APPENDIX A

Table A1 Sectoral distribution

	Nace Rev2	Freq.	Percent
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	4,422	1.90
26	Manufacture of computer, electronic and optical products	37,932	16.29
	Tot High-Tech (HT)	***42,354***	***18.19***
20	Manufacture of chemicals and chemical products;	25,896	11.12
27	Manufacture of electrical equipment	44,838	19.26
28	Manufacture of machinery and equipment n.e.c.	111,870	48.06
29	Manufacture of motor vehicles, trailers and semi-trailers	2,100	0.90
30	Manufacture of other transport equipment	5,736	2.46
	Tot Medium-High-Tech (MHT)	***190,440***	***81.81***
	Total	**232,794**	**100.00**

APPENDIX B

R&D EQUATIONS

To describe the firm research behaviour, it is necessary to take into account the latent amount of R&D expenditures. We follow the model developed by Crépon, Duguet and Mairesse in 1998. The econometric specification of equation (1) leads to the two equations (4) and (5), where the first equation accounts for the fact that the firm is engaged in research activities, and the second one for the intensity of these activities.

Let $D\_R\&D_i$* be the latent dependent variable whether to invest in R&D or not, and $LnR\&D_i$* be the latent or true intensity of R&D investment of firm i. $D\_R\&D_i$ and $LnR\&D_i$ are the corresponding observed variables.

The two-equation R&D investment model is written as follows:

4) $D\_R\&D_i^* = \beta_1 x_i^1 + u_i^1$

with $D\_R\&D_i$=1 if $D\_R\&D^*_i$>0, $D\_R\&D_i$=0 otherwise.

5) $LnR\&D_i^*/(D\_R\&D_i = 1) = \beta_2 x_i^2 + u_i^2$

with $LnR\&D_i$= $LnR\&D^*_i$ if $LnR\&D_i$>0, $LnR\&D_i$=0 otherwise.

The x^1_i and x^2_i are the explanatory variables, β_1 and β_2 the respective coefficients u^1_i and u^2_i follow a bivariate normal distribution with correlation coefficient ρ.

The independent variable explaining, first, the probability to engage in R&D activities and, second, the levels of these activities, is intangible assets. Investment in intangible assets can be considered a reliable proxy for predicting R&D activities. Indeed, the broad array of activities that are necessary to explore and recombine the existing stock of knowledge, both internal and external to each firm, and exploit it can be predicted using this measure. The selection equation (4) also includes a measure of firm size. Finally, both equations include a set of industry and time dummies to capture market and cycle conditions (see the following section for all variables' detailed specification).

We estimate equations (4) and (5) following the methodology proposed by Wooldridge (1995) and applying bootstrapping to both equations. This method can be used in a panel setting to take into accounts that there may be some unobserved time-variant factors that can affect selection and influence R&D levels through the error term. In this approach, the time-invariant effects are assumed to be linked with x^1_{it} through a linear function of k^1_i on the time averages of x^1_{it} (denoted with x^1_i_bar) and an orthogonal error term a_i that exhibits no variation over time and is independent of x^1_{it} and u^1_{it}:

$k^1_i = x^1_i\_bar + a_i$

Equation (4) can be rewritten as follows:

$$4a)\ D\_R\&D_{it}^{*} = \beta_1 x_{it}^{1} + \gamma_1 \bar{x}_i^{1} + v_{it}^{1}$$

with the composite error term $v^1_{it} = u^1_{it} + a_i$ being independent from x^1_{it} and normally distributed with zero mean and variance σ2. In this approach, to obtain estimates for the Inverse Mill's Ratio, a standard probit on the selection equation (4a) is estimated for each t relying on bootstrapped standard errors (200 repetitions).

In this approach, equation (5) can be rewritten as follows:

$$5a)\ LnR\&D_{it}^{*}/(D\_R\&D_{it} = 1) = \beta_1 x_{it}^{2} + \gamma_2 \bar{x}_i^{2} + \zeta \lambda_{it}^{2} + v_{it}^{2}$$

where λ_{it} is the Invers Mill's Ratio and v^2_{it} is an orthogonal residual.

According to Wooldridge (1995), equation (5) can be estimated by including the t IMRs obtained from the selection equation for each time period along with the regressors. This method allows the error term to be correlated with the IMRs, Equation (5a) can thus be consistently estimated by pooled OLS. Following Wooldridge (2010), we calculate panel bootstrapped standard errors (200 repetitions) clustered by firm in order to obtain standard errors corrected for first stage probit estimates and robust to heteroskedasticity and serial correlation.

By applying this approach, we are able to predict the potential R&D for non-reporting firms (*lnR&D_hat*). Our model indeed is based on the assumption that all firms perform innovative activities, although some of them do not report the related R&D investments.

THE INNOVATION EQUATION

The econometric specification of the innovation equation (See equation (2)) is equation (6), where the knowledge output is measured in terms of number of patents while all terms on the right hand side enter in logarithmic form. Equation 6 is formalised as follows:

6) $NPAT_{it}^{*} = \beta_l lnR\&D_{it}^{*} + \beta_3 x_{it}^{3} + u_{it}^{3}$

Where $lnR\&D^{*}_{it}$ is our latent research variable, x^{3}_{it} is a vector of other explanatory variables, β_l and β_4 are the respective coefficients and u^{3}_{it} is the error term.

Here, the output measure is explained by a set of independent time varying variables that aim at capturing the specific relevant characteristics of the size of the internal knowledge base and its interaction with the amount of external knowledge. Also specific firms' characteristics are taken into account (see section 4.3 for their detailed specification).

