares applied to the equation (10) and holding α_{it}, β_1 and β_2 fixed.

3. We recompute the estimates of α_{it} by plugging in the estimates of α_{if} and the estimates of β_1 and β_2 into the expression (11). If the new value of the estimate of α_{it} is sufficiently close to the old value, we declare convergence. If not, we iterate the previous step until convergence.

This iterative procedure yields estimates of the parameters α_{if}. We then generate our estimate of the quality of research opportunities using the estimates $\hat{\alpha}_{if}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ and the formula (9).

[7] We assume that $\alpha_{if} = 1$ and $\alpha_{it} = 1$ for all i, t, f. The estimating equation therefore becomes $p^*_{itf} = e^{-\beta_1(t-f)} \times \left[1 - e^{-\beta_2(t-f)}\right] + \varepsilon_{itf}$. Omitting a multiplicative constant in this specification is both innocuous and necessary because the true value of the parameter β_2 is typically very small and the variation in $t - f$ is limited which make the factor $\left[1 - e^{-\beta_2(t-f)}\right]$ approximately equal to $\beta_2 \times (t - f)$ in the sample.

[8] We assume that $\alpha_{if} = 1$ for all i, f.

5 Data and Measurement of Variables

5.1 Publications Data

We measure research effort in medicine from the MEDLINE publications database on approximately 16 million biomedical publications, generally from 1950 to the present. Publications in this database are indexed by the 2007 version of the Medical Subject Headings (MESH) vocabulary, which is a hierarchical medical vocabulary of over 20000 different terms. We use these MESH codes to identify the disease or diseases examined in each publication.

We limit the analysis to the 127 diseases which we have matched to disease prevalence data (see Section 5.2). To measure the extent of the total research effort related to a disease we count the number of publications that are matched to the disease. A publication may be indexed to multiple diseases. We allow for this possibility by counting publications that are matched to more than one disease the same way we would count the matches if each match was from a separate publication. We construct three measures of drug-related medical research, denoted *DRUG 1*, *DRUG 2*, and *DRUG 3*, and three measures of other medical research, denoted *OTHER 1*, *OTHER 2*, and *OTHER 3*. The construction of these measures is discussed in the Data Appendix.

5.2 Match of Publications Data and Disease Prevalence Data

An important contribution of our analysis is the construction of a match between the MEDLINE publications data, which is indexed by the MESH medical vocabulary, and the disease prevalence data (see Section 5.4), which is indexed by the ICD-9 disease classification system. We limit this matching effort in several ways that are detailed in the Data Appendix. The match, which we included in its entirety in the previous version of this paper (Bhattacharya and Packalen, 2008a), yields 127 separate matches between a disease or a group of diseases and a MESH entry/entries. The 127 diseases belong to 12 disease classes.

5.3 Measurement of Research Inputs and their Cohort

5.3.1 Citation Based Approach to Measuring Research Inputs

Using patent data, existing research (e.g. Popp, 2002, Caballero and Jaffe, 1993, and Jaffe and Trajtenberg, 1996) has relied on citations to previous patents to determine research inputs and their cohort. In principle, an analysis of research inputs in scientific research could similarly be based on citations in scientific publications. However, such citation data are not widely available to researchers. Moreover, an important input in scientific research is private-sector technological innovation (Rosenberg, 1982), and citations in scientific publications would seldom reveal the presence of such inputs in scientific research.

5.3.2 Textual Information Based Approach to Measuring Research Inputs

An alternative approach is to determine research inputs and their cohorts from textual information in research publications. This is the approach that we develop and apply here. Related research by Azoulay et al. (2007a, 2009) is discussed at the end of Section 2.2.

To limit the scope of the required data extraction exercise, we construct the measure of the quality of research opportunities only for drug-related medical research.[9] Within drug-related medical research, there are of course many different types of research inputs. We focus on approved active ingredients (new drugs) as research inputs. To determine which ingredients are used as research inputs in each medical research publication, we search through the titles and abstracts of all publications in the MEDLINE publications database for all approved active ingredients.[10] The match of an ingredient name and a publication

[9] The measurement of research inputs associated with different research opportunities is considerably more intense an exercise for non-drug-related "other medical research" than for drug-related medical research, as only for drug-related research is an important subset of the research inputs associated with each research opportunity easily available in the form of approved active ingredients (drugs).

[10] We identify active ingredients from the Federal Drug Administration (FDA) data on drug approvals during 1939-2006. As we generally cannot distinguish between active ingredients and their derivatives in the biomedical publications data, we consider the first word of each entry in the list of approved active ingredients to be the ingredient name that we use in our study. This yields a list of 1448 ingredients. This list includes the drugs that were in use pre-FDA. We do not use the drug approval year information in the

indicates that the ingredient was an input to the research process that led to the publication.

It is important to note that pharmaceutical research that leads to the discovery of new active ingredients precedes the type of drug-related medical research examined here. When the active ingredient name is used in publications the intended therapeutic use of the drug is already known and the associated patent application has already been filed.[11] Consequently, our measure of drug-related research captures applied drug-related medical research which is mostly conducted by academic researchers and published in academic journals as opposed to basic drug development research conducted by either pharmaceutical firms or academics.

5.3.3 Measurement of the Cohort of Research Inputs in Our Analysis

We set the cohort f of each measured research input (active ingredient) to equal the year before the year in which the research input is first mentioned in the publications data. We lump together research inputs with the same cohort f. A publication may mention research inputs from multiple cohorts. We count such multiple matches from one publication the same way we would count the matches if each match was from a separate publication.

5.4 Measurement of Disease Prevalence

To construct population aging and obesity epidemic related measures of disease prevalence over time we combine cross-sectional disease prevalence data with panel data on population characteristics (see Section 6). We estimate cross-sectional disease prevalence for each age and BMI group from the Medical Expenditure Panel Survey data for years 1996-2005. We use the Surveillance Epidemiology and End Results data for years 1975-2004 to estimate the share of people in each age group in each year. For each age group we use the National Health Interview Survey data for years 1976-2005 to estimate the share of people in each

FDA data because that information is unreliable in this administrative data.

[11] Pharmaceutical manufacturers apply for an active ingredient name for a drug during phase I or phase II clinical trials which happen after the pre-clinical testing has been completed and the drug has received an investigational drug application.

BMI group in each year. Details are relegated to the Data Appendix.

5.5 Sample Period

For reasons discussed in the Data Appendix, we set 1975-2005 as the sample period in the regression analyses, and in estimating the quality of research opportunities we limit the range of research input cohorts f to 1960-2001 and the range of publication years t to 1970-2002.

