

# R&D and market size: who benefits from orphan drug legislation?

Simona Gamba\* Laura Magazzini<sup>†</sup> Paolo Pertile<sup>‡</sup>

October 8, 2020

## Abstract

Since the early 80s, incentives have been introduced to stimulate R&D for rare diseases. We develop a theoretical model to study the impact of *push* and *pull* incentives on the intensive and extensive margin of optimal R&D investments. The model describes the mechanisms through which the type of incentives deployed may favor R&D for orphan diseases with comparatively high prevalence, increasing inequality within orphan diseases. Our empirical analysis uses data on orphan drug designations from the Food and Drug Administration. In line with the theoretical results, we find evidence supporting the idea that the incentives adopted may have contributed substantially to the widening of the gap between more and less rare diseases. Our theoretical and empirical findings together suggest that, if providing some therapeutic option to patients with very rare diseases is a priority, a revision of the current system of incentives should be considered.

**Keywords:** pharmaceutical innovation; orphan drug regulation; Gumbel distribution; market size; inequality

**JEL:** I14; I18; O31; O38; C35

## 1 Introduction

Orphan diseases are those that affect a small number of individuals, with the exact definition varying from one institutional context to another. Despite the fact that each of these diseases

---

\*Department of Economics and Finance, Università Cattolica del Sacro Cuore, Milan, Italy; e-mail: simona.gamba@unicatt.it

<sup>†</sup>Department of Economics, University of Verona, Verona, Italy. Current address: Institute of Economics, Sant'Anna School of Advanced Studies, Pisa, Italy; e-mail: laura.magazzini@santannapisa.it

<sup>‡</sup>Corresponding author. Department of Economics, University of Verona, Via Cantarane 24, 37129, Verona, Italy; e-mail: paolo.pertile@univr.it

often affects very few people, currently 7,000 orphan diseases are described in the literature, and it is estimated that 25 to 30 million US citizens and 27 to 36 million EU residents suffer from an orphan disease (Health and Safety, 2015). However, a treatment with a specific indication is available for less than 10% of known rare diseases (Melnikova, 2012; Tambuyzer, 2010).

Given that the pharmaceutical industry is mainly responsible for R&D investments for new drugs, the allocation of resources to diseases is affected by the expected return on investments. Hence, market size is a critical dimension. The empirical and theoretical analysis of the effect of market size on innovation indicates a positive correlation. Acemoglu and Linn (2004) find that a 1% increase in potential market size is associated with a 6% increase in the total number of new drugs launched in the US market. Similar results are obtained by Dubois et al. (2015), who find that R&D efforts are directed towards larger markets and estimate that, on average, additional revenues of \$2.5 billion are required to support the invention of one new chemical entity. Jobjörnsson et al. (2016) propose a theoretical model to study how the interaction between the regulation of marketing approval by institutions such as the FDA and EMA and reimbursement decisions by payers affects R&D investment, showing that R&D investments are less likely for rare diseases. Using data on marketing authorizations, Barrenho et al. (2019) find that innovation is concentrated on diseases with a larger market potential, i.e. with higher prevalence or greater willingness to pay.

In order to address the lack of incentives, policy makers have introduced a number of tools to foster R&D for orphan diseases. The main tools are tax credits on R&D expenditure, market exclusivity for new products, protocol assistance and lower fees for marketing authorization. The necessary formal step to access these incentives is to obtain an *orphan drug designation* (ODD) from the relevant regulatory authority. The first special legislation was introduced in the US, with the Orphan Drug Act (ODA) approved in 1983. Since then, several other countries have established regulations for the development of orphan drugs at different points in time. The economic rationale for these incentives is plausibly related to inequality aversion, given the huge differences in the availability of treatments between rare and common diseases. The problem can also fit an equality of opportunity framework, given that disease prevalence is clearly beyond individual control (Roemer, 1998).

Overall, there seems to be a general consensus that special regulations adopted over the world have contributed to closing the gap between orphan and non-orphan diseases. Braun et al. (2010), Lichtenberg and Waldfogel (2009) and Yin (2008) provide evidence of a positive impact of the ODA on R&D directed to orphan diseases. A positive impact on the number of designations and approvals for orphan drugs is also found in the EU (Westermarck et al., 2011).

Although most of the literature has addressed the question whether special regulations are effective in reducing the gap between R&D for orphan and non orphan diseases, far less at-

attention has been devoted to the possibly heterogeneous impact on different orphan diseases. However, this is extremely relevant, given the huge number of orphan diseases and the large variability in the level of research among them; this variability can be ascribed to differences in therapeutic class, prevalence, number of scientific publications (Heemstra et al., 2009; Pammolli et al., 2009), and population affected, e.g. diseases with onset in childhood (Raïs Ali and Tubeuf, 2019). If the motivation for the orphan drug legislation is the lack of or insufficiency of R&D investment for some diseases, an assessment of the effectiveness of the legislation should go beyond the estimate of an average impact on the class of orphan diseases, to study how additional R&D efforts are distributed across different diseases. This type of analysis is particularly relevant from the patient perspective, especially for those patients with no therapeutic option available. In the analysis by Yin (2008) on the effect of the ODA on the flow of clinical trials in the early years following the Act, differences in impact emerge when the prevalence of rare diseases is accounted for, with a smaller impact for most rare diseases.

The existence of a variety of incentives within the orphan drug legislation also raises the question which of them is most suitable to achieve the intended objectives. The main distinction is between *pull* programs (e.g., market exclusivity) and *push* programs (e.g., tax credits).

Our article contributes theoretically and empirically to the literature in three dimensions. This analysis is the first to characterize the impact of orphan regulations on inequality between diseases with different levels of prevalence over a long time-span (1983-2016). Second, we study the impact of orphan regulation on both the extensive and the intensive margin, by studying: *i*) the probability of having any investment in R&D for a certain disease, *ii*) the intensity of the R&D effort. Third, we contribute to the literature investigating the different impacts of *pull* versus *push* R&D incentives (Kremer, 2002; Rietzke and Chen, 2020). We show theoretically that in our framework both types of incentive have an unambiguously stronger effect on the extensive margin of investment for less rare diseases, meaning that the impact on the probability of obtaining any investment is larger for less rare orphan diseases. In terms of investment intensity, it is not possible to conclude unambiguously whether more or less rare diseases benefit more from the incentives.

We extend the theory-based distinction between the intensive and the extensive margin of R&D to the empirical analysis by means of a zero-inflated count data model, where the dependent variable is the yearly number of ODDs granted by the FDA at the disease level, as a proxy for R&D effort. For the sake of consistency with the theoretical model, the excess of zeros is modeled using the Gumbel distribution, to replace the standard Logit or Probit model. Our empirical approach exploits the fact that different countries (or geographical regions) have introduced the reforms at different points in time and that, according to our theoretical results, diseases with different prevalence may have benefited differently from the regulations.

We find that, within the class of orphan diseases, R&D efforts have increased over time substantially more for less rare diseases, thus increasing inequality. According to our baseline specification, the difference between the predicted number of orphan designations per year for a disease in the highest and the lowest class of prevalence was 5.6 times larger after 2008 than in the period 1983-1992. The main conclusion remains valid even when controlling for a number of other factors potentially affecting the relative convenience of investing in less rather than more rare diseases. To the best of our knowledge, no evidence of this dynamic has previously been reported. Based on our theoretical analysis and the results of model calibration, we argue that the way in which orphan incentives were designed may have contributed to widening this gap. By relying almost exclusively on *pull* incentives, European legislation may have exacerbated this tendency.

In terms of policy implications, our results suggest that, if inequality aversion is a fundamental motivation for orphan legislation, then a revision of the incentive tool-kit should be considered, with the objective of curbing the widening of the gap between less and more rare orphan diseases. One way of mitigating this tendency could be to shift the balance of incentives towards *push* tools. A more radical reform might be to consider abandoning the idea of setting an arbitrary threshold of prevalence, below which all diseases benefit from the same type of incentives, and to move towards prevalence-dependent incentives.

The structure of the article is as follows. Section 2 describes the various regulations that have been adopted over time. Section 3 describes the model, which is solved in Section 4. Section 5 presents the main results of a model calibration. Sections 6 and 7 describe, respectively, data and methodology for the empirical analysis, with results presented in Section 8. Section 9 concludes and discusses policy implications.

## 2 Institutional framework

Over the last 35 years, orphan drug regulations have been adopted in several countries around the world. The US was the first country to develop a specific legislation. In 1983, Congress signed the ODA, according to which a drug is considered *orphan* if it treats a rare disease or condition affecting fewer than 200,000 persons in the US (about 6.25 in 10 thousand persons) or if it is not expected to be profitable within seven years following approval by the FDA.<sup>1</sup> The incentives for drugs designated as orphan are (1) assistance from the Office of Orphan Product Development during the development process; (2) tax credits (up to 50% of clinical development costs); (3) exemption or waiver of application (filing) fees; (4) seven years of marketing

---

<sup>1</sup>Originally, the ODA defined as orphan any disease or condition occurring “infrequently” in the United States without reasonable expectations on profits. The Health Promotion and Disease Prevention Amendments of 1984 specified the prevalence requirement (Herder, 2017).

exclusivity and (5) subsidies for clinical trials from the Orphan Products Grant Program. Although multiple orphan designations may be granted for a particular disease (Gibson and von Tigerstrom, 2015), for 7 years no marketing approval will be given to a subsequent sponsor of a drug containing the same active moiety or principal molecular features as a previously approved drug intended for the same therapeutic indication unless this can be shown to be clinically superior.

