Bottazzi, Giulio; Plotnikova, Tatiana

Working Paper

Productivity and heterogeneous knowledge: Exploring the relationship in a sample of drug developers

Jena Economic Research Papers, No. 2010,044

Provided in Cooperation with:
Max Planck Institute of Economics

JENA ECONOMIC RESEARCH PAPERS

2010 – 044

Productivity and Heterogeneous Knowledge: Exploring the Relationship in a Sample of Drug Developers

by

Giulio Bottazzi
Tatiana Plotnikova

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Productivity and Heterogeneous Knowledge: Exploring the Relationship in a Sample of Drug Developers*

Giulio Bottazzi† Tatiana Plotnikova‡

May 2010

Abstract

This paper aims to investigate the effect of knowledge characteristics on the total factor productivity of firms developing drugs in the pharmaceutical industry. We decompose knowledge into knowledge associated with the technological firm portfolio and knowledge related to R&D projects, which represent drug development at the clinical testing stage. The latter is attributed to the knowledge of relevant markets where the drugs will be sold. The results show that the effect of technological coherence vs. market coherence and of accumulated knowledge on the productivity of firms differs. Productivity increases with the number of patents and decreases with the patent diversity and project portfolio coherence. When considering only the project knowledge, the diversity of the project portfolio positively affects productivity.

Keywords: total factor productivity, diversity, coherence, knowledge
JEL Classification: D24, O32, L25, L65

*We thank German Science Foundation (DFG) and DIME network for financial support. Moreover, we want to express our gratitude to Sant'Anna School for Advanced Studies for hosting one of the authors. We are grateful to the participants of the JER Workshop in December 2009, especially to Arianna Martinelli, Vera Popova and Sebastian von Engelhardt, and the participants of DIME conference "Organizing for networked innovation" in April 2010, particulary Diego D'Adda, Bart Leten and Ferran Vendrell for useful comments, expressed interest and concerns. We appreciate greatly Chad Baum's help with English editing. The usual caveats apply.

†Sant'Anna School for Advanced Studies, piazza Martiri della Liberta, 33, 56127 Pisa, Italy; e-mail: bottazzi@sssup.it

‡DFG Research Training Program "The Economics of Innovative Change", Friedrich-Schiller-University Jena, Carl-Zeiss-Strasse 3, D-07743 Jena , Germany; e-mail: tatiana.plotnikova@uni-jena.com

1 Introduction

Extensive empirical studies have focused on the significant heterogeneity among firms in terms of productivity. Furthermore, as demonstrated by the survey of the literature in Bartelsman and Doms (2000), productivity differences can be very large, while other important finding of these studies is that the differences in the productivity of firms can persist over time. One explanation for this persistent heterogeneity could be that firms have different knowledge sets, which determine their relative abilities to efficiently produce output from their given inputs. Specifically, knowledge in the form of R&D (Lichtenberg and Siegel, 1991; Hall and Mairesse, 1995), as well as advanced technology application (McGuckin et al., 1998; Chennells and Van Reenen, 1998) or computerization (Brynjolfsson and Hitt, 1995, 2003), has been found to explain the heterogeneous productivity of various firms.

Most studies have connected productivity with the amount of knowledge. In doing so, it is usually assumed that knowledge enters into the productivity function as an additional factor (e.g. Grilliches (1979); Hall and Mairesse (1995)). Furthermore, there is a broad research, which arguing that the relation between the amount of knowledge, usually approximated by accumulated R&D, and productivity or output growth is positive (see Nadiri (1994) for a survey).

What the majority of research on the relation between productivity and knowledge does not consider, however, is that knowledge might be heterogeneous. The knowledge of one firm could be considered as being comprised of different components, such as varying technologies and experiences across different areas. Accordingly, one of the necessary functions of a firm would be knowledge integration (Grant, 1996), in order to achieve efficient production. In other words, depending on a firm's success in knowledge integration, the firm could experience either an increase or decrease in productivity. The successful integration of new knowledge into an already existing pool thus depends on the cognitive distance between various types of knowledge (Nooteboom, 2000). Therefore, we argue that characteristics of knowledge portfolio, such as diversity and coherence, affect productivity by reflecting synergies among pieces of heterogeneous knowledge.

Within the relevant literature, some evidence of firm behavior aimed at the exploitation of knowledge synergies has been highlighted. For example, Scott and Pascoe (1987); Montgomery and Hariharan (1991); Teece et al. (1994), among others, find that the distribution of a firm's activities is not random. In fact, firms tend to distribute their activities into areas where they can apply their knowledge (Nerkar and Roberts, 2004; Cantner and Plotnikova, 2009). This intuition has proved to hold for technological relations (Breschi et al., 2003) as well as in relation

to the firm's previous experience in technology and product markets (Nerkar and Roberts, 2004). The existence of these synergies suggests that there might be a potential spillover effect between different types and groups of knowledge which would make any firm which expands its knowledge into related areas more efficient compared to other firms.

Consequently, this paper empirically investigates how the coherence and diversity of knowledge stock affects firm productivity using a data set of firms in the pharmaceutical industry. We develop a model which relates the total factor productivity of a firm to the characteristics of its heterogeneous knowledge. In addition, we also test whether the contribution of different types of knowledge (in our case technological and project development knowledge) varies in relation to productivity.

The rest of the paper is organized as follows: Section 2 describes our view on how characteristics of knowledge such as diversity and coherence may affect firm productivity; Section 3 describes the data set; Section 4 explains how our model can be tested empirically; Section 5 reports the results of the empirical investigation; Section 6 summarizes and concludes.

2 Knowledge valuation function

In this section, we try to sketch some possible approaches construct econometric specifications in order to assess the economic value of a firm's knowledge structure. The discussion is not meant, however, to derive a strict functional form for the valuation of knowledge portfolios. Rather, we are interested in identifying a small set of indicators which are able to capture independent and complementary effects. The obtained indicators will be used in the next sections, inside flexible parametric specifications, to measure the relative strength of these effects. The advantage of deriving a set of statistical indicators through formal analysis is that, in this manner, their relative meaning becomes more apparent and, consequently, the interpretation of the results becomes more straightforward.

Consider the problem from an abstract perspective. Assume a firm exists that is active in L "technological classes". These classes can represent different domains of technical or scientific knowledge, in principle associated with different designs and productive capabilities. Let $k_1, \ldots, k_L$ be the degrees of "knowledge" a given firm possesses in the different classes, with $k_l \geq 0$ for $l \in \{1, \ldots, L\}$. Even if how it is measured is unspecified, the important concept is that the degree of knowledge, k_l, is a (continuous or discrete) variable having a cardinal character, so that $2k_l$

represents two times the knowledge k_l. Thus, the spread of a firm's knowledge, at a certain point in time, is captured by a "knowledge portfolio" $(k_1, \ldots, k_L)$. Let's denote the economic value of this portfolio by $V(k_1, \ldots, k_L)$. In principle, many possible functional forms could be utilized for V. We will restrict the set of possibilities by requiring that all classes have the same "intrinsic" value; that is, that a technical advancement in one class, captured by an increase in the level of a specific k, is not intrinsically more valuable than an improvement in a different class. This amounts to assuming that the function V is completely symmetric in its arguments, so that

$$V(k_1, \ldots, k_i, \ldots, k_j, \ldots, k_L) = V(k_1, \ldots, k_j, \ldots, k_i, \ldots, k_L) \ , \quad \forall 1 \leq j \leq L \ . \quad (1)$$

Another reasonable assumption is that the addition of a new class may not decrease the overall value of a firm's knowledge portfolio. Furthermore, if the firm expands its activity in a new class $L+1$ by any amount k_{L+1}, the portfolio value is not reduced, i.e.