6 The Empirical Models

The empirical models that we apply are based on equation (7) and the associated discussion in Section 3.3. We rely on population aging and obesity epidemic induced exogenous changes in disease prevalence to identify the "induced innovation effect" (see Section 7.2). The construction of the population aging and obesity epidemic related disease prevalence variables M_{it}^{AGING} and $M_{it}^{OBESITY}$ is discussed below.[12] The resulting regression model is

$$\ln N_{it} = \beta_K \ln \hat{K}_{it} + \beta_A \ln M_{it}^{AGING} + \beta_O \ln M_{it}^{OBESITY} + \alpha_i + \alpha_t + u_{it}. \tag{14}$$

The variable N_{it} is a measure of medical research effort (publications) on the disease i in year t. We construct a measure of the quality of research opportunities variable K_{it} only for drug-related medical research (see Section 5.3.2). However, we include this variable also in our analyses of other types of medical research because this strategy enables us to examine whether the results on the "research opportunity effect" in drug-related medical research are affected by reverse causality (see Section 7.1.2). The variable u_t is the unobserved error term. The parameters α_i and α_t represent disease and year fixed effects, respectively. Using this fixed effects specification the identifying variation for each parameter is the variation in the regressor within each disease relative to the corresponding variation within all other

[12] See the Data Appendix to Section 5.4 for the derivation and rationale for this decomposition.

diseases. We also employ an alternative fixed effects strategy in which we include disease fixed effects α_i and the disease-class specific year fixed effects $\alpha_{d,t}$. Using this alternative fixed effects specification the identifying variation for each parameter is the variation in the regressor within each disease relative to the corresponding variation within all other diseases in the same disease class. The parameters of interest will be different in the two fixed effects specifications if the elasticity of substitution of research effort is different between diseases within each disease class than it is between diseases in different disease classes.

The constructed population aging related disease prevalence variable

$$M_{it}^{AGING} \equiv \sum_{j=1}^{5} \sum_{k=1}^{3} \mu_{i,j,k} \times s_{j,t}^{AGE} \times s_{j,k,t_0}^{BMI}, \tag{15}$$

is the prevalence of disease i in year t when the body weight distribution in year t is set to be the same as the body weight distribution is in the initial year t_0 in the sample and only the age distribution varies over time. Similarly, the constructed obesity epidemic related disease prevalence variable

$$M_{it}^{OBESITY} \equiv \sum_{j=1}^{5} \sum_{k=1}^{3} \mu_{i,j,k} \times s_{j,t_0}^{AGE} \times s_{j,k,t}^{BMI}. \tag{16}$$

is the prevalence of disease i in year t when the age distribution in year t is set to be the same as the age distribution is in the initial year t_0 in the sample and only the body weight distribution varies over time. A positive estimate of the parameter β_A (parameter β_O) is therefore evidence of aging (obesity) induced research. The parameter $\mu_{i,j,k}$ is the prevalence of disease i among people in the age group j who are in the Body-Mass-Index (BMI) group k, the parameter $s_{j,t}^{AGE}$ is the share of people in the age group j in year t, and the parameter $s_{j,k,t}^{BMI}$ is the share people in the age group j who are in the BMI group k in year t.[13]

[13]The parameters $\mu_{i,j,k}$, $s_{j,t}^{AGE}$ and $s_{j,k,t}^{BMI}$ are estimated from data on the disease prevalence and from data on demographics (see Section 5.4). The age groups are 0-18, 18-35, 35-50, 50-65 and 65+. The BMI groups are 18.5-25, 25-30 and 30-50. See the Data Appendix to Section 5.4 for an explanation on why age group 0-18 is excluded here and the implications. As we use disease and year fixed effects we can ignore population size and population growth in estimating disease prevalence. Due to space constraints and to keep the analysis accessible we also relegate the discussion of several additional issues to the Data Appendix to Section 5.4.

7 Identification Strategy

7.1 Identification of the Research Opportunity Effect

7.1.1 Correlated Unobservables

Variation in the quality of research opportunities is likely to be correlated with the variation in the unobserved determinants of medical research over time. Accordingly, we employ fixed effects approaches in which the research opportunity effect is identified by comparing changes in the quality of research opportunities across diseases (either across all diseases or across diseases in the same disease class) with the corresponding changes in research effort.

7.1.2 Reverse Causality

With a different functional form for the benefit from medical research the optimal distribution of research effort across opportunity cohorts within a disease would depend on the extent of research on the disease. Consequently, changes in the level of research on a disease would impact estimates of the quality of research opportunities. There might thus be a positive empirical relationship between these two variables even if there was no causal effect from the quality of research opportunities on the extent of research effort.

To address this potential concern we take advantage of the fact that we examine two categories of medical research, namely drug-related medical research and other medical research. The unobserved effects that influence the level of drug-related research effort and the unobserved effects that influence the level of other types of medical research effort are likely to be correlated. Consequently, if there is indeed reverse causality from the level of drug-related research effort to the measure of quality of research opportunities in drug-related research, this measure of the quality of research opportunities will likely also be correlated with the level of other medical research. In contrast, if there is no reverse causality from the level of drug-related research effort to the measure of the quality of research opportunities, this

measure of the quality of research opportunities will likely be uncorrelated with the level of other medical research. We can therefore test for the presence of reverse causality in our estimates of the "research opportunity effect" by including the estimate of the quality of research opportunities in drug-related medical research also as a regressor in the analyses of the determinants of other medical research. If the estimate of the research opportunity parameter β_K is close to zero when the dependent variable is a measure of other medical research, it is an indication that a positive estimate of the coefficient β_K when the dependent variable is a measure of drug-related medical research is not a result of reverse causality.

7.2 Identification of Induced Innovation Effects

7.2.1 Exogenous Variation in Disease Prevalence

As is well recognized in the literature on induced innovation, the causal effect of the potential market size on the extent of innovation cannot be inferred from the relationship between observed innovation and the observed market size due to the endogeneity of the observed market size. Acemoglu and Linn (2004) circumvent this problem by examining the relationship between changes in pharmaceutical innovation and changes in potential market size that are caused by population aging induced exogenous changes in disease prevalence. The key conditions to the success of this identification strategy are that the effect of aging on disease prevalence varies across diseases, the age demographics of the population have changed over time, and the changes in the age demographics are mostly caused by changes in fertility and are therefore mostly exogenous to the rate of pharmaceutical innovation.

We follow this general identification strategy in our analysis. However, we take into account both the effect that the change in age demographics has had on disease prevalence over time and the effect that the obesity epidemic has had on disease prevalence over time.[14]

[14] Acemoglu and Linn (2004) use also changes in the share of income of each age group. As the changes in income shares and the changes in population shares are relatively similar, it is not surprising that the inclusion of income movements does not alter their results.

The key conditions for using the obesity epidemic to identify the "induced innovation effect" are similar to the above mentioned conditions for the identification of the "induced innovation effect" from population aging induced changes in disease prevalence. First, the effect of obesity on disease prevalence must vary across diseases. Second, the body weight distribution in the population must have changed over time. And third, the obesity epidemic must be mostly exogenous to the rate of medical innovation. Our discussion of the descriptive statistics shows that the first two conditions hold (see Section 8.1). It is also reasonable to expect that the third condition holds. We are certainly not aware of any empirical research attributing the obesity epidemic–not to mention a non-negligible part of the obesity epidemic–to pharmaceutical and other medical innovation that would have made being obese or overweight more attractive choices compared to remaining normal weight.