Special regulations with the same objectives were subsequently introduced in several countries, such as Singapore (1991), Japan (1993), Australia (1998), South Korea (1998), the EU (2000) and Taiwan (2000). Among these, we focus on those applied in the areas with the largest markets: Japan and the EU.

In April 1993, Japan substantially revised its orphan medicinal product system, introduced in 1985, in order to extend the tools used to incentivize research into orphan diseases. So, in addition to the existing (1) reductions in the data required for applications, and (2) the accelerated review process, the following incentives were introduced: (3) protocol assistance; (4) tax credits (up to 6% of clinical and non-clinical costs); (5) subsidies for clinical and nonclinical studies and (6) ten years of market exclusivity. In order to be designated as orphan a drug has to treat a rare and serious disease or condition affecting less than 50,000 persons in Japan (about 4 in 10 thousand persons); no appropriate alternative treatment should be available on the market or the expected efficacy and safety must be higher compared with existing products. As in Japan the incentives which are the main focus of our analysis were introduced in 1993, we refer to this as the date when the special legislation was introduced.

In December 1999, the EU also introduced specific incentives for the development of orphan medicinal products through Regulation (EC) No 141/2000. The incentives include (1) protocol assistance; (2) access to a centralized procedure allowing immediate marketing authorization in all member states; (3) reduced fees for regulatory procedures and (4) ten years of market exclusivity. In order to benefit from the incentives, orphan drugs have to be designated as such before receiving marketing authorization. Moreover, when the application is made, the drug must treat a condition affecting no more than 5 in 10 thousand persons in the Community, or a life threatening or chronically debilitating condition for which it is unlikely, without incentives, that the marketing of the medicinal product in the Community would generate sufficient return to justify the necessary investment;<sup>2</sup> finally, there should be no satisfactory alternative methods authorized in the Community or, if such method exists, the medicinal product must be expected to bring significant benefit to those affected by that condition (article 3 of the Regulation). In addition to the incentives mentioned in the regulation, France and the Netherlands provide tax credits (Health and Safety, 2015).

---

<sup>2</sup>According to Tambuyzer (2010), more than 99.5% of orphan designations in the EU are granted because of the prevalence criteria.

	US (1983)	Japan (1993)	EU (2000)
<b>Orphan disease:</b>	< 200,000 in US (6.25/10,000) or not profitable	< 50,000 in Japan (4/10,000)	< 5 in 10,000 in EU or both not profitable & life-threatening
<b>Main incentives:</b>			
Tax credit	Yes (50% clinical costs)	Yes (6% clinical and non-clinical costs)	Member state specific
Market exclusivity	Yes (7 years)	Yes (10 years)	Yes (10 years)
Reduced applic. fees	Yes (waved)	No	Yes (reduced)
Protocol assist.	Yes	Yes	Yes
Subsidies for clinical trials	Yes	Yes	No

Table 1: Comparison of orphan drugs regulations in the US, Japan and EU.

The incentives provided by the US, Japan and the EU are summarized in Table 1, together with the requirements for drugs to be considered orphans.

Since November 2007, the European Medicines Agency (EMA) and the FDA have been collaborating to encourage joint applications for orphan drug status both in the EU and the US. A shared application form has been developed, in an effort to reduce the administrative burden on the orphan drug sponsor (Braun et al., 2010; Mariz et al., 2016). This reduced the cost of being eligible for incentives in both geographic areas. Parallel applications in Japan and the EU are also encouraged, although a shared application form is not yet in place (Mariz et al., 2016).

### 3 The model

Let  $N^f$  firms be free to decide on the size of an R&D investment,  $I \geq 0$ , targeting disease  $j$ , which affects  $n_j$  individuals. For an orphan drug, there are two key regulatory steps in the development process. In the first step, the firm that developed the molecule applies for an ODD. If granted, the ODD makes the firm eligible for incentives related to the development of the orphan drug. If the development process is successfully completed, the firm goes on to the second regulatory stage: marketing authorization.<sup>3</sup> From the perspective of the firm, both stages entail uncertainty. Let  $p_j^d(I)$  be the probability that the firm obtains an ODD, given the R&D

<sup>3</sup>In principle, it is also possible that a drug reaches the market without having previously obtained an ODD. However, this is the case for only 2% of US marketing authorizations in our dataset. For the sake of simplicity, we assume that only drugs with an ODD receive marketing authorization.

investment  $I$ . For the function  $p_j^d(I)$  we assume that  $\frac{\partial p_j^d}{\partial I} > 0$ ,  $\frac{\partial^2 p_j^d}{\partial I^2} < 0$  and  $\lim_{I \rightarrow 0} \frac{\partial p_j^d}{\partial I} = +\infty$ . Moreover,  $p_j^d(0) = 0$  and  $\lim_{I \rightarrow \infty} p_j^d(I) = 1$ .

Conditional on obtaining an ODD, the firm carries on with the development process. With probability  $p_j^m$ , assumed to be independent of  $I$ , this leads to the marketing approval of the product. Given disease specific per patient net revenue  $m_j$ , conditional on obtaining an ODD, the per patient expected net revenue is  $p_j^m m_j$ . To simplify the notation, we define  $M_j(\Omega_j) = p_j^m m_j$ . The parameter  $\Omega_j$  is a vector of disease specific characteristics that may affect the probability  $p_j^m$  and/or the net revenue  $m_j$ . For example, some regulators grant a price premium to drugs targeting life threatening conditions.

The expected profit for firm  $i$  ( $i = 1, 2, \dots, N^f$ ) investing  $I$  to develop a drug for disease  $j$  is:

$$E\Pi_{ij} = p_j^d(I)[M_j n_j] - I + \delta_{ij}. \quad (1)$$

The term  $\delta_{ij}$  is an idiosyncratic component aiming at capturing any additional positive or negative component of the expected profit known only to the firm. This may be related, for example, to the impact on R&D costs of other projects undertaken by the firm, simultaneously or in the past. It is assumed that  $\delta_{ij}$  is known to the firm before deciding the investment strategy. From the perspective of the researcher,  $\delta_{ij}$  is the realization of a random variable, with density  $f(\Delta)$ . According to Eq. 1, a new drug that obtains market authorization captures the entire market. This simplifying assumption is justified by the fact that, in our data, only for a small fraction of orphan diseases (6%) is more than one treatment authorized over the entire time span. If we could look at a single point in time, this proportion would be reduced further. We also introduce the simplifying assumption that firms make their investment decisions independently. However, in Appendix A we relax the assumption that a new drug captures the entire market and show that the quality of the relevant results is unchanged.

The aim of our analysis is to study the impact of different forms of incentives among those introduced as part of the special legislation on: *i*) the probability of obtaining an investment in a rare disease, *ii*) the probability of obtaining an orphan designation. Our analysis is carried out *within* the class of orphan diseases. In other words, we do not contrast rare versus non-rare diseases, but more versus less rare diseases within the class of orphan diseases. As a result, we assume that all diseases are eligible for incentives. Our focus is on how the impact of different types of incentives is affected by the prevalence of an orphan disease.

Our comparison of alternative incentives focuses on the typical distinction between *pull* and *push* programs. *Pull* incentives are those that aim to increase the net market revenue of investments made in orphan diseases. The best known instance of such an instrument is market exclusivity, to which all products with orphan designation are entitled. We model this as a markup,  $z$  ( $z \geq 0$ ), on net revenues. This way of modeling *pull* incentives is sufficiently flexible to

account for other types of incentives, such as a price premium to which all orphan drugs are equally entitled.

*Push* incentives reduce the cost of R&D investment in rare diseases. Examples of such incentives include tax credits, reduced application fees for market authorization and protocol assistance. We model this type of incentive as an allowance on investment costs, such that, conditional on obtaining an ODD, the investment cost borne by the firm is  $I(1 - \gamma)$ , with  $0 \leq \gamma \leq 1$ .

To take the role of these incentives into account, the expected profit function can be written as:

$$E\Pi_{ij} = p_j^d(I)[M_j n_j](1 + z) - (1 - p_j^d(I)\gamma)I + \delta_{ij}. \quad (2)$$

Eq. 2 shows that access to *pull* incentives is conditional upon obtaining an ODD and reaching the market (recalling  $M_j = p_j^m m_j$ ), whereas obtaining an ODD is sufficient for eligibility for *push* incentives.

## 4 Optimal investment policy

We start by characterizing optimal decisions from the perspective of a single firm and then move to the analysis of the outcome of these decisions at the market (disease) level.

### 4.1 The firm's decisions

The firm aims to maximize the expected profit in Eq. 2 with respect to  $I$ . The first order condition can be written as:

$$\frac{\partial p_j^d(I)}{\partial I}[M_j n_j(1 + z)] + \gamma \left( \frac{\partial p_j^d(I)}{\partial I} I + p_j^d(I) \right) = 1. \quad (3)$$

The first order condition in Eq. 3 requires equating the marginal cost of investment on the right hand side to the marginal benefit, which is made up of two components: *i*) the increased probability to reach the market and its revenues, *ii*) the positive impact on the expected revenue provided by the *push* incentive. The assumption that  $\lim_{I \rightarrow 0} \frac{\partial p_j^d}{\partial I} = +\infty$  ensures the existence of a strictly positive value of  $I$  that solves the equation. We assume the marginal benefit to be strictly decreasing in  $I$ , meaning that Eq. 3 defines a unique optimum.<sup>4</sup> Eq. 3 also highlights the well known role of market size as an incentive for R&D investments: a reduction in  $n_j$  reduces the marginal benefit of investment and leads to a lower optimal level of investment. According

---

<sup>4</sup>Note that this condition is directly implied by the assumption on the strict concavity of  $p_j^d(I)$ , for sufficiently small values of  $\gamma$ .



to Eq. 3, the optimal investment level ( $I_j^*$ ) depends only on characteristics at the disease level, but not on  $\delta_{ij}$ .