$$V(k_1, \ldots, k_L, k_{L+1}) \geq V(k_1, \ldots, k_L) \ . \quad (2)$$

At the same time, the value of a portfolio is non decreasing in any knowledge level, so that

$$\frac{\partial V}{\partial k_l} \geq 0 \ , \quad \forall 1 \leq j \leq L \ .$$

It cab be immediately verified that the above conditions are fulfilled by using the simple additive linear form

$$V(k_1, \ldots, k_L) \sim \sum_{l=1}^{L} k_l \quad , \quad (3)$$

where the total value of the portfolio is proportional to the sum of the knowledge level in all the classes. Notice that this specification is equivalent to assuming that there are no economies of scope from technical capabilities. Indeed, under the previous functional form, the value of the knowledge portfolio $(k_1, \ldots, k_L)$ is equivalent to a portfolio made up of a single class, but in which the knowledge level is the sum of the knowledge levels in the L different classes

$$V(k_1, \ldots, k_L) = V(\sum_{l=1}^{L} k_l) \quad .$$

Therefore, the linear specification in (3) implies a lack of scope economies.

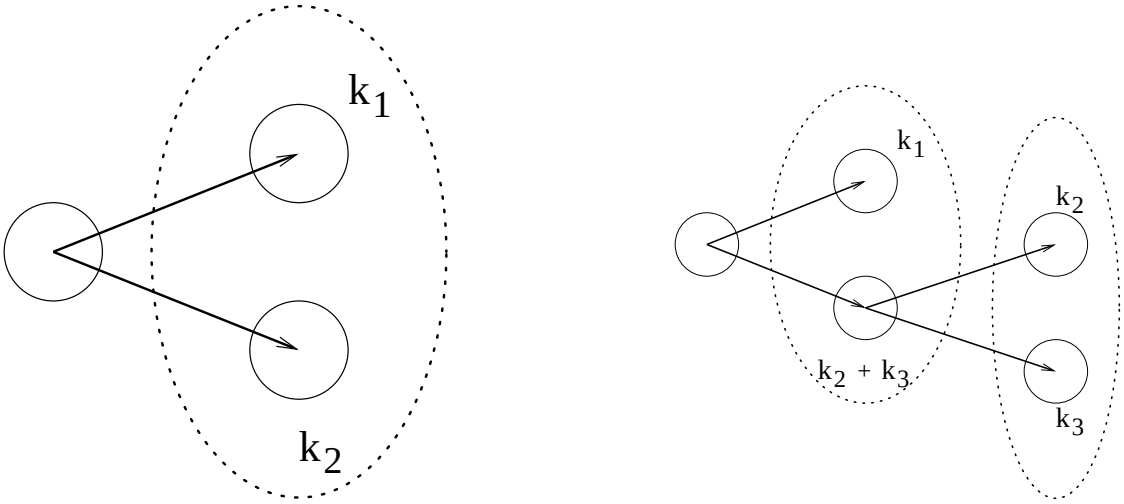

Figure 1: Hierarchical structure of the knowledge portfolio
In the right-hand panel two new classes stem from an original"branch", creating additional value for the firm.

Accordingly, assume instead that the contribution of a newly added knowledge class increases the total value of the portfolio, not by an amount proportional to its level, but rather so that the value of the new portfolio includes both the knowledge in the new class and the sum of all previous knowledge levels

$$V(k_1, \ldots, k_L, k_{L+1}) = V(k_1, \ldots, k_L) + V(k_1 + \ldots + k_L, k_{L+1}) \; . \tag{4}$$

Notice that (4) provides a recursive definition of the valuation function V: the L-dimensional function is defined in terms of a mixture of $(L-1)$-dimensional and 2-dimensional functions. Therefore, knowing the function V for two classes is enough to provide the valuation of any portfolio with a generic number of classes L. In fact, the fulfillment of (4) restricts the possible functional form of V according to the following proposition:

Proposition 2.1. *Consider a positive, continuous and symmetric valuation function V, which satisfies* (2) *and* (4)*. Then, for any portfolio $(k_1, \ldots, k_L)$, its expression would be*

$$V(k_1, \ldots, k_L) = -\sum_{l=1}^{L} k_l \log \frac{k_l}{k}, \tag{5}$$

where $k = \sum_{l=1}^{L} k_l$ is the total amount knowledge.

Proof. Define the function

$$S(k_1, \ldots, k_L) = V(k_1, \ldots, k_L)/k \quad .$$

Property (4), rewritten in terms of S, reads

$$S(k_1,\ldots,k_L,k_{L+1}) = S(k_1,\ldots,k_j+k_{L+1},\ldots,k_L) + \frac{\sum_{l=1}^{L} k_l}{\sum_{l=1}^{L+1} k_l} S(k_1,\ldots,k_L) \quad .$$

The equation above, together with the desired continuity and symmetry, is equivalent of Theorem 2 in Shannon (1948). Consequently, S is proportional to the entropy function and expression (5) immediately follows. □

Equation (4) implies that the value added to the portfolio by the development of knowledge k_{l+1} in a new class is not proportional to the knowledge level by itself, but is achieved by the interaction of the new class with the previously existing ones. Indeed, the value added to the portfolio by the inclusion of a new class is equal to a portfolio comprised of the new knowledge level and all previous knowledge as if it belonged to a single class. In other words, previously existing knowledge seems to have the potential to interact with newly generated knowledge to create an additional value. This feature, which directly introduces economies of scope into the technology valuation function, is consistent with the presence of a "hierarchical" structure which connects knowledge in different classes to each other. Indeed, it immediately follows from (5) that

$$V(k_1,\ldots,k_L,k_{L+1}) = V(k_1,\ldots,k_j+k_{L+1},\ldots,k_L)+V(k_j,k_{L+1}) \ , \quad \forall j \ , \quad 1 \leq j \leq L \ . \tag{6}$$

To understand the meaning of the previous relation, look at the picture in Fig. 1. Let us consider a firm having developed knowledge in two classes (k_1, k_2). Let $V(k_1, k_2)$ denote the value of its portfolio. This situation is similar to the left panel if Fig. 1. Now, assume the same firm develops a new knowledge level, k_3, in a third class. This new class can be thought as a new field branching out from an existing one, as in the right panel of Fig. 1. The effect on the value of the portfolio is then twofold: the previous value is increased by the addition of the new knowledge to the original "branch", so that its level increases from k_2 to $k_2 + k_3$. This would amount to a term equal to $V(k_1, k_2 + k_3)$. In addition, we assume that the two branches stemming from the original one do in fact "interact", to create a new value for the firm equal to that it would have if it possessed only those branches. This generates an additional value equal to $V(k_2, k_3)$ so that the total value of the portfolio becomes

$$V(k_1, k_2, k_3) = V(k_1, k_2 + k_3) + V(k_2, k_3) \ .$$

The "branching" that caused the amount of knowledge $k_2 + k_3$ to be spread across two different classes is actually generating value for the firm portfolio.

To summarize, from a novel approach to the way in which newly developed knowledge classes impact the value of knowledge portfolio, we have found new and interesting expressions for the knowledge valuation function. Under the approach which implies of absence of economies of scope and an equality between all possible distributions of knowledge, the linear expression (3) immediately follows. However, if instead we assume that the new value is not merely produced by an increase in the knowledge level of one specific class, but rather derived from the interaction of the knowledge in this class with the other classes, we are quite naturally led to the expression (5). Factorizing by a term proportional to the total knowledge in the portfolio k, (5) can be rewritten as

$$V(k_1, \ldots, k_L) = kS(k_1/k, \ldots, k_L/k) \ , \tag{7}$$

where S represents the Shannon entropy of the distribution of knowledge levels across different classes.