7.2.2 Omission of Other Patient Population Characteristics

Of course, disease prevalence is not the only factor that influences the benefit from medical research, and aging and obesity are not the only factors that have influenced disease prevalence. Potentially important omitted factors include changes in disease severity, which exclusion from the analysis we have addressed in Section 3.1.2, and changes in insurance coverage. Also, as a population becomes wealthier its willingness to invest in developing treatments to diseases that affect mainly financially vulnerable populations may change. It is also possible that the attitudes toward these or other populations such as children change over time for other not yet understood reasons and that these changes influence the allocation of research. We do not dispute the existence of these other factors that potentially influence the extent medical research. Our focus on population aging and the obesity epidemic in the identification of the induced innovation effect is merely dictated by the availability of data on how these factors have influenced disease prevalence over time and the fact that both of these factors are known to have had a large impact on the prevalence of many diseases.

8 Results

To shorten the length of the paper we refer the reader to the earlier version Bhattacharya and Packalen (2008a) (hereafter BP (2008a)) for the figures.

8.1 Descriptive Statistics

Figure 1a in BP (2008a) shows the age and body weight distributions during the sample period. While the change has been gradual for both distributions, the change in the body weight distribution began more recently. Figure 1b in BP (2008a) shows the effect that the changes in the two distributions have had on the prevalence of each disease from the beginning of the sample period (1975) to the end of the sample period (2005). For both variables there is considerable variation in the effect (from -10% to +20%). These identifying variations are also not too correlated for the effects to be separately identified in most cases.

Figure 2a in BP (2008a) depicts the count of all publications (*All Publications*) and the count of publications with an abstract (*Publications with an Abstract*) by the year of publication. The graph also shows the count of publications that are indexed with a disease (*Publications Indexed with a Disease*) and the count of publications that are indexed with a disease that is matched to an ICD-9 disease by our match (*Publications Matched*). A publication may be indexed with more than one disease and, consequently, our match may match a publication to more than one ICD-9 disease. Therefore, the count of matches of publications to a disease (*Publication-Disease Matches*) is higher than the number of publications matched to at least one disease (*Publications Matched*). Figures 2b and 2c in BP (2008a) depict the count of matches to one of the 127 diseases for each of the three measures of drug-related medical research in each year and for each of the three measures of other medical research in each year.

The count of publications for each measure is an important determinant of the precision of our estimates because the variance of the share of publications that are matched to a disease

is expected to be inversely related to the count of publications that are matched to the disease and the estimated effects are identified from the effects on the share of publications that are related to each disease. This is also the reason why we report weighted regression results. While we do not report the unweighted regressions, the reported residual graphs serve the same purpose and also show that the results are not the product of outliers.

8.2 Estimates of the Quality of Research Opportunities

To construct the measure of the quality of research opportunities $\hat{K}_{it}$ using formula (9), we first estimate the parameters α_{if} and the parameters β_1 and β_2 using the procedure described in Section 4.2. The estimates of the parameters β_1 and β_2 are $\hat{\beta}_1 = 0.0628$ (s.e. 0.0045) and $\hat{\beta}_2 = 0.003$ (s.e. 0.0004). Figure 2d in BP (2008a) shows that the predicted probability that is calculated based on the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ as a function of the ingredient age $t-f$ tracks the mean of the observed probability closely except for when the ingredient age is 35 and over. The share of publications that use ingredients aged 35 and over is artificially inflated by the fact that the publications data consists mostly of publications published after 1950 and our methodology of assigning the year of discovery of each ingredient thus assigns the year of discovery between 1950 and 1965 for a disproportionate number of ingredients, as can be seen from Figure 2e in BP (2008a).

8.3 Induced Innovation and Research Opportunity Effects

8.3.1 All Medical Research

The count of all publications, denoted by N_{it}^{ALL}, that we use in this analysis corresponds to the measure *Publication-Disease Matches* in Figure 2a in BP (2008a).[15]

[15] The observations are weighted by the total count of publications matched to the disease during the sample period. That is, each observation is weighted by $\sum_{t=1975}^{2005} N_{it}^{ALL}$. The number of observations varies across columns because an observation is omitted if either $\hat{K}_{it} = 0$ or $N_{it}^{ALL} = 0$.

The results are shown in Table 1. Columns 1 and 2 show that aging-induced increases in disease prevalence have increased the medical research effort. In contrast, there is no evidence of a corresponding effect for obesity-induced changes in disease prevalence. Columns 3 and 4 show a positive relationship between the quality of research opportunities variable and the amount of total research. We re-iterate that to limit the scope of our analysis we have constructed the quality of research opportunities variable from research inputs that are relevant only for drug-related research (see Section 5.3.2). Accordingly, we postpone discussion of the magnitude of any research opportunity effect estimates until the analysis of the determinants of drug-related medical research.

Columns 3 and 4 also show that the inclusion of the quality of research opportunity variable renders the effect of aging-induced changes in disease prevalence statistically insignificant. However, as can be seen from Figure 3c in BP (2008a), which depicts the fixed effects specification analyzed in column 3, with the exception of the outlier disease 299 there is a robust positive relationship between aging-induced changes in disease prevalence and the changes in the overall research effort in the disease. Columns 5 and 6 show that when the disease 299 and the two other children's mental health diseases (314 and 315) are excluded, the relationship between aging-induced changes in disease prevalence and the overall research effort is again statistically significant.

Because the change in the age distribution has had such an unusual effect on the predicted disease prevalence for the disease 299 (see Figure 1.2) and because the dramatic increases in the number of diagnoses and research interest in the children's mental health diseases have been well recognized but without agreement over the causes of this, in the subsequent analyses we exclude the children's mental health diseases.[16]

[16]Research on children's mental health diseases has increased dramatically since the early 1990s and this increase is undoubtedly tied with the increase in the number of diagnoses for these diseases during the same period. While the unusual increase in the interest in these diseases is well known there is no agreement on why the increase has occurred. One explanation is that the increase in the diagnoses and the increase in research to the children's mental health diseases are consequences of the availability of dramatically better treatment options for these diseases, especially in the form of better knowledge of the effects of several

The magnitude of our estimates of the impact of aging-induced changes in disease prevalence on medical research in columns 5 and 6 is similar to the magnitude of the corresponding estimate in Acemoglu and Linn (2004) for aging-induced pharmaceutical innovation. These estimates in our study and in Acemoglu and Linn (2004) are larger than the estimates in Finkelstein (2004) for the impact of vaccine policies on the number of new clinical trials. As our estimates are obtained using a difference-in-difference methodology, the estimates reflect impacts on the composition of innovation rather than impacts on the total extent of innovation. This is aspect of the applied methodology is also noted by Acemoglu and Linn (2004) who emphasize that the response of innovation to changes in relative market sizes across diseases can be quite different from the response of innovation to changes in the total market size. For this reason we do not attempt to use our estimates to construct an estimate of the dynamic welfare impact of induced medical research.

8.3.2 Drug-Related Medical Research

The results for the three measures of drug-related medical research are shown in Table 2.[17] There is robust evidence across the three measures of drug-related medical research and the two fixed effects specifications for the hypothesis that the quality of research opportunities is a determinant of the allocation of drug-related medical research effort across diseases.[18] As expected, the estimates of this effect are now larger compared to the case when the dependent variable is constructed from all medical research (see Columns 3-6 in Table 1). However,

drugs such as methylphenidate (ritalin). Methylphenidate was discovered in the 1950s and our measure of technological opportunity is unable to predict the increase in research to these diseases because the increase happens 40 years after the discovery of the drug. An alternative explanation for why the disease 299 and to a lesser extent also the two other children's mental health diseases (314 and 315) are outliers is that during the sample period there may have been a general disproportional increase in research to diseases that primarily affect the children. We plan to explore this possibility in future research.