#### 4.1.1 Impact of *pull* incentives

We can use the implicit function theorem to study the impact of an increase in  $z$  on the optimal level of investment:

$$\frac{dI_j^*}{dz} = -\frac{\frac{\partial p_j^d(I)}{\partial I} M_j n_j}{\frac{\partial^2 p_j^d(I)}{\partial I^2} [M_j n_j (1+z) + \gamma I] + 2\gamma \frac{\partial p_j^d(I)}{\partial I}} > 0. \quad (4)$$

From the perspective of our analysis, it is particularly interesting to investigate how the marginal impact on  $I_j^*$  of an increase in  $z$  varies with  $n_j$ . Differentiating Eq. 4 with respect to  $n_j$  obtains:

$$\begin{aligned} \frac{\partial^2 I_j^*}{\partial z \partial n_j} = & \frac{\frac{\partial p_j^d(I)}{\partial I} M_j n_j \left[ \frac{\partial^3 p_j^d(I)}{\partial I^3} \frac{\partial I_j^*}{\partial n} [M_j n_j (1+z) + \gamma I_j^*] + \frac{\partial^2 p_j^d(I)}{\partial I^2} \left( M_j (1+z) + 2\gamma \frac{\partial I_j^*}{\partial n} \right) \right]}{\left[ \frac{\partial^2 p_j^d(I)}{\partial I^2} [M_j n_j (1+z) + \gamma I_j^*] + 2\gamma \frac{\partial p_j^d(I)}{\partial I} \right]^2} + \\ & - \frac{M_j \left( \frac{\partial^2 p_j^d(I)}{\partial I^2} \frac{\partial I_j^*}{\partial n} n + \frac{\partial p_j^d(I)}{\partial I} \right)}{\frac{\partial^2 p_j^d(I)}{\partial I^2} [M_j n_j (1+z) + \gamma I_j^*] + 2\gamma \frac{\partial p_j^d(I)}{\partial I}}. \end{aligned} \quad (5)$$

Eq. 5 cannot be unambiguously signed, meaning that the size of the impact of a *pull* incentive on  $I_j^*$  may be increasing or decreasing in  $n_j$ . The ambiguity is due to the existence of two impacts going in opposite directions. There is a positive impact because, for a given level of  $I$ , the increase in expected revenues due to an increase in  $z$  is proportional to the market size. However, given  $z$ ,  $I_j^*$  is greater when  $n_j$  is larger. As the marginal productivity of an increase in  $I$ , in terms of increased probability of obtaining an ODD, is decreasing, this impact goes in the opposite direction.

Given  $I_j^*(M_j, n_j)$ , the firm invests only if the expected profit at the time of investment is non-negative, i.e.:

$$p_j^d(I_j^*) [M_j n_j] (1+z) - (1 - p_j^d(I_j^*) \gamma) I_j^* + \delta_{ij} \geq 0. \quad (6)$$

It is then possible to define a minimum value of  $\delta_{ij}$ ,  $\hat{\delta}_j$ , such that the firm makes any investment in R&D for disease  $j$ :

$$\hat{\delta}_j = (1 - p_j^d(I_j^*) \gamma) I_j^* - p_j^d(I_j^*) [M_j n_j] (1+z). \quad (7)$$

To investigate the impact of  $n_j$  on the decision whether to invest or not, we study the depen-

dency of  $\hat{\delta}_j$  on  $n_j$ . Observing that

$$\hat{\delta}_j = -E\Pi_{ij}(I_j^*) + \delta_{ij}, \quad (8)$$

simplifies calculations through the application of the envelope theorem, so:

$$\frac{\partial \hat{\delta}_j}{\partial n_j} = -p_j^d(I_j^*)M_j(1+z) < 0. \quad (9)$$

Hence, other things being equal, for a comparatively rare disease the value of  $\delta_{ij}$  must be larger for the firm to decide to undertake any investment (Eq. 9). Thus, it is less likely to observe R&D investment in comparatively rare diseases.

Using a similar approach, we can study the impact of an increase in  $z$  on  $\hat{\delta}_j$ . This leads to

$$\frac{\partial \hat{\delta}_j}{\partial z} = -p_j^d(I_j^*)M_j n_j < 0, \quad (10)$$

which shows the role of  $z$  in making it more likely that there is investment for disease  $j$ , by reducing the value of  $\hat{\delta}_j$ . Also in this case, we are interested in the heterogeneous impact of this incentive tool across different classes of prevalence. By differentiating Eq. 10 with respect to  $n_j$ , we obtain:

$$\frac{\partial^2 \hat{\delta}_j}{\partial z \partial n_j} = -M_j \left[ \frac{\partial p_j^d}{\partial I} \frac{\partial I_j^*}{\partial n_j} n_j + p_j^d(I_j^*) \right] < 0. \quad (11)$$

The negative sign of the expression means that the impact on the probability that the firm undertakes an investment of an increase in  $z$  is larger for less rare diseases.

#### 4.1.2 Impact of *push* incentives

As for  $z$ , the impact on  $I_j^*$  of an increase in  $\gamma$  is positive (see Eq. 3). Concerning the heterogeneity of the impact, using the same approach as above, we find that:

$$\begin{aligned} \frac{\partial^2 I_j^*}{\partial \gamma \partial n_j} = & \frac{\left( \frac{\partial p_j^d(I)}{\partial I} I_j^* + p_j^d \right) \left[ \frac{\partial^3 p_j^d(I)}{\partial I^3} \frac{\partial I_j^*}{\partial n} [M_j n_j (1+z) + \gamma I] + \frac{\partial^2 p_j^d(I)}{\partial I^2} \left( M_j (1+z) + \gamma \frac{\partial I_j^*}{\partial n} \right) + 2\gamma \frac{\partial^2 p_j^d(I)}{\partial I^2} \frac{\partial I_j^*}{\partial n} \right]}{\left[ \frac{\partial^2 p_j^d(I)}{\partial I^2} [M_j n_j (1+z) + \gamma I] + 2\gamma \frac{\partial p_j^d(I)}{\partial I} \right]^2} + \\ & - \frac{\frac{\partial^2 p_j^d(I)}{\partial I^2} \frac{\partial I_j^*}{\partial n} I_j^* + 2 \frac{\partial p_j^d(I)}{\partial I} \frac{\partial I_j^*}{\partial n}}{\frac{\partial^2 p_j^d(I)}{\partial I^2} [M_j n_j (1+z) + \gamma I] + 2\gamma \frac{\partial p_j^d(I)}{\partial I}}. \end{aligned} \quad (12)$$

As for  $z$ , this term cannot be unambiguously signed, because two effects operate in opposite directions. Other things being equal, with a larger value of  $n_j$ ,  $I_j^*$  is larger and so is the gain

from an increase in  $\gamma$ . However, a higher value of  $I_j^*$  also implies a lower marginal productivity of additional investment, in terms of increased probability of obtaining an ODD. Therefore, also the impact of an increase in  $\gamma$  on  $I_j^*$  may increase or decrease in  $n_j$ .

Concerning the impact on the minimum value of the idiosyncratic term that makes an investment in disease  $j$  profitable, we obtain:

$$\frac{\partial \hat{\delta}_j}{\partial \gamma} = -p_j^d(I_j^*) I_j^* < 0 \quad (13)$$

and

$$\frac{\partial^2 \hat{\delta}_j}{\partial \gamma \partial n_j} = -\frac{\partial I_j^*}{\partial n_j} \left( \frac{\partial p_j^d(I)}{\partial I} I_j^* + p_j^d \right) < 0. \quad (14)$$

In this case too, the impact is greater for less rare diseases.

The following proposition states an important difference between a *pull* and *push* incentive, based on the comparison between Eq. 11 and Eq. 14:

**Proposition 1.** *For both types of incentives, the reduction in  $\hat{\delta}_j$  is greater for less rare diseases. However, while for a push incentive this is due only to an indirect effect, i.e. through the impact on  $I_j^*$ , for a pull incentive there is both a direct and an indirect effect.*

## 4.2 Market outcomes

We can now move on to study the impact of incentives at the disease level, assuming that the  $N^f$  firms make independent investment decisions, as characterized in the previous subsection. We focus on two outcomes:

1. the probability that at least one firm makes an R&D investment targeting disease  $j$ ;
2. the expected number of ODDs for disease  $j$ .