2.1 Introducing relatedness

The formal expressions above have been obtained by the assumption of equivalent and symmetric classes of knowledge. In practice, however, empirical data used to measure a firm's knowledge portfolio (patents, projects, R&D expenditure, products, etc.), generally, refers to classifications where the elements van rarely be considered independent. A good example is that of chemical classifications, where a given compound can in principle be related to very different industrial processes or chemical products. Furthermore, the same is true for the pharmaceutical sector, in which classifications based on either therapeutic or chemical properties can overlap in complex ways. Knowledge used in the production of a given molecule or chemical entity could be relevant in producing very different final compounds and drugs and, consequently, affect the ability of the firm to participate in seemingly distant markets.

This assumption of symmetry and equivalence can be relaxed by introducing a measure of "relatedness" in the conceptualization of the knowledge classes. Various measures have been suggested in the literature, most notably the coefficient of variation, measuring the departure from a theoretical hypergeometric distribution in Teece et al. (1994), and the correlation coefficient measure in Breschi et al. (2003). The common idea within this approach is to have, for any pair of classes l

and m, a number $\tau_{l,m}$, which measures how important the knowledge developed in class l is for the activities related to class m.

Assume again to have L classes and let $(e_1, \ldots, e_L)$ be the empirically observed activities (number of patents or projects, R&D expenses, etc.) that a given firm pursues in the different classes. However, the actual knowledge level in class l, k_l, is not only measured by the activities directly classified in this class, e_l, but also depends on the empirically observed activities in other classes, illustrated by the relatedness indices, τ's. Using a simple linear additive expression for the contribution of the different classes, this relationship can be written

$$k_l = e_l + \sum_{m=1, m \neq l}^{L} \tau_{l,m} e_m \ . \tag{8}$$

Recall that what is directly observed is not the knowledge levels ,k's, but the activity levels of the firm, e's. Therefore, in order to have a directly testable specification, the valuation function needs to be rewritten in terms of the latter. Substituting (8) into the linear specification of the knowledge value (3) gives

$$V(e_1, \ldots, e_L) \sim \sum_l k_l = e \ (1 + (L-1)\, R) \quad , \tag{9}$$

where $e = \sum_{m=1}^{L} e_l$ represents the total level of activity for the firm and R is the average relatedness level of the firm portfolio, obtained as a relatedness-weighted average of portfolio shares

$$R = \frac{1}{L-1} \sum_{l=1}^{L} \sum_{m=1, m \neq l}^{L} \tau_{l.m} \frac{e_m}{e} \ . \tag{10}$$

In accordance with our hypothesis of a hierarchical structure, the substitution of (8) into (5) yields

$$V(e_1, \ldots, e_L) \sim e \ (1 + (L-1)\, R) \ S(\phi_1, \ldots, \phi_L) \ , \tag{11}$$

where

$$\phi_l = \frac{k_l}{k} = \frac{e_l + \sum_{m \neq l} \tau_{l,m} e_m}{e(1 + (L-1)\, R)}$$

are the knowledge shares computed by taking into account the relatedness matrix $\tau_{l,m}$.

2.2 Empirical estimation

When using the above functional form in empirical analysis, it is important to remember that neither the linear nor the hierarchical hypothesis of the structure of the valuation function are exactly replicated in reality. For this reason, some degree of variability between the different expressions has to be taken into account. For instance, a linear combination of (3) and (5) could be considered. For this linear combination, in the case in which the relatedness among the different classes is weak, i.e. $(L-1)R << 1$, the correction 11 of the previous section can be ignored and observed e's can be simply considered as proxies for the knowledge levels. Taking a linear combination in terms of the activity levels of (3) and (7) the expression

$$V(e_1, \ldots, e_L) \sim c_1\, e + c_2\, eS(e_1, \ldots, e_L)$$

can be directly estimated. After taking the logarithms and adding a scale factor (as a further generalization) the equation becomes

$$\log V(e_1, \ldots, e_L) \sim \alpha_1\, \log e + \alpha_2 S(e_1, \ldots, e_L)\ , \tag{12}$$

where we consider $S(e_1, \ldots, e_L) >> 1$.[1]

If $(L-1)R >> 1$[2], or when the average degree of relatedness in firms portfolios is large, the above approximation cannot be applied and one has to consider a relatedness-corrected valuation function. In particular, the entropy of the portfolio should be computed according to 11. Ignoring sub-leading corrections, this reduces to the log-linear approximation

$$\log V(e_1, \ldots, e_L) \sim \log e + \log L + \log R + \log S(\phi_1, \ldots, \phi_L)\ . \tag{13}$$

This approximation contains a term corresponding to the total amount of activity, e; a term corresponding to the number of active sectors, L; a term corresponding to the overall degree of portfolio relatedness R; and a term corresponding to the entropy of the portfolio.

3 Data sources

In order to empirically test the effect of knowledge characteristics on firm productivity, we choose to use data from firms developing medicine. The motivation for

[1] The entropy spins the support $[0, \log L]$. As long as L is large and firms are sufficiently diversified, the assumption of large entropy remains safe.

[2] This case is applicable to our data, with respect to both patent and project knowledge

this choice is that, due to the availability of the information related to development of drug projects we can easily refer to the knowledge related to these projects. This feature of the data set contributes to the novelty of this research, compared to similar research where only the patent data is analyzed (Nesta, 2008).

The information on drug development projects originates from the BioPharmInsight[3] website, which provides publicly available information on new drug development projects worldwide. Each project in the data represents a drug-in-development at the stage of clinical testing.

The feature of drug development projects which makes it possible to estimate the characteristics of knowledge heterogeneity such as diversity and coherence, is the classification of them into 225 Indications. Each indication represents a condition (i.e. a disease, or a symptom), which causes a particular procedure or treatment to be advisable[4]. An example of an indication would be "Breast Cancer", "Influenza" or "Heart Stroke". The assignment of indications according to diseases to be treated makes it reasonable to assume that each indication represents a particular market niche within the general drug market. Consequently, we treat this piece of information as a reflection of the knowledge related to project development.

An example of the raw project data used is given in Table 1. In the table, different projects of the company Bayer AG are represented. In our analysis, projects belonging to a similar indication can be recorded for different years. For example, Bayer AG had two projects in the indication "Kidney Cancer" in 2005 and 2003. Additionally, in 2005, it had projects both in liver cancer and arterial/vascular disease.

Table 1: Example of Project Data.

Company	**Indicator**	**Year**
BAYER AG	Kidney Cancer	2005
BAYER AG	Kidney Cancer	2003
BAYER AG	Liver Cancer	2005
BAYER AG	Liver Cancer	2005
BAYER AG	Lung Cancer	2006
BAYER AG	Melanoma	2006
BAYER AG	Arterial / Vascular Disease	2006
BAYER AG	Arterial / Vascular Disease	2005

In order to provide a comparison between different classes of knowledge and their effect on productivity, the project data have been enriched by adding data

[3]http://www.infinata5.com/biopharm/

[4]This definition is from the online medical dictionary: http://www.medterms.com

on patent applications at the firm level. As the unit of analysis is a single firm, patent applications to the US Patent Office by the firms in our data set on drug development were assigned to the respective firm.

Each patent is then classified according to the International Patent Classification, which provides the classification of patents according to the different areas of technology where they belong. IPC, that is a 4-digit classification, was chosen to represent each technology. For each patent, the information included in the data is the time of application, company name (applicant), and IPC4 classification codes. We suggest that the patent data represents technological knowledge, as opposed to project knowledge.