[17] For the measure *DRUG k*, where $k \in \{1, 2, 3\}$, each observation is weighted by $\sum_{t=1975}^{2005} N_{it}^{DRUG\ k}$.

[18] That this result is not a result of outliers can be seen from Figure 3d in BP (2008a). In this figure the horizontal axis is labeled as "Technological Opportunity Residual" as previously we referred to the quality of research opportunities in scientific research as "the quality of technological opportunity" to emphasize its similarity with the technological opportunity concept in research on private-sector technological innovation.

the estimates of the research opportunity effect remain considerably below the estimates of the impact of aging-induced changes in disease prevalence. This can reflect either the possibility that the true response is indeed relatively weak, as changes in the quality of research opportunities may be hard to identify at the time, or measurement error due to the fact that the research opportunity variable is constructed from a relatively limited set of research inputs. Nevertheless, our estimates of this effect demonstrate–for the first time– that also the direction of scientific research responds to changes in the quality of research opportunities.

There is also robust evidence for aging-induced changes in the composition of drug-related medical research across diseases. In contrast, there is no evidence for a positive relationship between obesity-induced changes in disease prevalence and the amount of drug-related medical research on the disease. If anything, the results suggest that there may be a negative relationship between obesity-induced changes in disease prevalence and the extent of drug-related medical research on a disease. A potential explanation for this result is that changes in population demographics may change the composition of research across different types of research within diseases (see Section 3.3.4). We examine this issue in Section 8.3.4.

8.3.3 Other Medical Research

The results for the three measures of other medical research are shown in Table 3.[19] For all specifications the estimate of the coefficient on the measure of the quality of research opportunities in drug-related research is much smaller than the estimates of the coefficient on the same variable are in the analyses of drug-related medical research. Moreover, the relationship is also statistically insignificant, except in column 1 in which the dependent variable is the most inclusive measure of other medical research and which is thus the most

[19] For the measure $OTHER\ k$, where $k \in \{1, 2, 3\}$, each observation is weighted by $\sum_{t=1975}^{2005} N_{it}^{OTHER\ k}$. There are fewer observations in columns 5 and 6 because the measure $N_{it}^{OTHER\ 3}$ is zero for some cells and therefore the dependent variable $\log(N_{it}^{OTHER\ 3})$ is not defined for those cells.

likely of the three measures of other medical research to include also some drug-related publications. The finding of no relationship between the level of other medical research and the measure of the quality of research opportunities in drug-related medical research is evidence against the potential concern that reverse causality is the reason for the observed positive relationship between the measure of the quality of research opportunities in drug-related research and the extent of drug-related medical research (see Section 7.1.2).

The results reported in columns 1-4 also provide evidence of aging-induced changes in the composition of other medical research across diseases but show no evidence of obesity-induced changes in the composition of other medical research across diseases. The results for surgery-related research (the measure *OTHER 3*) in columns 5 and 6 show that the relationship between aging-induced changes in disease prevalence and the extent of surgery-related research on the disease is positive but not statistically significant. The results in columns 5 and 6 also suggest a possible negative relationship between obesity-induced changes in disease prevalence and the extent of surgery-related research on the disease. A potential explanation is again that changes in population demographics may change the composition of research within diseases (see Section 3.3.4). We examine this possibility next.

8.3.4 Effects on the Composition of Medical Research Within Diseases

Changes in disease prevalence may have effects also on the composition of research across different types of research within a disease, and such changes in the composition of research within diseases may influence the estimates of the determinants of the extent of research effort across diseases (see Section 3.3.4). Accordingly, in Table 4 we report estimates of the determinants of the composition of research within diseases.

In the analyses reported in columns 1 and 2 the dependent variable is the logarithm of the ratio of the most restrictive measure of drug-related research and all research.[20] As expected,

[20] Each observation is weighted by $\sum_{t=1975}^{2005} N_{it}^{DRUG\ 3}$.

the results indicate a positive relationship between the quality of research opportunities in drug-related medical research on a disease and the share of medical research on the disease that is drug-related. The results also indicate that aging has not shifted research away from drug-related medical research. The positive but statistically insignificant point estimate leaves open the possibility that drug-related medical research reacts to aging-induced changes in disease prevalence more strongly than all medical research. The results also indicate that obesity-induced changes in disease prevalence have a negative but statistically insignificant relationship with changes in the share of all research that is drug-related research. This suggests that an obesity-induced increase in the prevalence of a disease may decrease the share of research on the disease that is drug-related and increase the share of research on the disease that examines the physiology of the disease in the obese.

In the analyses reported in columns 3 and 4 the dependent variable is the logarithm of the ratio of surgery-related research and all research.[21] The results show a negative and statistically significant relationship between the share of surgery-related research on a disease and the measure of the quality of research opportunities in drug-related medical research on the disease. This is additional evidence against the aforementioned reverse causality explanation for the research opportunity effect estimate in drug-related medical research. This is also further evidence that an increase in the quality of research opportunities in drug-related research shifts research effort away from other types research to drug-related research. We find no relationship between aging-induced changes in disease prevalence and the ratio of research that is surgery-related. The negative relationship between obesity-induced changes in disease prevalence and the share of research on the disease that is surgery-related again suggests the possibility that an obesity-induced increase in the prevalence of a disease shifts resources away from general research to obesity-specific research on the disease.

[21] Each observation is weighted by $\sum_{t=1975}^{2005} N_{it}^{OTHER\ 3}$.

9 Conclusion

Our empirical results show that the composition of medical research across diseases responds to changes in the quality of research opportunities and population aging induced changes in disease prevalence. The results also suggest that an obesity epidemic induced increase in the prevalence of a disease may have shifted research away from more general drug-related medical research and from more general surgery-related research on the disease and likely toward obesity-specific research on the disease.

These results provide support for the hypothesis that, similar to private-sector technological innovation which has been the focus of existing research, also the direction of scientific research responds to changes in the quality of research opportunities and the benefit from research. The distinction between scientific research and private-sector technological innovation is important because the two can differ in many ways, including for-profit vs. non-profit status and the level of authority that individual researchers have.

While we do not examine the mechanisms that induce the direction of scientific research to respond to these factors, our analysis shows that these–yet largely unexplored–mechanisms have a virtuous property in the sense that they induce scientific research to respond to factors that in part determine the socially optimal allocation of research resources. Our analysis is an important input into analyses of these mechanisms as it refutes the view of scientific research as an ivory tower in which scientists' desire to influence other scientists is the only determinant of the direction of research. We expect that future research on these mechanisms, on how far or close the allocation of scientific research is from the socially optimal allocation, and on what factors besides opportunities and societal benefits are important drivers of the direction of scientific research, will be both fertile and worthy. We hope that the value of such analyses is further enhanced by our methodological innovations, which include a match between publications and disease prevalence data and new approaches to estimating the quality of opportunities and measuring the associated research inputs.

References

Acemoglu, D. and J. Linn, 2004, "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics*, 119, pp. 1049-90.