Starting with the first outcome of interest, investment by at least one firm occurs if

$$\max_i \{\delta_{ij}\} > \hat{\delta}_j. \quad (15)$$

For the most common types of distributions  $f(\Delta)$ , including the normal and the exponential, the Gumbel distribution is the limiting distribution of  $\max_i \{\delta_{ij}\}$  (Ahsanullah, 2016). We denote by  $f^G(\tilde{\delta})$  and  $F^G(\tilde{\delta})$  the probability density function and the cumulative density function (CDF) of  $\max_i \{\delta_{ij}\}$ , respectively. The indicator function  $\mathcal{I}_j^I$  can be used to define whether at least one firm invests in disease  $j$  ( $\mathcal{I}_j^I = 1$ ) or not ( $\mathcal{I}_j^I = 0$ ). We obtain

$$\mathcal{P}(\mathcal{I}_j^I = 1) = 1 - \int_{-\infty}^{\hat{\delta}_j} f^G(\tilde{\delta}) d\tilde{\delta}. \quad (16)$$

Following the analysis in the previous subsection, the focus is on how the impact of incentives changes with prevalence, i.e.:

$$\frac{\partial^2 \mathcal{P}(\mathcal{I}_j^I = 1)}{\partial z \partial n_j} = - \left[ \frac{\partial^2 F^G(\tilde{\delta})}{\partial \tilde{\delta}^2} \frac{\partial \hat{\delta}_j}{\partial n} \frac{\partial \hat{\delta}_j}{\partial z} + \frac{\partial F^G(\tilde{\delta})}{\partial \tilde{\delta}} \frac{\partial^2 \hat{\delta}_j}{\partial z \partial n} \right] \quad (17)$$

and

$$\frac{\partial^2 \mathcal{P}(\mathcal{I}_j^I = 1)}{\partial \gamma \partial n_j} = - \left[ \frac{\partial^2 F^G(\tilde{\delta})}{\partial \tilde{\delta}^2} \frac{\partial \hat{\delta}_j}{\partial n} \frac{\partial \hat{\delta}_j}{\partial \gamma} + \frac{\partial F^G(\tilde{\delta})}{\partial \tilde{\delta}} \frac{\partial^2 \hat{\delta}_j}{\partial \gamma \partial n} \right]. \quad (18)$$

Given the results of the previous subsection, the sign of the second term in brackets is negative for both expressions. As the derivatives of  $\hat{\delta}_j$  with respect to  $n$ ,  $z$  and  $\gamma$  are also negative, the following proposition holds.

**Proposition 2.**  $\frac{\partial^2 F^G(\tilde{\delta})}{\partial \tilde{\delta}^2} \leq 0$  is a sufficient condition for both a pull and a push incentive to increase the probability of having investment in disease  $j$  more for less rare diseases.

According to Eq. 17 and 18 the condition is not necessary, because if unsatisfied, the two terms in brackets have opposite signs. However, we argue that the condition is very likely to be satisfied for our market, because the relevant part of  $F^G(\tilde{\delta})$  is likely to be concave. Investment in disease  $j$  occurs if  $\tilde{\delta}_j \geq \hat{\delta}_j$ . Given that most orphan diseases attract no investments,  $\hat{\delta}_j$  is likely to be comparatively high, meaning that the relevant part of the Gumbel distribution is in the right tail, i.e. where the CDF is concave.

We can now move on to study the impact of incentives on the expected number of ODDs, conditional on  $\mathcal{I}_j^I = 1$ . Let  $\tilde{N}^f(\hat{\delta}_j)$  be the number of firms that decide to invest in  $j$ , because  $\tilde{\delta}_{ij} > \hat{\delta}_j$ . For each of these firms, the investment decision has a Bernoulli outcome, with probability of obtaining an ODD equal to  $p_j^d(I_j^*)$ . From Eq. 3, the optimal investment level, and hence the probability of success, is the same for all firms for which it is convenient to invest in disease  $j$ . The sum of  $\tilde{N}^f(\hat{\delta}_j)$  independent random variables with Bernoulli distribution has a *Binomial*( $\tilde{N}^f(\hat{\delta}_j), p_j^d(I_j^*)$ ) distribution, for which the limiting distribution is *Poisson*. If we take this approximation, the number of ODDs, conditional on investment, is distributed *Poisson*, with parameter  $\lambda_j = \tilde{N}^f(\hat{\delta}_j) \cdot p_j^d(I_j^*)$ .

The following proposition summarizes the results of the theoretical analysis of the impact of incentives on the expected number of ODDs, conditional on  $\mathcal{I}_j^I = 1$ , for different classes of prevalence.

**Proposition 3.** *Conditional on at least one firm investing in disease  $j$ , the impact of incentives on the expected number of orphan designations may be greater or lower for less rare diseases.*

The ambiguity of this impact follows from the fact that the expected number of designations is  $\lambda_j = \tilde{N}^f(\hat{\delta}_j) p_j^d(I_j^*)$ . The impact on the probability that at least one firm invests in market  $j$  has

	$\mathcal{P}(\mathcal{I}_j^I = 0)$	# ODD   $\mathcal{I}_j^I = 1$
$\uparrow z$	Negative; impact size increasing in $n_j$	Positive; role of $n_j$ ambiguous
$\uparrow \gamma$	Negative; impact size increasing in $n_j$	Positive; role of $n_j$ ambiguous

Table 2: Summary of theoretical results: impact of an increase in the policy parameters.

been shown to be greater for less rare diseases. However, this does not necessarily imply that the impact on  $\tilde{N}^f(\hat{\delta}_j)$  is also greater, as this depends on the distribution of  $\delta_{ij}$ . Moreover, the impact on  $I_j^*$  is also ambiguous (Eq. 5 and Eq. 12).

Table 2 summarizes our main theoretical results. For consistency with the empirical analysis that follows, the table refers to  $\mathcal{P}(\mathcal{I}_j^I = 0)$ .

## 5 Calibration

This section presents the main results of a calibration of the theoretical model.<sup>5</sup> The aim is to explore the difference in the impact of *push* versus *pull* incentives, with an emphasis on whether the advantage for less rare diseases is larger with one type of incentive or the other. The theoretical prediction for this comparison is ambiguous, due to inconclusive results related to the intensive margin. Nonetheless, in terms of policy implications, being able to compare the two types of incentives is important, because it might inform on whether a different mix of incentives can be used to influence the distribution of R&D efforts between less and more rare diseases.

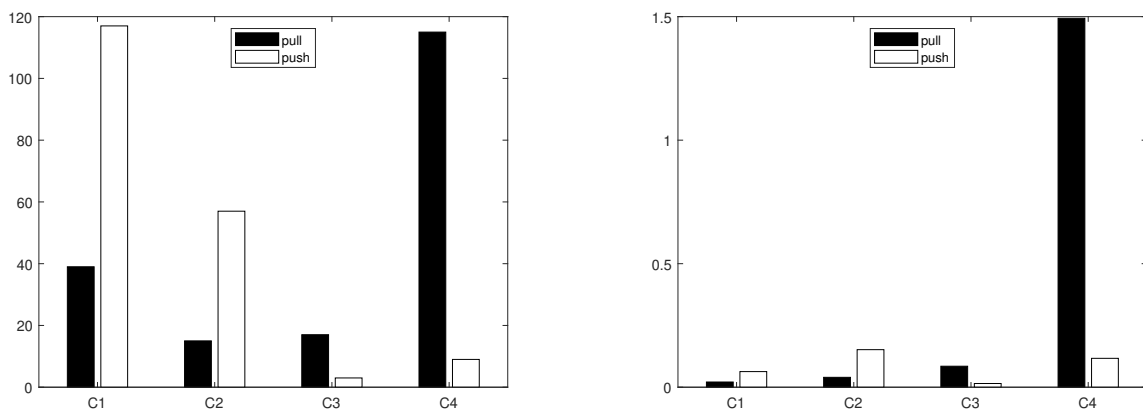
To operationalize the model, the probability function of obtaining an ODD is defined as  $p_j^d(I) = \sqrt{\frac{2 \arctan(\epsilon I)}{\pi}}$ , where  $\epsilon$  is a parameter. Disease prevalence and the idiosyncratic term ( $\delta_{ij}$ ) are assumed to have, respectively, a Log-Normal and a Normal distribution. Details of the calibration are presented in Appendix B.

The exercise undertaken with the calibrated model is based on three steps:

1. solution of the model with no incentive;
2. solution of the model with *pull* incentive ( $z$ ) only;
3. iteration of the solution with *push* incentive ( $\gamma$ ) only, to find the level of incentive such that the total number of ODDs gained is the same as in point 2.

For step 2, we set  $z = 0.2$ . Given the other parameter values, the iterated solution of the model shows that the condition defined in point 3 is satisfied with  $\gamma = 0.745$ .

<sup>5</sup>The calibration is implemented using the software MATLAB<sup>®</sup>.



(a) Additional ODD per class of prevalence ( $z = 0.2, \gamma = 0.745$ ). (b) Average number of additional ODD per disease by class of prevalence ( $z = 0.2, \gamma = 0.745$ ).

Figure 1: Distribution of additional ODD due to the incentives.

Figure 1(a) shows the number of additional ODD for each class of prevalence for the two types of incentive. By construction (step 3 of the calibration), the sum of the height of the bars for each type of incentive over the classes of prevalence is the same. More informative from the policy perspective, however, is the number of additional ODDs per disease (Figure 1(b)), which is narrowly related to the probability that a patient suffering from a certain disease may benefit from new therapeutic options in the future. Figure 1(b) shows that there is a clear difference in the distribution of additional R&D effort generated by *pull* and *push* incentives. With the *push* incentive, the increase in the number of ODDs per disease is greater in the two classes with lower prevalence, whereas there is a huge advantage for the class of less rare diseases when the incentive is *pull*. The calibration results suggest that the greater advantage offered by *pull* incentives to less rare diseases that our theoretical analysis showed for the extensive margin (Proposition 1) still exists when the overall investment effort, including both the intensive and the extensive margin, is considered. Additional output of the calibration exercise is presented in Appendix B.