As the dependent variable, i.e. *NPAT*, measuring the number of firm patent applications, is a count variable, equation (6) is estimated using count models that prove more appropriate in dealing with non-negative integers.

More precisely, equations (6) can be estimated by means of either a Poisson or a negative binomial model. Since our dependent variable is over-dispersed, as showed in Table 2 by the fact that its variance is far larger than the mean for the sampled firms, the negative binomial estimator seems to be more appropriate. However, since firms included in our sample belong to different industrial macro-sectors, they show a different patenting behaviour. For this reason a zero-inflated regression model seems appropriate to test equation (6). Zero-inflated models attempt to account for excess zeros by means of the estimation of two equations simultaneously, one for the count model and one for the excess zeros. In other words, zero-inflated models deal with two sources of over-dispersion: a qualitative part, which explains the presence or absence of patent count, and a quantitative part, which explains the positive patent count for firms having at least one patent in a given year time. Zero-inflated regression models might be a good option if there are more zeros than would be expected by either a Poisson or negative binomial model. We thus finally use a zero-inflated negative binomial regression estimator.[i] To account for the panel nature of our dataset, we cluster on firms identifiers to correct the standard errors for within cluster similar values.

THE PRODUCTIVITY EQUATION WITH ENDOGENOUS KNOWLEDGE COSTS AND IMITATION EXTERNALITIES

The econometric specification of equation (3), the productivity equation, leads to equations (7) and (8). In equation (7) the cost of knowledge is the endogenous variable. For each firm, the endogenous cost of knowledge is measured as the ratio of R&D expenditures (predicted from equations (4) and (5)) to the number of patent applications (predicted from equation (6)):

7) KCOST = R&D* / NPAT**

Knowledge costs are endogenous and specific to each observation as both the R&D (*R&D**) and the patent (*NPAT**) measures are the predicted values of the econometric estimates of the respective equations (4) and (5).

The econometric specification of the productivity equation (equation 3) is formalized by equation (8) as it follows:

8) $Y/L_{it} = \beta_l KCOST^*_{it} + \beta_4 x^4_{it} + u^4_{it}$

Here, the dependent variable is labor productivity measured as deflated value added per employee (in logarithm). x^4_{it} is a vector of explanatory variables other than the estimated including physical capital per employee, firms' size, R&D per employee and the interaction between the internal knowledge base and the amount of external knowledge. β_l is the elasticity of total factor productivity with respect to the cost of knowledge, β_4 is the

vector of coefficients for the explanatory variables and u^4_{it} is the error term.

The use of predicted innovation costs in the productivity equation instead of the predicted innovation success is a major departure from the standard CDM model. The classical CDM model tests the impact of innovation output 'given' the inputs. We extend the CDM model and account for the endogenous determinants of the costs of knowledge.

It is worth noting that the absolute amount of R&D expenditures of the firms considered is quite low and it represents an average of less than 2% of sales. Consequently, although it is true that from an 'accounting' point of view, the inclusion of both innovation output and R&D may lead to multiple counting of labour and capital used for research (in K/L, in R&D expenditure and in the innovative output), the risks to generate potentially biased estimates are negligible. In order to minimize them and to take into account the effects of the limited appropriability of knowledge and its uncontrolled leakage we include the flow of R&D expenditures instead of the stock. Knowledge spillovers limit the cumulability and reduce the time window into which the flows of R&D expenditures exert their effect internally, within the boundaries of the firm. The stock of external knowledge instead is augmented by the flows of R&D activities performed by each firm. What matters for productivity is not only the innovation output but also and primarily its cost (once they are accounted for in labour and capital input and in the innovation equation). The difference between equilibrium and actual innovation costs stemming from

knowledge externalities is the single plausible explanation for productivity growth. If knowledge were a standard economic good with high appropriability and exhaustibility, its marginal output would match its costs: there would be no relationship between innovation and productivity growth. Productivity growth stems not only from the multiplicative relationship between the internal and the external stock of knowledge but also from the internal generation of knowledge at costs that are below equilibrium levels. This is due to the full range of effects of the Arrovian limited appropriability of knowledge as an economic good of which Zvi Griliches saw the positive effects in terms of spillovers, in terms of reduced knowledge costs, rather than just the negative ones in terms of missing incentives (Antonelli, 2013; Antonelli and Gehringer, 2016).

Following the CDM approach, we could simply estimate equation (8) by ordinary least squares (OLS), with a robust covariance matrix. However, to take into account the panel nature of our data, we opted for a fixed effect estimator with a robust covariance matrix.[ii]

References

Antonelli, C. (2013). Knowledge governance: Pecuniary knowledge externalities and total factor productivity Growth. *Economic Development Quarterly*, *27*, 62–70. doi:10.1177/0891242412473178

Antonelli, C., & Gehringer, A. (2016). The cost of knowledge and productivity growth. In A. N. Link, & C. Antonelli (Eds.), *Strategic alliances. Leveraging economic growth and development*. London: Routledge.

Crépon, B., Duguet, E., & Mairesse, J. (1998). Research, innovation and productivity: An econometric analysis at the firm level. *Economics of Innovation and New Technology*, *7*, 115–158. doi:10.1080/10438599800000031

Wooldridge, M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, *68*(1), 115–132. doi:10.1016/0304-4076(94)01645-G

Wooldridge, M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

[i] We also implemented the Vuong test to compare the zero-inflated model with the negative binomial regression model. As shown in the Results section, the zero-inflated negative binomial is a better fit than the standard negative binomial in our estimations.

[ii] Equations (6) and (8) have also been estimated simultaneously by applying the 3-stages least square estimator. We obtained similar results. This confirms the robustness of our analysis.