Adams, J. D. and J. R. Clemons, 2006, "The NBER-Rensselaer Polytechnic Institute Scientific Papers Database: Characteristics and Purpose," Mimeo.

Adams, J. D. and J. R. Clemons, 2008, "The Origins of Industrial Scientific Discoveries," NBER working paper No. 13823.

Aghion, P., Dewatripont, M. and J. C. Stein, 2008, "Academic Freedom, Private-Sector Focus, and the Process of Innovation," *RAND Journal of Economics*, 39, pp. 617-35.

Azoulay, P., Ding, W. and T. Stuart, 2007a, "The Determinants of Faculty Patenting Behavior: Demographics or Opportunities?" *Journal of Economic Behavior & Organization*, 63, pp. 599-623.

Azoulay, P., Ding, W. and T. Stuart, 2009, "The Effect of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output," *Journal of Industrial Economics*, 57, pp. 637-76.

Azoulay, P., Michigan, R. and B. N. Sampat, 2007b, "The Anatomy of Medical School Patenting," *The New England Journal of Medicine*, 357, pp. 2049-56.

Bhattacharya, J. and M. Packalen, 2008a, "Is Medicine an Ivory Tower? Induced Innovation, Technological Opportunity, and For-Profit vs. Non-Profit Innovation," NBER working paper No. 13862.

Bhattacharya, J. and M. Packalen, 2008b, "The Other Ex-Ante Moral Hazard in Health," NBER working paper No. 13863.

Caballero, R. J. and A. B. Jaffe, 1993, "How High are The Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth," in Blanchard O. J. and S. Fisher, eds., *NBER Macroeconomics Annual 1993*. Cambridge: MIT Press, pp. 15-74.

Cameron, A. C., Gelbach, J. B. and D. L. Miller, 2007, "Bootstrap-Based Improvements for Inference with Clustered Errors," NBER technical working paper No. 344.

Cockburn, I. M. and R. M. Henderson, 1998, "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery," *The Journal of Industrial Economics*, 46, pp. 157-82.

Dasgupta, P. and P. A. David, 1994, "Towards a new Economics of Science," *Research Policy,* 23, pp. 487-521.

DellaVigna, S., and J. M. Pollet, 2007, "Demographics and Industry Returns," *American Economic Review*, 97, pp. 1667-702.

Finkelstein, A., 2004, "Static and Dynamic Effects of Health Policy: Evidence from the Vaccine Industry," *Quarterly Journal of Economics*, 119, pp. 527-64.

George, L. and J. Waldfogel, 2003, "Who Affects Whom in Daily Newspaper Markets?" *Journal of Political Economy*, 111, pp. 765-84.

Glaeser, E. L., ed., 2003, *The Governance of Not-for-Profit Organizations*. Chicago: The University of Chicago Press

Jaffe A. B. and M. Trajtenberg, 1996, "Flows of Knowledge from Universities and Federal Labs: Modeling the Flow of Patent Citations Over Time and Across Institutional and Geographic Boundaries," *Proceedings of the National Academy of Sciences*, 93, pp. 12671-7.

Hicks, J. R., 1932, *Theory of Wages*. London: Macmillan.

Lakdawalla, D. and T. Philipson, 2006, "The Nonprofit Sector and Industry Performance," *Journal of Public Economics*, 90, pp. 1681-98.

Lichtenberg, F. R., 1999, "The Allocation of Publicly Funded Biomedical Research," in *Medical Care Output and Productivity: Studies in Income and Wealth*, LXIII, Berndt, E. and D. Cutler, eds., Chicago: University of Chicago Press.

Lichtenberg, F. R., 2006, "Importation and Innovation," NBER working paper No. 12539.

Lichtenberg, F. R. and J. Waldfogel, 2003, "Does Misery Love Company? Evidence from Pharmaceutical Markets Before and After the Orphan Drug Act," NBER working paper No. 9750.

Mane, K. K. and K. Börner, 2004, "Mapping topics and topic bursts in PNAS," *Proceedings of the National Academy of Sciences*, 101, pp. 5287-90.

Merton, R. K. 1973 [1942], "The Normative Structure of Science," in *The Sociology of Science: Theoretical and Empirical Investigations*, R. K. Merton, ed., Chicago: The University of Chicago Press.

Murray, F. and S. Stern, 2007, "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? An Empirical Test of the Anti-Commons Hypothesis," *Journal of Economic Behavior and Organizations*, 63, pp. 648-87.

Newell, R. A., Jaffee, A. and R. Stavins, 1999, "The Induced Innovation Hypothesis and Energy-Saving Technological Change," *Quarterly Journal of Economics*, 114, pp. 907-40.

Popp, D., 2002, "Induced Innovation and Energy Prices," *American Economics Review*, 92, pp. 160-80.

Rosenberg, N., 1982, *Inside the Black Box: Technology and Economics*. Cambridge: Cambridge University Press.

Saha, S. B, and B. Weinberg, 2008, "The Economics of Ivory Towers," Mimeo.

Scherer, F. M., 1965, "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions," *American Economic Review*, 55, pp. 1097-125.

Schmookler, J., 1966, *Invention and Economic Growth*. Cambridge: Harvard University Press.

Stern, S., 2004, "Do Scientists Pay to Be Scientists?" *Management Science*, 50, pp. 835-53.

Waldfogel, J., 2003, "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated Product Markets," *RAND Journal of Economics*, 34, pp. 557-68.

Ward, M. R. and D. Dranove, 1995, "The Vertical Chain of Research and Development in the Pharmaceutical Industry," *Economic Inquiry*, 33, pp. 70-87.

Yin, W., 2008, "Market Incentives and Pharmaceutical Innovation," *Journal of Health Economics*, 27, pp. 1060-1077.

Table 1. Determinants of the Allocation of the All Medical Research Across Diseases.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	$\ln(N_{it}^{ALL})$	$\ln(N_{it}^{ALL})$	$\ln(N_{it}^{ALL})$	$\ln(N_{it}^{ALL})$	$\ln(N_{it}^{ALL})$	$\ln(N_{it}^{ALL})$
$\ln\left(\hat{K}_{it}\right)$			0.32 [0.16] $p_{wild} = 0.027$	0.22 [0.15] $p_{wild} = 0.137$	0.29 [0.14] $p_{wild} = 0.014$	0.22 [0.15] $p_{wild} = 0.149$
$\ln\left(M_{it}^{AGING}\right)$	2.74 [1.43] $p_{wild} = 0.037$	2.40 [1.29] $p_{wild} = 0.083$	1.76 [1.16] $p_{wild} = 0.125$	1.84 [1.18] $p_{wild} = 0.168$	2.66 [1.47] $p_{wild} = 0.062$	2.76 [1.28] $p_{wild} = 0.023$
$\ln\left(M_{it}^{OBESITY}\right)$	0.43 [1.18] $p_{wild} = 0.773$	−0.004 [0.71] $p_{wild} = 0.996$	−0.25 [0.98] $p_{wild} = 0.800$	−0.13 [0.68] $p_{wild} = 0.858$	−0.07 [1.05] $p_{wild} = 0.947$	−0.13 [0.68] $p_{wild} = 0.857$
Fixed effects	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year
Number of observations	3884	3884	3883	3883	3796	3796

Our statistical inference is based on p_{wild} which is calculated using the cluster-robust standard error (clustered at the class level) and the wild cluster bootstrapped distribution of the t-statistic (1000 iterations). Monte Carlo evidence favors this approach when the number of clusters is small and the clusters are unbalanced (Cameron et al., 2007). The wild cluster bootstrapped standard error (1000 iterations) is presented in brackets. In columns 5 and 6 children's mental health diseases (299, 314, 315) are omitted.