Firm-level data, in the form of balance sheet information, have been acquired from the compustat database for each firm. As better explained in the next section, we only use some information from the compustat data base, specifically that which is relevant for productivity estimation. Moreover, as stated previously, only firms that appear in project-in-development data have been selected for the analysis.

Before combining these three data sets[5], the variables characterizing the patent and the project portfolios have been calculated as described below. These variables correspond to the model in section 2.

Since the data set was formed by selecting firms operating in a certain market, we expect to observe fewer inter-firm differences in terms of the variability of productivity, capital, labor and knowledge than there would be in a inter-industry study (e.g. Nesta (2008)). Even though the number of estimated units and observations might be reduced, the more focused data set should provide more reliable and precise estimation results, due to removing the problem of industry specific effects at the high aggregation level.

4 Methodology

In our empirical estimation, we follow a two-step procedure. In the first step, we estimate productivity. In the second step, the derived productivity is assumed to be a function of knowledge characteristics, as described in section 2.

We chose to measure productivity as a Solow residual (or total factor productivity), due to our belief that knowledge should affect the overall efficiency the firm. Although knowledge may be considered as an additional factor of production, we think that it is different from other factors, such as labor and capital. Moreover, total factor productivity allows for more flexibility in econometric estimation, due

[5]The overall time span of the combined data is 1985-2004.

to the absence of any complex relationship with other production factors.

Total factor productivity is factor A in the standard Cobb-Douglas representation of the production function:

$$Q = AK^{\beta_K} L^{\beta_L} \tag{14}$$

Taking the log of both sides of this equation results in:

$$q = a + \beta_K k + \beta_L l \tag{15}$$

The lowercase letters represent the logs of the corresponding variables in the equation (14).

The logarithm of total factor productivity can be derived from equation (15) using the following specification:

$$a = q - \beta_K k - \beta_L l \tag{16}$$

The term a captures differences in output across firms that are not accounted for by changes in the input use. It is typically referred to as total factor productivity or multi-factor productivity. As we do not include knowledge characteristics explicitly as inputs in the production function (like we do with labor or capital), a will be our dependent variable in second step of the empirical estimation.

As a second step, we assume that the total factor productivity derived in equation (16) is a function of knowledge characteristics such as entropy, diversity, coherence and the total accumulated knowledge. Generally speaking, the dependence of productivity on knowledge can be expressed by equation (13) in section 2, where the logarithmic value of knowledge, $log(V)$, is assumed to be proportional to total factor productivity [6].

$$a \sim log(e) + log(L) + log(R) + log(S) \tag{17}$$

[6]Two estimation strategies have been proposed depending on the size of $(L-1)R$. Since $(L-1)R >> 1$ is true for our data, we apply only the second proposed specification (equation (13)).

4.1 Productivity estimation using compustat data[7] [8].

In our estimation of productivity using compustat data, we follow the methodology for measuring capital and labor proposed by Brynjolfsson and Hitt (2003).

Output

Value added is used as a proxy for output Q. Value added allows us to use only labor and capital on the right-hand side of the production function without controlling for materials. In other words, using value added in the productivity estimation should produce less biased results. Value added (VA) is calculated according to the formula:

$$VA = (operating\ income) + (labor\ cost) \tag{18}$$

Operating income corresponds to the compustat item "operating income before depreciation". Operating income is equal to operating revenue minus operating expenses. Therefore, the cost of materials is not present in value added and thus should not be controlled for in the further estimation.

As we utilize data from different years, our measures for output, labor and capital must be deflated. The deflated version of value added (in year 2000 dollars) is obtained applying the output deflator, *vapi* [9].

$$dVA(t) = VA(t) * 100/vapi(t). \tag{19}$$

Labor

Labor cost (LC) is computed as the number of employees ($EMPL$) multiplied by the average labor cost per person per year (ALC).

$$LC = EMPL * ALC. \tag{20}$$

The deflated version of labor costs (in year 2000 dollars) uses the employee cost index (*eci*) for US manufacturing

$$dLC(t) = LC(t) * eci(2000)/eci(t). \tag{21}$$

[7] In the estimation, we also applied a different approach to derive productivity measurements from our data. This other productivity measure produced largely similar results to those reported in the paper. See Appendix, section B for the alternative estimation of TFP.

[8] In this section, we describe the construction of the variables through available data. For all additional sources of data, as well as a description of the variables go to section A of the Appendix.

[9] Since the *vapi* deflator is available only for NAICS classification, in order to apply it to SIC, we use the nearest industry or, if necessary, a lower disaggregation level.

Capital

The definition of capital, K, which enters into the total factor productivity estimate is gross capital stock.

$$K(t) = GK(t). \tag{22}$$

The deflated version of capital (in year 2000 dollars) is computed by applying the deflator relative to the average age (average vintage) of the capital stock. The average age of the capital stock is calculated by dividing total amortization by annual amortization (Hall, 1990; Brynjolfsson and Hitt, 2003). This method is generally believed to be more precise in estimating the accumulated capital stock:

$$dK(t) = GK(t) * ppi(2000)/ppi(t - int((GK(t) - NK(t))/DEP(t))), \tag{23}$$

where $int()$ is the nearest integer function; NK is net capital stock, DEP is depreciation and ppi is producer price index.

Productivity

After deflating the value added, capital and labor proxies, the estimate of productivity is:

$$a = log(dVA(t)) - log(dK(t)) - log(dLC(t)) \tag{24}$$

4.2 Knowledge variables

We employ two different sources of knowledge (patents and projects) in our estimation. Therefore, each knowledge variable has two variants: one which is calculated on patent and the other using project data. The corresponding variables have a subscript t and p respectively (for a description of variables, see table 7 in Appendix).

Considering project and patent knowledge separately is motivated by the idea that there may be differences in the effect of these types of knowledge on productivity. Previous research points at the difference between creating new technologies and developing R&D projects aimed at certain market niches (Cantner and Plotnikova, 2009; Plotnikova, 2009). Since knowledge in our view is acquired through learning-by-doing (which is reflected in the model in section 2), different activities should generate different types of knowledge. In the case of projects, knowledge is generated not only through the creation of a drug, but also through its testing, development and market introduction. Conversely, knowledge generated through

patenting is more probable to contain rather technical know-how.

It is common to assume that the rate of knowledge depreciation is 20% (Pakes and Schankerman, 1979; Henderson and Cockburn, 1996). We follow this assumption, however we also assume a linear depreciation of knowledge. This last assumption is made in order to avoid complications in the calculation of the diversity and coherence measures. Consequently, accumulated knowledge and its characteristics are measured over five-year periods.

Accumulated knowledge (e)

Accumulated technological knowledge is measured as the number of patents over a five-year period. Similarly, accumulated project knowledge is the total number of projects a firm was developing over the last five years.

Diversity (L)

Diversity is the number of different groups of technologies or projects. Technological knowledge diversity reflects the number of different IPC classes in the firm patent portfolio over a five-year periods. Project knowledge diversity is the number of different indications in the firm's project portfolio over the same five years.

Entropy (S)

Knowledge entropy is a measure of the stock of technologies (projects) accumulated by a firm over the last five years. Technology is defined on the 4-digit level according to the International Patent Classification (IPC). A project is defined by its relevant indication (the condition or disease to be treated by a potential product/medicine). Accordingly, the portfolio of technologies consists of different groups, as defined by IPC. Meanwhile, a project portfolio consists of several potential products grouped by indication.