## 6 Data and measures

The first step in our analysis is the identification of the full list of orphan diseases, i.e. those for which a drug is eligible to obtain an ODD. For this purpose, we rely on the Orphanet database (INSERM, 1999), which is the standard reference for information on rare diseases. The list used in the empirical analysis was downloaded in October 2017. After excluding some items, following the criteria detailed in Appendix C, our list includes 5,132 diseases. Orphanet provides

information on the class of prevalence of the disease, if already documented:<sup>6</sup> “<1/1,000,000”, “1-9/1,000,000”, “1-9/100,000”, and “1-5/10,000”.<sup>7</sup> We refer to worldwide prevalence; where this information is missing, we consider prevalence in Europe or, failing that, in the US. From Orphanet additional information was retrieved at the disease level, including the therapeutic class(es) of each disease, the age of onset and age at death (however, the latter is available only for 28% of diseases). Ages are reported as antenatal, neonatal, infancy, childhood, adolescence, adulthood and elderly.

In order to gather information on the existing knowledge about each disease, a search was conducted on PubMed<sup>8</sup> to retrieve the number of articles published over the period 1970-2016 which contain the name of each disease in the title, abstract or content. This information was used to construct a measure for the stock of publications ( $SP$ ), following the perpetual inventory method:

$$SP_{jt} = P_{jt} + (1 - \rho)SP_{j,t-1},$$

where  $P_{jt}$  is the number of publications related to disease  $j$  at time  $t$  and  $\rho = 0.1$  is the rate of obsolescence of knowledge, as generally applied in the empirical literature (Keller, 2002).

The list of rare diseases is systematically updated, as approximately 250 new diseases are described each year (Westermarck et al., 2011; Wästfelt et al., 2006). Hence, the list of known diseases in October 2017 (the basis of our analysis) may include pathologies which were unknown earlier. A lack of ODD for a disease that has not yet been discovered cannot be interpreted as a lack of R&D targeting that disease. To account for this, we include in our baseline analysis disease  $j$  only if its stock of publications in  $t - 5$  is positive (i.e.,  $SP_{j,t-5} > 0$ ).<sup>9</sup> The information on publications is also used as a control variable to proxy the level of scientific information available on the disease.

A key decision concerns how to measure R&D efforts at the disease level. Our proxy of R&D efforts targeting rare diseases is the flow (number) of ODDs granted by the FDA per year between 1983 and 2016. An ODD represents the “successful translation of rare disease research into an orphan drug discovery and development program” (Heemstra et al., 2009). Obtaining an ODD is a necessary condition for the project, and eventually for the drug, to be eligible for the incentives. In comparison with proxies of R&D used in previous contributions, such as the number of clinical trials (see, for example, Yin, 2008), ODDs have the advantage of providing information for a long time span from a single administrative source. A potential limitation

---

<sup>6</sup>In a few cases (6.7% of diseases), a numeric value for prevalence is also provided. However, the availability of this information is unevenly distributed for classes of prevalence. Given these limitations, the point estimate of prevalence is not used in the empirical analysis.

<sup>7</sup>Information on prevalence refers to year 2017 and we are unable to track moves from one class to the other. However, they are very unlikely, given the width of the classes considered.

<sup>8</sup>See <https://www.ncbi.nlm.nih.gov/pubmed/>.

<sup>9</sup>Robustness of the results to this criteria is explored in Appendix D.

is that ODDs are only available for orphan diseases, meaning that they are not suitable for an analysis of the impact of the orphan legislation on orphan versus non orphan diseases. Given that we focus on the heterogeneous impact *within* the class of orphan diseases, this concern is less relevant.

We focus on ODDs granted in the US, because the ODA establishment in 1983 enables the dynamics in the number of designations over the selected time span to be studied, including 1993, when Japan substantially strengthened its orphan provisions, and 2000, when orphan legislation was introduced in the EU. As pharmaceuticals are a global industry, it is convenient for the inventor to apply for orphan drug status in several countries to benefit from additional incentives. Together with the size of the US market, this implies that FDA data provide a reliable picture of global R&D activity. For each drug, the FDA provides the date of orphan designation, marketing approval (if any), the designated indication, and the company sponsoring the request.

Great effort was made to match the indications of the FDA list of ODDs with the Orphanet list of diseases. Further details on assignment criteria and exclusion of some of the 3,996 ODDs granted by the FDA between 1983 to 2016 are provided in Appendix C.

All in all, our data comprise 136,036 observations (5,132 diseases over – at most – 34 years). The distribution of diseases included in the analysis among prevalence classes is shown in Table 3. Information on the prevalence is missing (or not yet documented) in Orphanet for a large share of the diseases: these are considered as a separate class. Among the classes with known prevalence, a significant majority of diseases are classified with a prevalence lower than 1 in 1 million (36.89%), with only 2.98% in the class 1-5/10,000. Table 3 also shows how the average yearly number of ODDs per disease changes from one class of prevalence to another. The reported numbers of ODDs are calculated taking the average over the years of the study period and for the diseases in each class of prevalence.

Figure 2 compares the yearly number of ODDs (solid line; year 2000 = 100) with the number of investigational new drugs (IND) applications for orphan and non-orphan diseases in the US (source: FDA; dashed line; year 2000 = 100).<sup>10</sup> Although the comparison may not be fully homogeneous, it clearly suggests that the increase in R&D effort was much greater for orphan diseases. The existing literature shows that the availability of incentives for orphan drugs, which accumulated over time, was a key determinant of the sharp increase in the number of ODDs (Yin, 2008; Westermark et al., 2011; Braun et al., 2010). Figure 3 shows the evolution in the average number of new ODDs per disease for each class of prevalence for selected years. The graph shows that the growth documented by Figure 2 is mainly driven by ODDs for less rare diseases.

---

<sup>10</sup>IND application is the regulatory step required to start clinical trials on humans in the US, for all drugs, including orphan drugs. Note that this measure would not be suitable to replace ODDs in our analysis, as the data available do not allow to link them to therapeutic indications.



Prevalence	number of diseases	% total	avg. number of ODDs per disease (yearly)
<i>C1</i> : <1/1,000,000	1,893	36.89	0.03
<i>C2</i> : 1-9/1,000,000	205	3.99	0.13
<i>C3</i> : 1-9/100,000	302	5.88	0.17
<i>C4</i> : 1-5/10,000	153	2.98	0.22
<i>C0</i> : Missing prev.	2,579	50.25	0.13
Total	5,132	100	–

Table 3: Distribution of the diseases among prevalence classes.

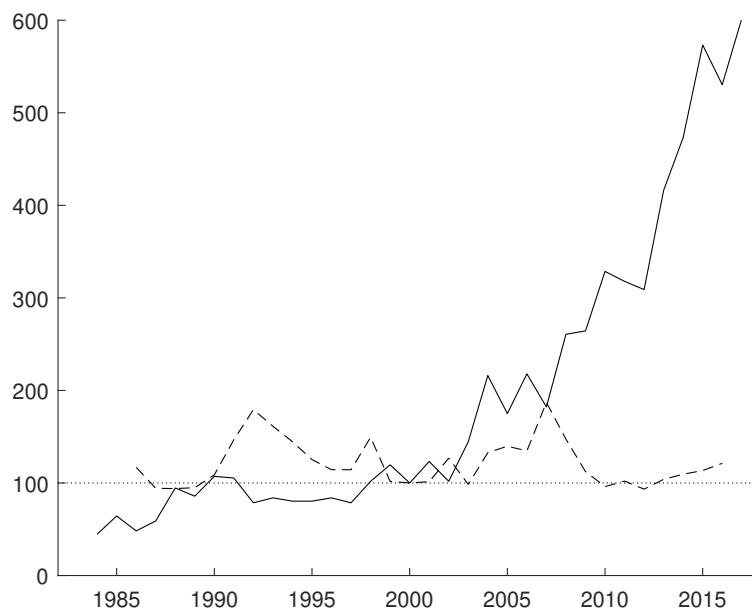


Figure 2: Yearly number of ODDs (continuous line) and IND applications (dashed line); year 2000 = 100.

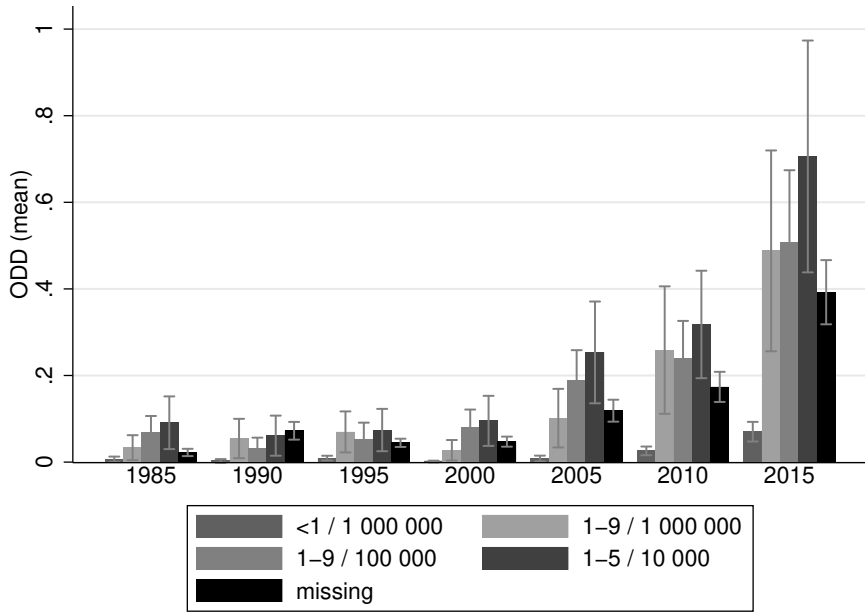


Figure 3: Average number of new ODDs per disease among prevalence classes for selected years.