Table 2. Determinants of the Allocation of Drug-Related Medical Research Across Diseases.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	$\ln(N_{it}^{DRUG\,1})$	$\ln(N_{it}^{DRUG\,1})$	$\ln(N_{it}^{DRUG\,2})$	$\ln(N_{it}^{DRUG\,2})$	$\ln(N_{it}^{DRUG\,3})$	$\ln(N_{it}^{DRUG\,3})$
$\ln\left(\hat{K}_{it}\right)$	0.64 [0.36] $p_{wild} = 0.021$	0.58 [0.31] $p_{wild} = 0.037$	0.59 [0.30] $p_{wild} = 0.022$	0.48 [0.28] $p_{wild} = 0.057$	0.85 [0.35] $p_{wild} = 0.006$	0.74 [0.31] $p_{wild} = 0.010$
$\ln\left(M_{it}^{AGING}\right)$	2.51 [1.81] $p_{wild} = 0.200$	2.32 [1.09] $p_{wild} = 0.008$	2.46 [1.17] $p_{wild} = 0.185$	2.73 [1.22] $p_{wild} = 0.015$	3.85 [1.92] $p_{wild} = 0.030$	4.06 [1.93] $p_{wild} = 0.019$
$\ln\left(M_{it}^{OBESITY}\right)$	−1.75 [2.02] $p_{wild} = 0.525$	−1.79 [1.29] $p_{wild} = 0.163$	−1.77 [1.91] $p_{wild} = 0.489$	−1.57 [1.15] $p_{wild} = 0.223$	−2.08 [2.11] $p_{wild} = 0.451$	−1.87 [1.49] $p_{wild} = 0.208$
Fixed effects	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year
Number of observations	3730	3730	3730	3730	3697	3697

Children's mental health diseases (299, 314, 315) are omitted. See the footnote to Table 1 for an explanation of the standard errors and p-values.

Table 3. Determinants of the Allocation of Other Medical Research Across Diseases.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	$\ln(N_{it}^{OTHER\,1})$	$\ln(N_{it}^{OTHER\,1})$	$\ln(N_{it}^{OTHER\,2})$	$\ln(N_{it}^{OTHER\,2})$	$\ln(N_{it}^{OTHER\,3})$	$\ln(N_{it}^{OTHER\,3})$
$\ln\left(\hat{K}_{it}\right)$	0.20 [0.12] $p_{wild} = 0.086$	0.15 [0.13] $p_{wild} = 0.339$	0.06 [0.09] $p_{wild} = 0.566$	−0.007 [0.09] $p_{wild} = 0.950$	−0.11 [0.18] $p_{wild} = 0.507$	−0.06 [0.16] $p_{wild} = 0.726$
$\ln\left(M_{it}^{AGING}\right)$	2.84 [1.53] $p_{wild} = 0.056$	2.82 [1.35] $p_{wild} = 0.018$	2.98 [1.57] $p_{wild} = 0.062$	2.79 [1.36] $p_{wild} = 0.025$	1.64 [1.24] $p_{wild} = 0.181$	2.59 [1.55] $p_{wild} = 0.114$
$\ln\left(M_{it}^{OBESITY}\right)$	0.23 [0.92] $p_{wild} = 0.798$	−0.11 [0.52] $p_{wild} = 0.830$	0.13 [0.89] $p_{wild} = 0.897$	−0.19 [0.69] $p_{wild} = 0.787$	−0.51 [1.20] $p_{wild} = 0.684$	−1.49 [0.81] $p_{wild} = 0.066$
Fixed effects	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year
Number of observations	3796	3796	3796	3796	3723	3723

Children's mental health diseases (299, 314, 315) are omitted. See the footnote to Table 1 for an explanation of the standard errors and p-values.

Table 4. Determinants of the Allocation of Medical Research Across Research Types Within Diseases.

	(1)	(2)	(3)	(4)
Dependent variable:	$\ln(\frac{N_{it}^{DRUG\ 3}}{N_{it}^{ALL}})$	$\ln(\frac{N_{it}^{DRUG\ 3}}{N_{it}^{ALL}})$	$\ln(\frac{N_{it}^{OTHER\ 3}}{N_{it}^{ALL}})$	$\ln(\frac{N_{it}^{OTHER\ 3}}{N_{it}^{ALL}})$
$\ln\left(\hat{K}_{it}\right)$	0.53 [0.21] $p_{wild} = 0.005$	0.54 [0.21] $p_{wild} = 0.001$	−0.25 [0.14] $p_{wild} = 0.043$	−0.20 [0.12] $p_{wild} = 0.089$
$\ln\left(M_{it}^{AGING}\right)$	0.79 [1.06] $p_{wild} = 0.494$	1.58 [1.05] $p_{wild} = 0.196$	−0.62 [1.04] $p_{wild} = 0.563$	−0.77 [0.92] $p_{wild} = 0.459$
$\ln\left(M_{it}^{OBESITY}\right)$	−1.69 [1.28] $p_{wild} = 0.242$	−1.76 [1.07] $p_{wild} = 0.105$	−0.65 [0.46] $p_{wild} = 0.248$	−0.89 [0.48] $p_{wild} = 0.073$
Fixed effects	Disease, Class×Year	Disease, Year	Disease, Class×Year	Disease, Year
Number of observations	3697	3697	3723	3723

Children's mental health diseases (299, 314, 315) are omitted. See the footnote to Table 1 for an explanation of the standard errors and p-values.

Data Appendix

Data Appendix to Section 5.1

We use several strategies to identify and measure drug-related medical research. The first is to classify all publications that are matched to an ingredient as being drug-related medical research and count a publication that is matched to n different cohorts of ingredients as n units of research. The second is to classify all publications that are matched to an ingredient as being drug-related medical research and count each such publication as one unit of research. The third is to classify all publications that have a MESH term indexed together with the "major topic" flag and the MESH qualifier term "drug therapy", "drug effects" or "pharmacology" as being drug-related research and count each such publication as one unit of research. We refer to these three constructed measures of drug-related medical research as *DRUG 1*, *DRUG 2*, and *DRUG 3*, respectively.

We also use several strategies to identify and measure other medical research. The first is to classify all publications that are not matched to an ingredient as being other medical research and count each such publication as one unit of research. The second is to classify all publications that are 1) not matched to an ingredient, 2) not indexed are indexed with any of the MESH qualifier terms "drug therapy", "drug effects" or "pharmacology", and 3) not indexed with the MESH term "Chemicals and Drugs", as being other medical research and count each such publication as one unit of research. This second method should therefore exclude also most of the research that is conducted using unapproved drugs that do not appear in our list of FDA approved ingredients. The third is to classify all research that is indexed with the MESH qualifier term "surgery" or "transplantation" as being other medical research and count each such publication as one unit of research. We call these three measures as *OTHER 1*, *OTHER 2*, and *OTHER 3*, respectively.