If x_k is the fraction of projects (technologies) of type k within the the total number of projects (technologies) in the firm portfolio, then the knowledge entropy can be calculated by the following equation:

$$S = Entropy = -\sum_k x_k ln(x_k) \tag{25}$$

In our estimation, we distinguish between the entropy of the project portfolio, S_p, calculated using project data, and the entropy of the technology portfolio, S_t, calculated using patent data. As a reminder, the calculation of entropy controls for the relatedness between knowledge pieces so that we can be sure that our entropy is not affected by the relationship between various knowledge (see formula for ϕ_l in

equation (11) in section 2). By doing so, we prevent possible multicellularity due to correlation between relatedness and entropy in the further estimation.

It is also worth mentioning that our entropy measurement captures the degree of evenness of the distribution of different projects or technologies within the firm portfolio. Therefore, the measure increases for a more uneven distribution of projects (technologies) across different indications (IPC codes).

Coherence (R)

In order to measure the relatedness between different pieces of knowledge (IPC classes for technological knowledge and indications for projects), we employ the cosine index as introduced in Jaffe (1986) and later used in Breschi et al. (2003). The cosine index in Breschi et al. (2003) measures "the angular separation between the vectors representing co-occurrences of technological fields i and j with all other fields", as shown in equation (26). The intuition behind this index is that the degree of relatedness captures the inverted distance between different projects or technologies. This index varies from zero to one, one representing the maximum possible relatedness; therefore, one represents the closest possible distance and zero represents the longest potential distance between two projects or technologies.

$$Cosine_{ij} = \frac{\sum_k C_{ik}C_{jk}}{\sqrt{\sum_k C_{ik}^2}\sqrt{\sum_k C_{jk}^2}}, \tag{26}$$

where C_{ij} is the number of patent documents classified in both technological fields i and j.

We followed a similar procedure to estimate the relatedness of project areas (defined by indications). To calculate the cosine index for projects, we considered C_{ij} to be the number of firm portfolios which contain both projects i and j.

The cosine index is then used to calculate the knowledge portfolio coherence for each firm. The knowledge portfolio coherence was calculated in accordance with Teece et al. (1994). However, instead of the relatedness index in Teece et al. (1994), we utilize the cosine index here:

$$R = \sum_j \left(\frac{P_j}{\sum_j P_j} \left(\frac{\sum_{i \neq j} Cosine_{ij} P_j}{\sum_{i \neq j} P_j} \right) \right) \tag{27}$$

In this equation, P_j is the number of products (technologies) j in a firm's portfolio; $Cosine$ is the cosine index, which acts as our measure of relatedness. All portfolios containing exactly one technology (project), meaning all specialized portfolios, were assigned a coherence value equal to one.

In order to separate technological knowledge from project knowledge, as well as

to account for a possible bias in relatedness estimation, technological relatedness was calculated via the number of co-occurrences of technologies (IPC codes) in patents.

Relatedness matrices were then calculated for the whole sample of projects and patents, a much wider sample than that used for the empirical analysis, since it also includes firms which are not listed in the compustat database. For this reason, relatedness and coherence measures are not expected to be biased due to a small sample size.

4.3 Econometric specification

As described earlier, the first part of our estimation follows equation (24). Therefore, the productivity estimator is derived as a residual after regressing the logarithm of value added on the logarithms of labor and capital, each variable is deflated to account for annual inflation. Next, we use this productivity measure as the dependent variable in the regression conducted in accordance with equation (17).

In the second part of our estimation, there are two main concerns when regressing productivity on knowledge characteristics: autocorrelation of residuals and unit heterogeneity. Autocorrelation generally arises from path dependency in the productivity development, as well as the knowledge variables. Since the sample represents an unbalanced panel of firms, we want to exploit this panel structure of our data in order to control for unobserved firm characteristics. Therefore, given these considerations, any econometric specification utilized would need to control for firm heterogeneity, at the same time accounting for the time structure of residuals.

Firstly, we apply an ordinary least squares (OLS) model to estimate equation (17), and then test for the presence of autocorrelation in the residuals. Secondly, we test the relevance of a least squares with unit dummy variables (LSDV) specification. For this specification, we use both firm and industry panel data as we expected that there might be both unobserved industry (4-digit SIC classification) and firm-level effects. Since our sample of firms deals with firms developing products in a single market (drug market), firm-level effects would seem to be more important, and therefore we report only those results where the firm specific effects are controlled for.

As a next step, we apply a model that accounts for potential autocorrelation in the residuals. More precisely, we apply a generalized least squares method with a Prais-Winsten transformation of residuals (Prais and Winsten, 1954). Finally,

fixed effects and random effects specifications, assuming no autocorrelation and controlling for panel autocorrelation in the residuals, are also considered. In the first method, the main purpose is to choose the correct specification which then allows us to control for any unobserved unit heterogeneity in the sample. The second method, however, is more concerned with capturing any existing panel autocorrelation, implying that autocorrelation patterns vary in different panels (i.e. for different firms).

This estimation strategy was applied to three samples: patent data, project data, and pooled patent and project data. This method of sampling is helpful for better understanding of the sample, because patent and project information is available for overlapping, but not always matching, observations. By estimating these types of information separately, more observations are available for each sample. Moreover, by separating the patent and project data, our estimation is able to distinguish between the effects of different types of knowledge on the dependent variable, firm productivity.

The dispersion of firms in the data across industries, according to SIC 4-digit industrial classification, is quite broad (note that our selection of firms depends on whether or not they were developing pharmaceutical products). Therefore, in order to test whether our results hold for a smaller sample, we restricted our estimation to the sample of firms belonging to the industry "Chemicals and allied products" (SIC28). However, since the results of this estimation proved to be similar to those for the whole sample, they are not reported in the paper.

Consequently, we do not report the results of every estimation performed. Instead, we only list those tables which we think best demonstrate the effect of knowledge characteristics on productivity. The correlation of variables is reported in table 9 of the Appendix. Moreover, a short description of variables is given in table 7 in section C of Appendix. The next section reports the results of our empirical estimation.

5 Results

This section describes the results of applying the estimation strategy described in the previous section. Three sets of estimations have been conducted in order to explore the relationship between productivity and knowledge heterogeneity separately for two different types of knowledge: technological knowledge (measured based on patent data) and development knowledge (measured via project data). Afterward, the characteristics of both technological and development knowledge have been

pooled into single estimation equation in order to further explain productivity.

5.1 Patents

We first report the results where only technological knowledge from the patent statistics, is taken into consideration. The following table (table 2) contains the estimated coefficients.

This table contains the results from the estimation of five distinct models. Autocorrelation has been detected in the ordinary least squares model (model 1) and the least squares dummy variable model (model 2). Consequently, the general least squares estimation (model 3) was applied in order to correct for the first-order autocorrelation. Lastly, model 4 and 5 assume a random effects specification. In accordance with the result of a Hausman specification test (Hausman, 1978), random effects specification is preferred to a fixed effects specification.