## 7 Empirical methods

Our theoretical analysis considers *i*) the probability of no R&D for a certain disease ( $\mathcal{I}_j^I = 0$ ), and *ii*) the expected number of ODDs conditional on  $\mathcal{I}_j^I = 1$ . Empirically, the two processes can be jointly modeled using a zero-inflated count data model. The unconditional expected number of ODDs is the result of the combination of the zero-inflated and count part of the model, which are jointly estimated via maximum likelihood.

The zero-inflated model indicates the determinants of the two different processes causing a zero outcome (Lambert, 1992): choice (the decision not to invest in R&D) and nature (the lack of innovative output, conditional on the level of effort) (Winkelmann, 2008). R&D effort, proxied by the number of ODDs targeting disease  $j$  granted in year  $t$ , is therefore modeled as:

$$y_{jt} = \begin{cases} 0, & \text{if } \mathcal{I}_{jt}^I = 0 \\ y_{jt}^*, & \text{if } \mathcal{I}_{jt}^I = 1 \end{cases} \quad (19)$$

where:

- $\mathcal{I}_{jt}^I$  is the binary variable introduced in Section 4. If  $\mathcal{I}_{jt}^I = 0$ , the outcome is a “certain zero”, also referred to as “strategic” or “structural” zero (Staub and Winkelmann, 2013). For the sake of consistency with the analysis in Section 4, we depart from the standard

assumption that the relevant probability distribution for the inflated part is either *Logistic* or *Normal* (hence, the estimated model is either Logit or Probit) and adjust the model for the *Gumbel* distribution;<sup>11</sup>

- $y_{jt}^*$  is a count variable, representing the number of ODDs targeting disease  $j$  granted by FDA in year  $t$ . From the analysis in Section 4, under the assumptions of our model, its distribution can be approximated by a *Poisson*, with parameter  $\lambda_j = \tilde{N}^f(\hat{\delta}_j) \cdot p_j^d(I_j^*)$ . However, given that  $\lambda_j$  is disease-specific, when several diseases are considered, it is natural to refer to the Negative Binomial distribution, to account for over-dispersion.

The density for  $y_{jt}$  is:

$$f(y_{jt}) = \begin{cases} \mathcal{P}(\mathcal{I}_{jt}^I = 0) + [1 - \mathcal{P}(\mathcal{I}_{jt}^I = 0)]f^{nb}(0) & \text{if } y_{jt} = 0 \\ [1 - \mathcal{P}(\mathcal{I}_{jt}^I = 0)]f^{nb}(y_{jt}^*) & \text{if } y_{jt} \geq 1, \end{cases} \quad (20)$$

where  $f^{nb}$  is the density function of the Negative Binomial distribution. The probability of being in the ‘‘certain zero’’ case ( $\mathcal{I}_{jt}^I = 0$ ) is estimated using the Gumbel distribution:

$$\mathcal{P}(\mathcal{I}_{jt}^I = 0) = \exp(-\exp(-x'_{jt}\beta_1)). \quad (21)$$

Conditional on  $\mathcal{I}_{jt}^I = 1$ , the expected number of ODDs is:

$$\lambda_{jt} = \exp(x'_{jt}\beta_2). \quad (22)$$

Combining the two processes, the unconditional expected number of ODDs can be expressed as:

$$E(y_{jt}|x_{jt}) = (1 - \mathcal{P}(\mathcal{I}_{jt}^I = 0)) \lambda_{jt} = (1 - \exp(-\exp(-x'_{jt}\beta_1))) \exp(x'_{jt}\beta_2). \quad (23)$$

Our identification strategy cannot rely on the existence of a set of untreated diseases, given that our measure of R&D effort is specific to orphan diseases and that incentives are in force over the whole observation period. However, incentives accumulated over time as new reforms were approved, so that the overall intensity of treatment (incentive) grew after each reform. Moreover, according to our theoretical analysis, disease prevalence affects the impact of incentives. We exploit these characteristics, in a difference-in-differences framework. Referring to equations 21 and 22, we specify:

$$x'_{jt}\beta = \alpha + \sum_{i=0}^4 \zeta_i C i_j + \sum_{p=1}^4 \tau_p D p_t + \sum_{i=0}^4 \sum_{p=1}^4 \kappa_{ip} (C i_j \times D p_t) + \theta W_{jt}, \quad (24)$$

---

<sup>11</sup>The STATA<sup>®</sup> code used for estimation is available from the authors upon request.

where  $\beta$  refers to both  $\beta_1$  and  $\beta_2$ , which are separately estimated. Note that we normally use the same set of variables in the Gumbel and Negative Binomial part of the model.  $Ci$  denotes the class of prevalence, from the rarest ( $C1$ ) to the least rare ( $C4$ ) (see Table 3). The binary variables  $Dp$  indicate relevant periods of time, related to the introduction of special legislation in the three geographical areas of interest, and to joint application in the US and EU, which enhanced access to the incentives: 1983-1992 (incentives available in the US only); 1993-1999 (incentives available in the US and Japan); 2000-2007 (incentives also available in Europe) and 2008-2016 (possibility of joint application in Europe and the US).  $C1$  and  $D1$  are taken as reference categories.

The coefficients  $\kappa_{ip}$  are the main parameters of interest, both in the Gumbel and Negative Binomial part of the model, representing the differential effect of the reforms for diseases in class of prevalence  $Ci$ , compared to those in the lowest class of prevalence. The sign of  $\kappa_{ip}$  in the Gumbel part of the model is related to the signs of  $\partial\mathcal{P}(\mathcal{I}_j^I = 0)/\partial z\partial n_j$  and  $\partial\mathcal{P}(\mathcal{I}_j^I = 0)/\partial\gamma\partial n_j$ ; in the count part of the model, conditional on positive investments, it is related to the compound effect of  $\partial I_j^*/\partial z\partial n_j$  and  $\partial I_j^*/\partial\gamma\partial n_j$ . Based on our theoretical analysis, we expect  $\kappa_{ip} < 0$  for all  $i$  and all  $p$  in the Gumbel part of the model, whereas the sign may be positive or negative in the count part of the model.  $W$  is a vector including additional control variables which, according to the analysis in Section 4, may have an impact on R&D effort:

- a dummy variable (*EarlyD*) indicating whether the disease causes premature death (at pediatric age or in adulthood; 9% in our sample). This variable might affect the per patient net revenue,  $m_j$ , as some regulators grant a price premium to drugs targeting life-threatening conditions, and pediatric drugs are granted additional market exclusivity;
- the stock of publications (*SP*), given that inputs from science can play a relevant role in stimulating R&D efforts (Pavitt, 1984; Mansfield, 1995);
- dummy variables for 26 therapeutic classes (*TC*) and a dummy variable identifying genetic diseases (*G*), as assigned to diseases by Orphanet.

## 8 Empirical Results

Table 4 shows, for each specification, the estimated coefficients for the interactions of interest for the zero-inflated (Gumbel) part and of the “count” part of the model.<sup>12</sup> Full results are reported in Appendix D, Table 7. Dummy variables for disease prevalence and time period, therapeutic class and genetic diseases are included in all specifications. Standard errors are clustered by pathology.

---

<sup>12</sup>Diseases with missing prevalence are included, but interactions are not shown.

Dep. variable	(1)		(2)		(3)		(4)	
	$y_{it}$		$y_{it}$		$y_{it}$		$y_{it+5}$	
	Gumbel	count	Gumbel	count	Gumbel	count	Gumbel	count
$C2 \times D2$	-3.415*** (1.043)	-1.732*** (0.525)	-3.335*** (0.966)	-1.695*** (0.566)	-3.191*** (0.963)	-1.738*** (0.546)	-3.487*** (1.137)	-1.989*** (0.492)
$C2 \times D3$	-3.139** (1.303)	-1.325** (0.575)	-3.010*** (1.049)	-1.269** (0.563)	-2.846*** (0.955)	-1.321** (0.560)	-3.746*** (0.966)	-1.436*** (0.497)
$C2 \times D4$	-3.568*** (1.329)	-1.347** (0.584)	-3.442*** (1.088)	-1.307** (0.586)	-3.274*** (0.993)	-1.406** (0.591)	-3.258*** (0.832)	-1.285*** (0.447)
$C3 \times D2$	-2.549*** (0.733)	-1.286*** (0.407)	-2.478*** (0.724)	-1.253*** (0.461)	-2.324*** (0.751)	-1.251*** (0.421)	-3.698*** (0.881)	-1.826*** (0.432)
$C3 \times D3$	-2.626*** (0.698)	-1.011*** (0.335)	-2.552*** (0.639)	-0.972*** (0.347)	-2.412*** (0.656)	-1.055*** (0.372)	-3.675*** (1.066)	-1.480*** (0.486)
$C3 \times D4$	-2.855*** (0.781)	-1.114*** (0.354)	-2.800*** (0.722)	-1.089*** (0.361)	-2.718*** (0.712)	-1.266*** (0.396)	-4.011*** (1.305)	-1.401*** (0.531)
$C4 \times D2$	-2.416 (1.624)	-1.239* (0.708)	-2.291 (1.784)	-1.193* (0.714)	-2.326 (1.870)	-1.275* (0.748)	-4.332*** (1.270)	-2.148*** (0.617)
$C4 \times D3$	-4.327*** (1.636)	-1.438** (0.688)	-4.183** (1.774)	-1.372** (0.686)	-4.187** (1.917)	-1.535* (0.793)	-6.202*** (1.322)	-2.126*** (0.646)
$C4 \times D4$	-4.561*** (1.612)	-1.607** (0.718)	-4.475** (1.747)	-1.565** (0.706)	-4.495** (1.881)	-1.803** (0.815)	-6.265*** (1.555)	-2.020*** (0.616)
$C0, C2, C3, C4$	yes		yes		yes		yes	
$D2, D3, D4$	yes		yes		yes		yes	
$TC\&G$	yes		yes		yes		yes	
$EarlyD$	no		yes		yes		no	
$\ln(SP_{j,t-5})$	no		no		yes		no	

Robust (clustered by pathology) standard errors in parentheses.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 4: Estimation results: interaction effects,  $\kappa_{ip}$  in Eq. 24.