Data Appendix to Section 5.2

We limit the match effort to diseases for which the MEPS disease incidence data (see Section 5.4) includes at least 100 observations.[22] We do not match ICD-9 codes that include either the word "Other" or the word "Unspecified" in the title because these ICD-9 codes typically include a variety of different diseases and are therefore difficult to match to the MESH vocabulary. Neither do we match diseases in the pregnancy category (class 11), in the

[22] We exclude HIV/AIDS because the disease does not appear in the publications database until the early 1980s and because the variations in the incidence of HIV/AIDS are obviously not mainly driven by aging or the obesity epidemic.

congenital category (class 14), in the perinatal category (class 15), in the symptoms category (class 16), in the injuries category (class 17) or in the services category (class V). These classes are excluded from the match effort both in order to limit the scope of our match effort and because of the difficulty of matching diseases in these categories. If a match from an individual disease to a MESH entry/entries is not possible we try to match a group of ICD-9 codes to a MESH entry/entries. The 127 matched diseases account for 377,482 of the 745,355 disease mentions in the MEPS disease incidence data.

Because MESH is a hierarchical vocabulary, we also count all research that is indexed to any subnode of a matched MESH term as research that is related to the matched disease or group of diseases.[23] As the MESH vocabulary has changed over the years we make an effort to check that the MESH terms for the matched diseases have not changed in a way that would influence the research effort estimate. For the diseases for which the related publications from a year during the sample period are likely to have been indexed by terms other than the matched MESH entry/entries we exclude the observations from such years and from any of the preceding years. In Bhattacharya and Packalen (2008a) the match for such diseases is marked with an asterisk and the year prior to which any observations are excluded.

Data Appendix to Section 5.4

To estimate disease incidence for each age and BMI group we use the Medical Expenditure Panel Survey (MEPS) data from years 1996-2005.[24] Each subject is followed in MEPS for two years. For each subject we aggregate the observations in each year into one observation. MEPS includes a list of self-reported diseases that are coded by the International Classification of Diseases, Ninth Revision (ICD-9). MEPS does not include BMI information for years 1996-2000. We therefore use the National Health Interview Survey (NHIS) data from years 1996-2000 and the match between NHIS and MEPS to obtain BMI information for the observations in those years. Except for subjects in the age group 0-18 we exclude subjects without either age or BMI information.[25] The resulting MEPS data includes 262,958

[23] We manually remove several matches of ICD-9 diseases to terms for neoplasms in MESH when the same neoplasm term is also mapped to a disease in the ICD-9 disease class 2 (neoplasms). MESH has 4982 disease terms. The match maps 1338 terms in MESH to the 127 diseases. 51 of the matched terms are mapped to 2 diseases and one term in MESH is mapped to 3 diseases. All other terms are mapped to only 1 disease.

[24] Because the trends in the changes in the age and body weight distributions have been similar across the developed nations we do not believe that using data on disease incidence, age demographics and obesity for the United States but data on world-wide publications is a significant concern.

[25] Interpreting BMI of children is not as straightforward as interpreting BMI of adults. Hence, we do not distinguish the disease incidence by body weight for the age group 0-18. Consequently, we set $s^{BMI}_{1,1,t} = 1$,

observations on 149,737 subjects.

We use the Surveillance Epidemiology and End Results (SEER) data from years 1975-2004 to estimate the share of people in each age group in each year.[26] For each age group we use the NHIS data from years 1976-2005 to estimate the share of people in each BMI group in each year.[27]

In estimating the disease incidence parameters $\mu_{i,j,k}$ (see Section 6) we allow these parameters to vary by sex, race (black/non-black), insurance status (private/not private) and year but for expositional simplicity we omit these issues in the main text. As we don't measure changes in insurance coverage across time we do not examine the effect that changes in the insurance coverage across time may have on the benefit from medical research and on the extent of research.

The decomposition of changes in disease incidence to population aging and obesity epidemic induced changes (see Section 6) arises as follows. Let M_{it_0} denote the incidence of disease i in the initial year t_0. Let R_{it}^{AGING} denote the effect of aging alone on the incidence of disease i so that if only population aging affected the incidence of disease i the incidence of disease i would be $M_{it_0}R_{it}^{AGING}$ in year t. Let $\tilde{R}_{it}^{OBESITY}$ denote the additional effect of the obesity epidemic on the incidence of disease i so that if only aging and obesity affected the incidence of disease i the incidence of disease i would be $M_{it} = M_{it_0}R_{it}^{AGING}\tilde{R}_{it}^{OBESITY}$ in year t. Let $R_{it}^{OBESITY}$ denote the effect of obesity alone on the incidence of disease i so that if only obesity affected the incidence of disease i the incidence of disease i would be $M_{it_0}R_{it}^{OBESITY}$ in year t. Because R_{it}^{AGING} is small, $R_{it}^{OBESITY} \approx \tilde{R}_{it}^{OBESITY}$. Therefore, $\ln\left(M_{it_0}R_{it}^{AGING}\tilde{R}_{it}^{OBESITY}\right) \approx \ln\left(M_{it_0}R_{it}^{AGING}R_{it}^{OBESITY}\right)$. We can therefore decompose the total effect $\ln\left(M_{it_0}R_{it}^{AGING}\tilde{R}_{it}^{OBESITY}\right)$ into an aging effect $\ln\left(R_{it}^{AGING}\right)$ and an obesity effect $\ln\left(R_{it}^{OBESITY}\right)$. Because the empirical specifications include either disease fixed effects, we can use the variables $\ln\left(M_{it_0}R_{it}^{AGING}\right)$ and $\ln\left(M_{it_0}R_{it}^{OBESITY}\right)$–instead of the variables $\ln\left(R_{it}^{AGING}\right)$ and $\ln\left(R_{it}^{OBESITY}\right)$–as regressors. In the text these variables $\ln\left(M_{it_0}R_{it}^{AGING}\right)$ and $\ln\left(M_{it_0}R_{it}^{OBESITY}\right)$ are denoted by $\ln\left(M_{it}^{AGING}\right)$ and $\ln\left(M_{it}^{OBESITY}\right)$, respectively.

This decomposition reflects the fact that the effect that an obesity-induced change in disease incidence has had on the extent of research may be different than the effect that a

$s_{1,2,t}^{BMI} = 0$ and $s_{1,3,t}^{BMI} = 0$ for all t. Because people the age group 0-18 have small average expenditures and also the effect of the obesity epidemic on disease incidence is small for this age group, ignoring the effect of the obesity epidemic on the disease incidence of this age group has a negligible influence on the potential market size variable M_{it}^{TOTAL}.

[26] We impute the values for 2005 by assuming that the change in the population in each age group from 2004 to 2005 was the same as it was from 2003 to 2004.

[27] We impute the values for 1975 by assuming the the body weight distribution was the same in 1975 as it was in 1976.

corresponding aging-induced change in disease incidence has had on the extent of research. These two effects would be different, for example, if the implications of aging on disease incidence have been better understood than the implications of obesity on disease incidence, or if the change in age demographics was more expected than the obesity epidemic.