Table 2: Estimation Results for Patents

Productivity	**(1) OLS**	**(2) LSDV**	**(3) GLS**	**(4) RE**	**(5) RE AR1**
$log(e_t)$	**0.400*** **	**0.243*** **	**0.214*** **	**0.258*** **	**0.192*** **
	(0.0535)	(0.0699)	(0.0816)	(0.0600)	(0.0736)
$log(L_t)$	**-0.660*** **	**-0.230***	**-0.163**	**-0.289** **	**-0.212**
	(0.0867)	(0.129)	(0.148)	(0.114)	(0.130)
$log(R_t)$	**0.0449**	**-0.0417**	**-0.0605**	**-0.0335**	**-0.0600**
	(0.0428)	(0.0556)	(0.0400)	(0.0453)	(0.0392)
$log(S_t)$	**0.0322**	**-0.213***	**-0.311** **	**-0.163**	**-0.279**
	(0.127)	(0.121)	(0.141)	(0.137)	(0.178)
Firm dummies		**yes**	**yes**		
Constant	**0.408*** **	**-1.062*** **	**-1.442*** **	**-0.161**	**-0.204**
	(0.135)	(0.169)	(0.166)	(0.175)	(0.172)
Observations	439	439	439	439	439
R-squared	0.149	0.684	0.549		
Number of firm				54	54
Autocorrelation†	yes	yes		yes	
Durbin-Watson‡			1.208261	0.877297	
Hausman				RE	RE,10%
Baltagi-Wu LBI				1.555463	

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
† Wooldridge test for panel autocorrelation
‡ modified Durbin-Watson (Bhargava et al., 1982)

Autocorrelation in the panel data has been tested for through a Wooldridge test (Wooldridge, 2002), a modified Durbin-Watson (Bhargava et al., 1982) and the test for autocorrelation in unequally spaced panel data developed by Baltagi and Wu (1999). Since our panel of firms is not balanced, we hypothesize that the Baltagi and Wu (1999) test should be the most reliable to reveal whether

autocorrelation is present in our sample. In the model 4, the test statistic for Baltagi-Wu test is not very close to two[10], therefore it is safe to assume some (although not very pronounced) autocorrelation in our panel data. Consequently, the last model, which allows for random panel effects and first order autocorrelation (model 5), is considered to be the most robust specification for the estimation for the current data.

The results of the application of models 1-5 suggest that, in the case of patents, accumulated knowledge ($log(e_t)$) positively affects productivity, since the coefficient on the number of patents is positively significant. At the same time, the diversity of technologies in the firm knowledge portfolio ($log(L_t)$) is found to have a negative impact on productivity. However, this result does not hold for every model. For instance, in the most robust model, model 5, the coefficient on technological diversity is not significant.

The entropy of firm technological knowledge ($log(S_t)$) is significant and negative in the model from which an unbiased estimator was not expected due to autocorrelation (model 2), and where the autocorrelation correction is not very successful (in model 3, the Durbin-Watson statistic is far from two). Therefore, we cannot provide any meaningful conclusions concerning this variable.

The coherence of technological knowledge ($log(R_t)$) is not significant in any specification. Therefore, we can conclude that the coherence of technological knowledge is not correlated with firm productivity. In other words, there are no significant differences in productivity among firms with different levels of technological knowledge coherence.

5.2 Projects

Table 3 reports the empirical results for the sample of projects-in-development. Autocorrelation was not detected in the least squares dummy variable model (model 2). Therefore, we use a random effects specification and compare it with its fixed effects counterpart. According to a Hausman specification test (Hausman, 1978), the random-effects specification should be preferred over a fixed-effects specification. Moreover, since the Baltagi-Wu test statistic (Baltagi and Wu, 1999) is close to two, we conclude that there is no significant autocorrelation in our panel data. Consequently, in the case of project data, we do not need to proceed any further with a panel estimation that corrects for panel autocorrelation.

As opposed to technological knowledge, product-in-development accumulated knowledge ($log(e_p)$) does not have a significant impact on productivity. None of the

[10]Since our panel is not strongly unbalanced, we use Durbin-Watson significance intervals.

Table 3: Estimation Results for Projects

Productivity	(1) OLS	(2) LSDV	(3) RE
$log(e_p)$	**-0.145**	**-0.345**	**-0.294**
	(0.346)	(0.265)	(0.254)
$log(L_p)$	**0.473**	**0.577***	**0.511***
	(0.445)	(0.310)	(0.291)
$log(R_p)$	**-1.829**	**-2.705*****	**-2.739*****
	(1.215)	(1.035)	(0.924)
$log(S_p)$	**-0.606**	**-0.482****	**-0.445**
	(0.471)	(0.243)	(0.300)
Firm dummies		**yes**	
Constant	**-0.715***	**-0.937*****	**-0.769****
	(0.421)	(0.322)	(0.319)
Observations	228	228	228
R-squared	0.059	0.657	
Number of firm			35
Autocorrelation†	yes	no	yes
Durbin-Watson‡			1.0984
Hausman			RE
Baltagi-Wu LBI			1.810134

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
† Wooldridge test for panel autocorrelation
‡ modified Durbin-Watson (Bhargava et al., 1982)

reported models contains a significant coefficient for the corresponding variable.

Project knowledge diversity ($log(L_p)$) is positively significant after controlling for unit heterogeneity by introducing random effects (model 3), as well as in the LSDV specification (model 2). The sign on the diversity of project knowledge variable is positive, suggesting that firms with varied project knowledge tend to be more productive. In other words, more productive firms tend to possess more diversified project knowledge.

The coherence of project knowledge ($log(R_p)$) is negatively significant in both models 2 and 3. Recall that the coherence of technological knowledge was insignificant in the estimation of the impact of technological knowledge on productivity. The negative coefficient for project knowledge coherence means that firms with coherent project knowledge tend to be less productive. In other words, specializing in one indication is generally correlated with lower productivity.

The coefficient on the entropy of project knowledge ($log(S_p)$) is negatively significant in the model 2. Consequently, there is weak evidence that the entropy of project knowledge negatively affects productivity. Therefore, the uneven distribution of a project knowledge portfolio across an increasing number of project areas (indications) is negatively correlated with productivity.

5.3 Patents and projects

In the next estimation we combine project and technology knowledge characteristics, as variables affecting firm productivity, into one equation. Here we assume a simple specification, where the dependence of productivity on the value of technological and project knowledge is expressed as a multiplication of the values of technological and project knowledge. Accordingly, the logarithmic version of this dependence is expressed through the following equation:

$$a \sim log(e_t)+log(L_t)+log(R_t)+log(S_t)+log(e_p)+log(L_p)+log(R_p)+log(S_p) \quad (28)$$

Note that due to logarithmic specification, the potential interaction between different types of knowledge is not formally excluded.

Table 4: Estimation Results for Patents and Projects

Productivity	**(1)** OLS	**(2)** LSDV	**(3)** GLS	**(4)** RE
$log(e_t)$	**0.186***	**0.167****	**0.167****	**0.193****
	(0.0959)	(0.0791)	(0.0813)	(0.0932)
$log(L_t)$	**-0.272**	**-0.308****	**-0.299****	**-0.310***
	(0.175)	(0.122)	(0.126)	(0.183)
$log(R_t)$	**0.0390**	**0.109**	**0.106**	**0.0197**
	(0.114)	(0.0899)	(0.0929)	(0.109)
$log(S_t)$	**0.107**	**0.0767**	**0.0777**	**0.141**
	(0.242)	(0.202)	(0.207)	(0.300)
$log(e_p)$	**-0.0581**	**-0.274**	**-0.274**	**-0.249**
	(0.378)	(0.262)	(0.269)	(0.210)
$log(L_p)$	**0.215**	**0.439**	**0.434**	**0.381**
	(0.468)	(0.306)	(0.314)	(0.249)
$log(R_p)$	**-0.838**	**-1.704**	**-1.620**	**-1.490***
	(1.484)	(1.192)	(1.193)	(0.901)
$log(S_p)$	**0.0542**	**-0.191**	**-0.195**	**-0.142**
	(0.373)	(0.249)	(0.251)	(0.275)
Firm dummies		**yes**	**yes**	
Constant	**-0.141**	**-0.259**	**-2.988*****	**-0.208**
	(0.413)	(0.501)	(0.454)	(0.373)
Observations	157	157	157	157
R-squared	0.116	0.772	0.758	
Number of firm				25
Autocorrelation†	yes	yes, 10%		yes
Durbin-Watson‡		1.115054	1.168077	1.234805
Hausman				RE
Baltagi-Wu LBI				1.939832