Column (1) reports results for our baseline specification. In the Gumbel part of the model, all but one of the coefficients associated with the interaction terms are negative and statistically significant: the increase in incentives provided over time brings an advantage for less rare diseases with respect to the most rare in terms of the extensive margin of R&D, in line with Proposition 2.

As a complement to these results, Figure 4(a) plots the predicted probability of a “certain zero” for the classes of prevalence  $C1$  and  $C4$  conditional on the controls of Column (1) (see Figure 7 in Appendix D for an illustration including all classes of prevalence). From the second (93-99) to the third period (00-07, when market exclusivity is additionally provided in the EU), a larger variation in probability is detected for  $C4$ . The difference of -34 percentage points is statistically different from zero (p-value = 0.004).

The coefficients of the interaction terms are also significant in the count part of the model of Column (1), indicating a heterogeneous impact of the reforms according to prevalence. In this case, the sign of the difference, for which the theoretical prediction is ambiguous, is in favor

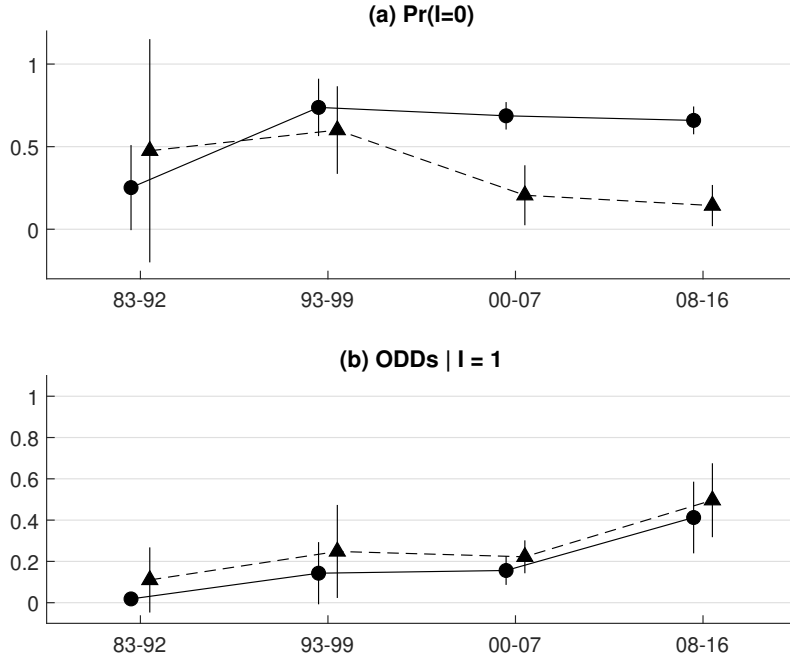


Figure 4: Average values for class of prevalence  $C1$  (continuous line) and  $C4$  (dashed line) in each period (with 95% confidence interval): (a) predicted probability that  $\mathcal{I}_j = 0$  (Gumbel; Eq. 21); (b) predicted number of ODDs, conditional on  $\mathcal{I}_j > 0$ , computed as  $\exp(x_{jt}\hat{\beta}_2)$  (count; Eq. 22).

of the class with the lowest prevalence. In Figure 4(b) we plot the average of the exponential values of  $x_{jt}\hat{\beta}_2$ . This corresponds to the predicted number of ODDs per disease conditional on  $\mathcal{I}_{jt}^I > 0$  (see Eq. 22). The difference between  $C1$  and  $C4$  is never statistically different from zero.<sup>13</sup>

We combine the estimated coefficients in the Gumbel and count part of the model to compute the predicted number of ODDs per year per disease (Eq. 23). Figure 5 depicts, in each period, (a) the estimated number of ODDs obtained by combining the two parts of the model, and (b) the difference between the estimated number of ODDs per disease in  $C4$  and  $C1$ . Over time, there has been an increase in the number of ODDs for all classes of prevalence. This was led by a decrease in the probability of no positive investment for diseases belonging to  $C4$  and by an increase in the intensity of investments, conditional on  $\mathcal{I}_{jt}^I = 1$ , for diseases belonging to  $C1$ . However, the magnitude of the heterogeneous impact on the probability of making an investment (Gumbel part) outweighs the effects on research intensity (count part) which go in the opposite direction. Thus, over time, we observe a widening gap in the predicted number of ODDs for a disease in the lowest *versus* the highest class of prevalence. The difference in the

<sup>13</sup>Note that coefficients in the table represent semi-elasticities, whereas the figure shows the exponential of the linear combination corresponding to the predicted number of ODDs conditional on  $\mathcal{I}_j > 0$ .

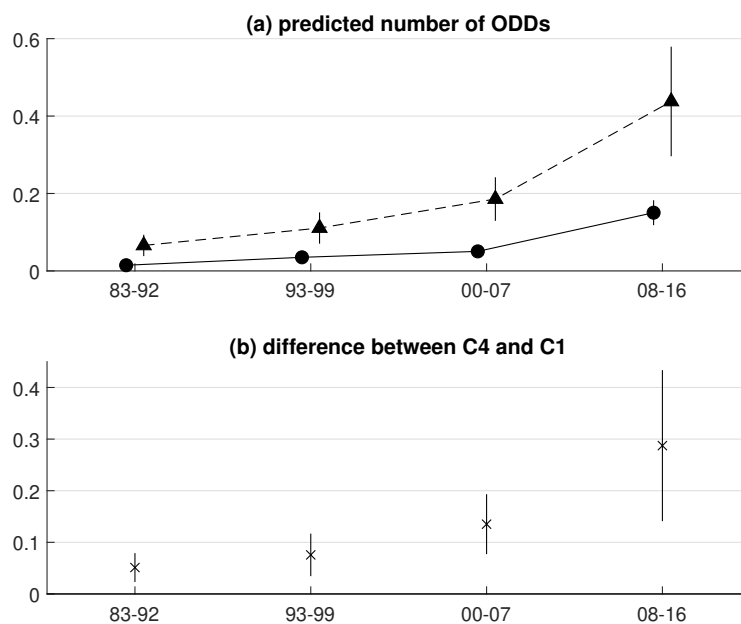


Figure 5: (a) Predicted number of ODDs (see Eq. 23) for  $C1$  (continuous line) and  $C4$  (dashed line); (b) difference between  $C4$  and  $C1$  (95% confidence interval reported in the graphs).

number of ODDs is 5.6 times larger in the last period as compared to the first one.<sup>14</sup>

Columns (2) and (3) of Table 4 include additional control variables (respectively,  $EarlyD$  and  $\ln(SP_{j,t-5})$ ). Column (4) takes into account the possibility that, even with an immediate impact of the reform on R&D effort, the increase in the number of ODDs may be delayed. Therefore, the effect of independent variables at time  $t$  on the number of ODDs in  $t + 5$  is considered. The five-year lag has been selected as it is the average time span from the beginning of clinical trials to an ODD application.<sup>15</sup> Importantly, results about the heterogeneous effect of Orphan Regulations across classes of prevalence shown in Column (2)-(4) confirm the results of the baseline specification of Column (1). Moreover, when the time lag is taken into account, the estimated effect of the reforms is larger. This is in line with the idea that, when not accounting for the lag, the impact of the reform is estimated in a period in which it has yet to deploy its

<sup>14</sup>As the Japanese law considers the lowest threshold for defining an orphan drug (about 4 in 10 thousand), a subset of diseases in  $C4$  do not benefit from incentives in this country. Hence, starting from period  $D2$ , the estimated coefficients of  $C4$  and its interactions may represent a lower bound.

<sup>15</sup>The length of the lag was estimated by combining our own calculations using FDA data with data on the length of drug development provided by DiMasi et al. (2016). According to our calculations, the *average* time lag between designation and marketing approval for drugs designated before 2005 is 68 months. Analysis is not extended beyond 2005 to avoid data censoring. DiMasi et al. (2016) indicate a time period of 126 months from synthesis to approval. Taking the difference between these two numbers, designations take place on average five years after the synthesis of the compound. This result is in line with Hay et al. (2014), who find that ODDs are most often given to a drug in phase 2, that, according to DiMasi et al. (2016), is roughly five years after synthesis.

effects. If this is the case, impacts estimated not accounting for the lag may be a lower bound.

The empirical analysis does not allow us to formally distinguish between the impact of *pull* versus *push* incentives, as both have been part of US and Japan regulations since their introduction. However, the European legislation provides solely *pull* incentives, whereas tax-related provisions are delegated to single countries. The results of Proposition 1, coupled with the fact that only two European countries provide tax incentives (see Section 2), suggest that the design of the European legislation may have contributed substantially to the widening of the gap between less and more rare diseases that we observe in Figure 5 after 2000.