Data Appendix to Section 5.5

As our discussion of the descriptive statistics in Section 8.1 shows, there is a discontinuous jump in the share of publications with abstracts in the database from 1974 to 1975. Moreover, a number of diseases are indexed with different MESH terms before 1975 and especially before 1970 than they are after 1975. For these reasons we choose 1975-2005 as our sample period.

When we determine the cohort of an ingredient (the year before the first mention of the ingredient–see Section 5.3.3) from the publications in years 1906-2005. In estimating the parameters that govern the quality of research opportunities (see Section 4) we limit the range of cohorts f to years 1960-2001 because there is a discontinuous jump in 1950 in the number of publications that are indexed in MEDLINE and because there is a discontinuous fall in the number of ingredients in a cohort from 2001 to 2002 due to the lag between the year in which an ingredient is first mentioned in the publications database and the year of FDA approval of the ingredient.[28] Because of this lag, because many of the diseases are indexed with different terms before 1970, and because in the subsequent analysis our focus is on the sample period 1975-2005, in estimating the quality of research opportunities (see Section 4) we limit the range of the years t to 1970-2002.

[28] We multiply the initially estimated research opportunity by a factor that compensates for truncation. We assume that the average baseline productivity is the same before and after any truncation point. That is, the estimates are multiplied by $\{\sum_{t-f=1}^{\infty} e^{-\hat{\beta}_1(t-f)} \times [1-e^{-\hat{\beta}_2(t-f)}]\}/\{\sum_{t-f=1}^{t-1960} e^{-\hat{\beta}_1(t-f)} \times [1-e^{-\hat{\beta}_2(t-f)}]\}$ for all years $t \leq 2001$. For $t > 2001$ we also compensate for truncation due to the upper bound.

Background Appendix: Non-Profit Nature of Publications in Medicine

The intertwined nature of industrial R&D activity and academic research activity is well established. The two are connected in many ways, especially in the biomedical sector. In this appendix we discuss each of these connections. And in each case we argue that despite the connection, new drug introductions reflect largely the functioning of for-profit incentives and biomedical publications reflect largely the functioning of non-profit incentives.

1. Pharmaceutical Innovation Reflects Mostly For-Profit Incentives

First, many of the innovations that are introduced by pharmaceutical companies are based on knowledge generated in the public sector (see e.g. Cockburn and Henderson, 1998 and Ward and Dranove, 1995). This fact is at odds with the interpretation in the related literature on the determinants of pharmaceutical innovation (e.g. Acemoglu and Linn, 2004), where research in the pharmaceutical industry is often taken as an example of innovation in the for-profit sector. However, we believe that, to a first approximation, it is reasonable to consider the majority of pharmaceutical innovation as reflective of the functioning of for-profit incentives. The alternative–that pharmaceutical firms do not make substantial choices about which lines of research to pursue but decide their research agenda mainly on the basis of prior basic research done in the public sector–seems to us less reasonable.

2. Research Publications Are Mostly from Academic Research Institutions

Second, individuals employed in pharmaceutical companies also publish in academic journals and co-author research papers with university researchers (see e.g. Cockburn and Henderson, 1998 and Adams and Clemons, 2008). Biomedical publishing will therefore reflect, in part, how for-profit incentives respond to determinants of innovation. Unfortunately no comprehensive study exists on what part of biomedical publishing can be attributed to industry. However, Adams and Clemons (2006) present summary statistics on the origin of scientific publications in a database of over 5000 journals across the sciences during the time period 1980-1999. Their analysis shows that during this time period the top 110 U.S. universities published 800,000 papers in medicine and the top 200 U.S. R&D firms published less than 30,000 papers in medicine. While this comparison is not comprehensive because it does not compare the biomedical publications of all U.S. universities with the biomedical publications of all pharmaceutical and biotechnology firms, the comparison suggests that the contribution

of industry to academic publishing during this time period is substantially less voluminous than the contribution of research universities to academic publishing.

This conclusion is also supported by a National Science Foundation study which mentions that in 2003 in the science and engineering sector the academic sector accounted for almost three quarters of the publications originating in the U.S.[29] The remaining one quarter of the publications is attributed to industry, government and non-profits. The study also finds that only 6.0% of the publications that have at least one academic author have an industry co-author. The available evidence thus suggests that the results in our analysis and the results in any other analysis that examines the determinants of publications in medicine in a comprehensive manner mainly reflect the publishing behavior of academic research institutions and the associated non-profit incentives as opposed to the publishing behavior of industry and the associated for-profit incentives.

3. Academic Researchers in Medicine Mostly Do Not Patent

Third, in addition to publishing their work in academic publications university researchers also apply for patents. If patents and the associated for-profit incentives are a significant driving force behind academic bio-medical research then biomedical publishing as the other product of biomedical research would also reflect the functioning of for-profit incentives. However, the analysis of patenting in medicine by Azoulay et al. (2007b) shows that during the period from 1981 to 2000 only 5% of faculty members in medical schools applied for a patent that was successfully granted.[30] Especially when this already low percentage figure is combined with the fact that most patented innovations bring no revenue to the patentee, we conclude that patenting and the associated for-profit incentives are likely not a significant determinant of biomedical research and publishing. Moreover, this conclusion is even stronger for our analysis as we only consider biomedical publications that are applied biomedical research in the sense that the publication is related to a specific disease (see Sections 5.1-5.2) and Azoulay et al. (2007b) find that that biomedical patenting is much more common for basic research than for applied research.

Because patenting is not very common in medicine and particularly in the applied research that is our focus, patenting by other researchers is unlikely to influence the direction of research much in medicine. This is somewhat in contrast with the analysis of the anti-commons hypothesis in biotechnology by Murray and Stern (2007) who find a modest anti-

[29] National Science Foundation, Division of Science Resources Statistics (NSF/SRS) 2006, "Industrial Funding of Academic R&D Continues to Decline in FY 2004," NSF 06-315.

[30] The analysis also shows that academic biomedical patents accounted for only 25% of total biomedical patenting even during the peak period of academic biomedical patenting (the late 1990s).

commons effect. However, their analysis was based on selecting the biomedical publication (Nature Biotechnology) in which, *ex ante*, patenting by other researchers was the likeliest to have an effect on publication behavior.

4. Academic R&D is Mostly Funded by Non-Industry Sources

Fourth, some of the research activities of academic institutions are financed by industry. If industry funding is a major source of funding for academic R&D, the direction of academic R&D might simply reflect the direction of industry R&D and the associated for-profit incentives. However, a National Science Foundation study shows that in science and engineering during 1993-2004 academic R&D funds provided by industry have been less than 8% of all R&D funding.[31] The industry funding of academic R&D was $2.1 billion in 2004. This figure is substantially less than total university research expenditures ($43.0 billion in 2004) and federal support for university research expenditures ($27.4 billion in 2004). That figure is also substantially less than university research expenditures in medical sciences alone ($14 billion in 2004) and federal support for medical sciences ($9.4 billion in 2004). Considering the balance of the evidence, we conclude that research in academic medicine largely (though not exclusively) exemplifies the products of non-profit incentives.

[31] National Science Foundation, Division of Science Resources Statistics (NSF/SRS) 2006, "Where Has All the Money Gone? Declining Industrial Support of Academic R&D," NSF 06-328.