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
† Wooldridge test for panel autocorrelation
‡ modified Durbin-Watson (Bhargava et al., 1982)

The first two models in table 4 indicate the presence of autocorrelation. Conse-

quently, a generalized least squares estimation is applied, with the results reported as model 3 estimation. Although the significance of detected autocorrelation decreases (judging from the change in the Durbin-Watson statistic from 1.115054 in LSDV (model 2) to 1.168077 in GLS (model 3)), the residuals from GLS estimation are still autocorrelated. However, when we used a random-effects specification, instead of a fixed-effects specification (the latter produces results similar to LSDV) there is no autocorrelation detected in the panel data, according to Baltagi-Wu test for an unbalanced panel. Taking into account that the Hausman test results suggested a random-effects estimation should be preffered to a fixed-effects estimation, model 4 seems to provide the best fit for our data.

As seen in table 4 we are now left with only three significant coefficients for the variables which correspond to the results of previous estimations. In specific, the amount of accumulated technological knowledge ($log(e_t)$) has a positive effect on productivity. Moreover, the diversity of technological knowledge ($log(L_t)$) and the coherence of project knowledge ($log(R_p)$) both affect productivity negatively. In other words, the more productive firms tend to be those with a large amount of technological knowledge, a lesser degree of project specialization and a reduced diversification of their technological portfolios.

6 Discussion and conclusion

There are some caveats that should be discussed prior to summarizing the main results of the paper. For instance, this analysis is performed on a relatively small data set, especially the last part of estimation which was performed using data from both the project and patent side. However, the advantage of this approach is that it contains information on a relatively homogeneous group of firms, which enables us to make conclusions which suffer less from various industry biases. Consequently, it is worth noticing that our results differ from similar analysis made by Nesta (2008) on a sample of firms from different industries.

Secondly, we did not model the interaction between project and patent knowledge explicitly, potentially leaving room for further investigation into the subject matter. This extension on the presented results could be further justified given the importance of the finding that various types of knowledge cause different effects on productivity.

We should again note that the productivity estimation was performed using compustat data including firms with mostly above average size. This, of course, poses the question of the effect of heterogeneous knowledge on the productivity of

small firms. Although extending the data to include smaller firms would also be an interesting way for further research, restricting the data to the sample of big firms still provides interesting results.

The results of our regressions allow us to conclude that there are some significant differences between the various types of a firm's knowledge and their effect on productivity. Specifically, the estimation performed utilizing only patent data demonstrates the significant effect of accumulated knowledge on productivity. The positive correlation between productivity and accumulated technological knowledge allows to claim that firms with larger sets of accumulated knowledge tend to be more productive. Note though, that because unobserved unit heterogeneity was controlled for in the estimation, it cannot be stated that this finding is equivalent to claiming that big firms are more productive. Instead, it just signifies that more technological knowledge is associated with higher productivity.

The strong finding on the project side is that high project knowledge coherence is associated with low productivity. Considering that projects represent future products in the process of development, this finding suggests that there may be certain disadvantages in developing a coherent project portfolio. It could be that coherent or specialized project knowledge does not allow firms to either take advantage of every opportunity, maintain flexibility and so to enrich its production process with new ideas. These facts may actually lead to lower productivity in those firms with more coherent project knowledge. On the other hand, the lower productivity of highly coherent firms could also be explained by the higher risks born by those firms due to poor diversification of their products-to-be portfolio.

Concerning the weaker findings of this study, the diversity of project knowledge positively affects productivity, suggesting that more diversified firms are more productive. However, this finding does not appear in either joint project or the patent regression, possibly due to the more complicated nature of the interaction between these two types of knowledge. Nevertheless, technological diversity becomes significant in the last regression.

We can thus conclude from the comparison of these regression results that different types of knowledge should be considered separately, and through possibly different theories and interpretations. The knowledge of a firm appears to be highly heterogeneous, not only with respect to different technology classes, but also and most importantly with respect to the various fields of a firm's activity.

References

Baltagi, B. H. and Wu, P. X. (1999). Unequally spaced panel data regressions with AR(1) disturbances. *Econometric Theory*, 15:814–823.

Bartelsman, E. J. and Doms, M. (2000). Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic Literature*, 38(3):569–594.

Bhargava, A., Franzini, L., and Narendranathan, W. (1982). Serial correlation and the fixed effects model. *The Review of Economic Studies*, 49(4):533–549.

Breschi, S., Lissoni, F., and Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1):69–87.

Brynjolfsson, E. and Hitt, L. M. (1995). Information technology as a factor of production: The role of differences among firms. *Economics of Innovation and New Technology*, 3(3):183–199.

Brynjolfsson, E. and Hitt, L. M. (2003). Computing productivity: Firm-level evidence. *The Review of Economics and Statistics*, 85(4):793–808.

Cantner, U. and Plotnikova, T. (2009). Technological diversity and future product diversity in the drug industry. *Jena Economic Research Papers*, 29-031.

Chennells, L. and Van Reenen, J. (1998). Establishment level earnings, technology and the growth of inequality: Evidence from britain. *Economics of Innovation and New Technology*, 5(2):139 – 164.

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal*, 17:109–122. Winter Special Issue.

Grilliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, 10(1):92–116.

Hall, B. H. (1990). The manufacturing sector master file: 1959-1987. *NBER Working Paper*, 3366.

Hall, B. H. and Mairesse, J. (1995). Exploring the relationship between R&D and productivity in french manufacturing firms. *Journal of Econometrics*, 65(1):263–293.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–71.

Henderson, R. M. and Cockburn, I. (1996). Scale, scope, and spillovers: The determinants of research productivity in drug discovery. *RAND Journal of Economics*, 27:32–59.

Jaffe, A. B. (1986). Technological opportunity and spillovers of r&d: Evidence from firms' patents and market value. *American Economic Review*, 76(5):984–1001.

Lichtenberg, F. R. and Siegel, D. (1991). The impact of R&D investment on productivity - New evidence using linked R&D–LRD data. *Economic Inquiry*, 29(2):203 – 229.

McGuckin, R. H., Streitwieser, M. L., and Doms, M. (1998). The effect of technology use on productivity growth. *Economics of Innovation and New Technology*, 7:1–26.

Montgomery, C. A. and Hariharan, S. (1991). Diversified expansion by large established firms. *Journal of Economic Behavior & Organization*, 15(1):71–89.

Nadiri, I. M. (1994). Innovations and technological spillovers. *NBER Working Paper*, 4423.

Nerkar, A. and Roberts, P. W. (2004). Technological and product-market experience and the success of new product introductions in the pharmaceutical industry. *Strategic Managemet Journal*, 25:779–799.

Nesta, L. (2008). Knowledge and productivity in the world's largest manufacturing corporations. *Journal of Economic Behavior and Organization*, 67:886–902.

Nooteboom, B. (2000). Learning by interaction: Absorptive capacity, cognitive distance and governance. *Journal of Management and Governance*, 4(1-2):69–92.

Pakes, A. and Schankerman, M. (1979). The rate of obsolescence of knowledge, research gestation lags, and the private rate of return to research resources. *NBER Working Paper Series*, (346).

Plotnikova, T. (2009). Technology, competition and the time of entry: Diversification patterns in the development of new drugs. (2009-078).

Prais, S. J. and Winsten, C. B. (1954). Trend estimators and serial correlation. *Cowles Foundation, Discussion Paper*, 383.

Scott, J. T. and Pascoe, G. (1987). Purposive diversification of r&d in manufacturing. *The Journal of Industrial Economics*, 36(2):193–205.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994). Understanding corporate coherence : Theory and evidence. *Journal of Economic Behavior & Organization*, 23(1):1–30.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

A Sources of data for productivity estimation

Table 5 lists variables used for productivity estimation. Their labels correspond to those in the main part of the paper.

Table 5: Variables

Name	Description	Source	Period
EMPL	Number of employees	COMPUSTAT ITEM 29	'50-'04
LRE	Labor and related expenses	COMPUSTAT ITEM 42	'50-'04
ALC	Average labor cost per person per year	BLS [1]	'88-'00
eci	Employee cost index, constant dollar	BLS [2]	'79-'05
GK	Gross capital stock	COMPUSTAT ITEM 7	'50-'04
NK	Net capital stock	COMPUSTAT ITEM 8	'50-'04
DEP	Current depreciation	COMPUSTAT ITEM 14	'50-'04
ppi	Producer price index	BLS [3] SIC	'84-'03
		NAICS	'67-'06
TS	Total sales	COMPUSTAT ITEM 12	'50-'04
OIBD	Operating income before depreciation	COMPUSTAT ITEM 13	'50-'04
vapi	Value added price index	BEA [4]	'47-'07

[1]File 'ntaa8800.zip' obtained from ftp://ftp.bls.gov/pub/special.requests/cew/SIC/history/national/ntaa8800.zip It contains Annual Average Employment (EMP), and Total Wages (WAGES) for 4-digit SIC industrial sectors. The average labor cost can be computed as ALC = WAGES/EMP.

[2]File 'eci.ecconst.txt' can be downloaded from ftp://ftp.bls.gov/pub/suppl/eci.ecconst.txt It contains 'price index' for compensation at constant dollar (2005=100).

[3]From the website http://www.bls.gov/ppi it is possible to obtain producer price index for 3 or 4 digit SIC sectors. Base price is December '85. Longer series can be obtained looking at "commodity" rather than "industry" data.

[4] From the website http://www.bea.gov/industry/gpotables. Chain-type price indexes for value added by industry. Reference year 2000=100. Industries are classified according to NAICS.

B Alternative productivity estimation

Value added ($VA2$) is calculated according to formula:

$$VA2 = (operating\ income\ in\ dollars) + (labor\ expenses\ in\ dollars) \tag{29}$$

$$Labor\ expenses = (number\ of\ employees) * (hour\ wage) * (number\ of\ working\ hours) \tag{30}$$

Operating income, capital and labor expenses were deflated using indexes of output, capital income and labor compensation respectively. All values are expressed in 2000 dollars.

Estimation of a measure of total factor productivity was attained from running regressions following equation 15 and then predicting residuals. Productivity attained using value added as output measure is called $mprod1$.

Compustat items

Table 6: Variables and their Correspondence to Compustat Items

Variable	Compustat correspondence
Net Sales	DATA12 in compustat: Net sales
Operating Income	DATA13 in compustat: Operational income
Number of Employees	DATA29 in compustat: Number of employees
Net Capital Stock	DATA8 in compustat: Capital stock net (less depreciation)

Table 6 contains variables used for productivity estimation together with the corresponding items in compustat database. Number of employees is a measure of labor and Net Capital Stock is a measure of capital in equation 14.

Note that all compustat items are expressed in millions of dollars, except for number of employees. Therefore, Net Sales, Operating income and net capital stock volumes were multiplied by 1.000.000 for estimation.

Deflators

Labor expenses, capital, net sales and operating income volumes were deflated according to labor compensation, capital income and output indexes from the historical multifactor productivity measurement tables from the US Bureau of Labor Statistics[11].

The set of tables "Historical multifactor productivity measures (SIC 1948-87 linked to NAICS 1987-2007)" (the name of the file is prod3.mfptablehis.zip) was downloaded from the web site of the US Bureau of Labor Statistics. These tables are available for the private business sector and private nonfarm business sector. We used the table "Net Multifactor Productivity and Cost, 1948-2007. Basic measures." for the private nonfarm business sector, which reports current dollar output, labor compensation and capital income as index values with year 2000 as the base year. From this table, we extracted the columns for Current Dollar Output, Labor Compensation in current dollars and Capital Income in current dollars. These three

[11] http://www.bls.gov/mfp/#tables

indexes were used to adjust output, net sales, employee expenditure and capital measures in the calculation of value added and productivity.

Wage and working hours

Hourly wage was attained from the US Bureau of Labor Statistics web site[12] by requesting table B-4, "Average hourly earnings of production and nonsupervisory workers on private nonfarm payrolls by industry sector and selected industry detail, seasonally adjusted" for non-durable goods production. As monthly (not yearly) data was derived, data for one month (July) was chosen to represent hourly wage.

Number of working hours was assumed to be 2040 per year, following Brynjolfsson and Hitt (2003).

C Tables

Table 7: Description of Variables in Regressions

Variable	Description
a	logarithm of productivity as in equation 24
e_t	number of patents accumulated over 5 years
L_t	variety of patent portfolio over 5 years
S_t	entropy of patent portfolio over 5 years
R_t	coherence of patent portfolio over 5 years, following Breschi et al. (2003)
e_p	number of projects accumulated over 5 years
L_p	variety of project portfolio over 5 years
S_p	entropy of project portfolio over 5 years
R_p	coherence of project portfolio over 5 years, following Breschi et al. (2003)

[12] http://www.bls.gov/webapps/legacy/cesbtab4.htm

Table 8: Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
a	957	3.67E-10	0.951735	-4.40589	3.801741
e_t	6136	35.78113	179.3873	0	3087
e_p	2932	4.213847	17.38699	0	231
L_p	2932	2.773874	9.466475	0	99
L_t	6136	5.565026	18.64114	0	264
R_t	1731	0.241517	0.259368	0	1
R_p	691	0.795174	0.061246	0.412048	1
S_p	956	0.747732	0.55828	0	2.554782
S_t	1932	1.128958	0.623785	0	2.932398

Table 9: Correlation

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	a	**1**								
2	e_t	**0.0625**	**1**							
		0.0534								
3	e_p	**0.1229**	**0.1351**	**1**						
		0.0003	0							
4	L_p	**0.133**	**0.172**	**0.9803**	**1**					
		0.0001	0	0						
5	L_t	**-0.0882**	**0.7906**	**0.0764**	**0.0965**	**1**				
		0.0063	0	0	0					
6	R_t	**0.2344**	**-0.0791**	**0.0427**	**0.0579**	**-0.2499**	**1**			
		0	0.001	0.12	0.0352	0				
7	R_p	**-0.1478**	**-0.0614**	**-0.0321**	**-0.0527**	**-0.078**	**-0.0204**	**1**		
		0.0225	0.1068	0.3996	0.1665	0.0403	0.6943			
8	S_p	**0.2066**	**0.1387**	**0.5145**	**0.5972**	**0.0785**	**0.2162**	**-0.2272**	**1**	
		0.0002	0	0	0	0.0152	0	0		
9	S_t	**-0.1943**	**0.0663**	**-0.0225**	**-0.0341**	**0.3414**	**-0.5459**	**0.0325**	**-0.1666**	**1**
		0	0.0035	0.3887	0.1925	0	0	0.5157	0.0001	

Correlation values are bold. significance levels are reported below correlation values.