There are clearly some challenges in the identification of a causal impact of orphan drug regulations adopted over time in our empirical analysis. One concern is that, due to the long time span considered, other events, in addition to those for which we can control for, may have had an impact on the development of new orphan drugs. One example is the Human Genome Project completed in 2003: this may have increased R&D for genetic diseases, which are a large share of orphan diseases. However, in our data, 90% of diseases in *C1* have a genetic origin against 41% in *C4*. If the Project increased R&D for genetic diseases, our estimate of the greater impact of the European legislation for *C4* versus *C1* would be biased downward. What may limit the validity of our conclusions is the presence of events other than orphan legislation, which had a stronger impact on diseases with a comparatively high level of prevalence. Although we are not aware of any relevant event with these characteristics during the period of analysis, in Appendix D (Table 9) we show that our results are robust to changes in the specification aiming to control for the impact of such events.

Additional robustness checks related to sample selection and the way in which we count the number of ODDs, are set out in Appendix D. The main results presented in this Section are confirmed.

## 9 Concluding remarks

Since the early 80s, regulators have started to address the lack of incentives to invest in innovation for rare diseases by means of specific provisions. As the pharmaceutical market is global, these incentives for the development of orphan drugs have increased over time as new countries have introduced them. There is ample evidence that this has increased investments in projects targeting rare diseases, with a potential reduction in inequality between orphan and common diseases. In this article, we study the impact of these incentives on the distribution of R&D efforts within the class of orphan diseases, with a focus on heterogeneity in relation to prevalence.

We develop a theoretical model considering both *pull* programs, such as market exclusivity,



and *push* programs, such as tax credits. The model shows that, under plausible assumptions, both types of incentive increase the *probability of observing investment* more for a less rare disease. At the firm level, this is due to both a direct and an indirect impact of *pull* incentives, whereas for *push* incentives the impact is only indirect. In terms of the *optimal level of R&D investment*, it is not possible to conclude unambiguously whether the impact of the incentives increases or decreases with the prevalence of the disease. A calibration exercise suggests that the impact on the extensive margin prevails, meaning that, overall, the relative disadvantage for very rare diseases is greater with *pull* than with *push* incentives.

The number of orphan designations, a condition for eligibility for incentives, is used as a proxy of R&D effort to investigate empirically the impact of the introduction of incentives on different geographical areas over time. We utilize a Gumbel zero-inflated negative binomial model which, in line with the theoretical analysis, allows us to study separately the intensive and extensive margin of R&D effort. We find that the number of new designations per year has increased over time for all orphan diseases, but inequality within orphan diseases has also increased: the difference between the predicted number of new orphan designations per year for a disease in the highest and the lowest class of prevalence is 5.6 times larger in the last than the first period of analysis. This result is driven by the extensive margin of R&D. Our empirical analysis suggests that the introduction of the European legislation, mainly providing *pull* incentives, contributed substantially to this outcome. Although one may argue that the causal interpretation of our results may be challenged by the presence of events, other than the orphan legislation, which had a heterogeneous impact according to prevalence, our findings are supported by several robustness checks.

The combination of our theoretical and empirical results leads to clear-cut policy implications. If providing as many patients as possible with some therapeutic option is an objective of European policy makers, an increase in the relative weight of *push* incentives should be considered. However, the adoption of some of these incentives, such as tax credits, may be more challenging at the European level than in other regulatory frameworks, due to the fact that single EU member countries are still responsible for the definition of fiscal policies. An alternative approach might consider the introduction of provisions tailored to the disease prevalence (Jobjörnsson et al., 2016). Finally, it should be noted that these policy implications are based on equity considerations. A challenge for future research is the exploration of the relationship between equity and efficiency issues in the field of orphan diseases.

### Acknowledgments

We would like to thank participants at the Essen Health Conference 2020, the 4<sup>th</sup> meeting of the American-European Health Economics Study Group (Vienna, 2019), the 12<sup>th</sup> Annual Conference on Innovation Economics (Chicago, 2019), the 18<sup>th</sup> Journées LAGV (Aix-en-Provence, 2019), the 8<sup>th</sup>

ZEW/MaCCI Conference on the Economics of Innovation and Patenting (Mannheim, 2019), the 10<sup>th</sup> Annual Meeting of the Network on Economics of Regulation and Institutions (Rome, 2019), the 8<sup>th</sup> Italian Congress of Econometrics and Empirical Economics (Lecce, 2019), the European Health Economics Association Conference (Maastricht, 2018), and at the seminar series of the University of Milan Bicocca (2019). We are particularly grateful to Marco Caliari, Annika Herr, Giuseppe Moscelli and Van Ahn Vuong for their helpful suggestions.

### **Funding**

This work was supported by the University of Verona under the program ‘Bando di Ateneo per la Ricerca di Base’.

### **Conflict of interest**

The authors declare that there is no conflict of interest regarding the publication of this article.

## **References**

- Acemoglu, D. and Linn, J. (2004). Market size in innovation: theory and evidence from the pharmaceutical industry. *The Quarterly Journal of Economics*, 119(3):1049–1090.
- Ahsanullah, M. (2016). *Extreme Value Distributions*. Tsokos, C.P (ed.), Atlantis Studies in Probability and Statistics, Vol. 8.
- Barrenho, E., Miraldo, M., and Smith, P. C. (2019). Does global drug innovation correspond to burden of disease? The neglected diseases in developed and developing countries. *Health Economics*, 28(1):123–143.
- Braun, M. M., Farag-El-Massah, S., Xu, K., and Coté, T. R. (2010). Emergence of orphan drugs in the United States: a quantitative assessment of the first 25 years. *Nature Reviews Drug Discovery*, 9(7):519.
- DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47:20–33.
- Dubois, P., De Mouzon, O., Scott-Morton, F., and Seabright, P. (2015). Market size and pharmaceutical innovation. *The RAND Journal of Economics*, 46(4):844–871.
- Gibson, S. and von Tigerstrom, B. (2015). Orphan drug incentives in the pharmacogenomic context: policy responses in the us and canada. *Journal of Law and the Biosciences*, 2(2):263–291.

- Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40.
- Health and Safety, F. (2015). *Inventory of Union and Member State incentives to support research into, and the development and availability of, orphan medicinal products*. European Commission.
- Heemstra, H. E., van Weely, S., Büller, H. A., Leufkens, H. G., and de Vruet, R. L. (2009). Translation of rare disease research into orphan drug development: disease matters. *Drug Discovery Today*, 14(23-24):1166–1173.
- Herder, M. (2017). What is the purpose of the Orphan Drug Act? *PLoS Medicine*, 14(1):e1002191.
- INSERM (1999). *Orphadata: Free access data from Orphanet*. available on <http://www.orphadata.org> (data version: 1.2.4/4.1.6).
- Jobjörnsson, S., Forster, M., Pertile, P., and Burman, C.-F. (2016). Late-stage pharmaceutical R&D and pricing policies under two-stage regulation. *Journal of Health Economics*, 50:298–311.
- Keller, W. (2002). Geographic localization of international technology diffusion. *American Economic Review*, 92(1):120–142.
- Kremer, M. (2002). Pharmaceuticals and the developing world. *Journal of Economic Perspectives*, 16(4):67–90.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.
- Lichtenberg, F. R. and Waldfoegel, J. (2009). Does misery love company? Evidence from pharmaceutical markets before and after the Orphan Drug Act. *Michigan Telecommunications and Technology Law Review*, 15:335–357.
- Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *The Review of Economics and Statistics*, 77:55–65.
- Mariz, S., Reese, J. H., Westermarck, K., Greene, L., Goto, T., Hoshino, T., Llinares-Garcia, J., and Sepodes, B. (2016). Worldwide collaboration for orphan drug designation. *Nature Reviews Drug discovery*, 15(6):440.
- Melnikova, I. (2012). Rare diseases and orphan drugs. *Nature Reviews Drug Discovery*, 11:267–268.

- Pammolli, F., Riccaboni, M., and Magazzini, L. (2009). Nuove politiche per l'innovazione nel settore delle scienze della vita. *Rapporto CERM*.
- Pavitt, K. (1984). Sectoral patterns of technical change: towards a taxonomy and a theory. *Research Policy*, 13(6):343–373.
- Raïs Ali, S. and Tubeuf, S. (2019). (In-)Equality in the allocation of R&D resources for rare diseases. *Social Justice Research*, 32:277–317.
- Rietzke, D. and Chen, Y. (2020). Push or pull? performance-pay, incentives, and information. *The RAND Journal of Economics*, 51(1):301–317.
- Roemer, J. (1998). *Equality of Opportunity*. Harvard University Press.
- Salop, S. C. (1979). Monopolistic competition with outside goods. *The Bell Journal of Economics*, 10(1):141–156.
- Staub, K. E. and Winkelmann, R. (2013). Consistent estimation of zero-inflated count models. *Health Economics*, 22(6):673–686.
- Tambuyzer, E. (2010). Rare diseases, orphan drugs and their regulation: questions and misconceptions. *Nature Reviews Drug Discovery*, 9(12):921.
- Wästfelt, M., Fadeel, B., and Henter, J.-I. (2006). A journey of hope: lessons learned from studies on rare diseases and orphan drugs. *Journal of Internal Medicine*, 260(1):1–10.
- Westermarck, K. et al. (2011). European regulation on orphan medicinal products: 10 years of experience and future perspectives. *Nature Reviews Drug Discovery*, 10(5):341.
- Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.
- Yin, W. (2008). Market incentives and pharmaceutical innovation. *Journal of Health Economics*, 27(4):1060–1077.

## **A The impact of incentives in a competitive model**

Following Yin (2008), the market of an orphan disease  $j$  is modeled as a unit circle, with patients uniformly distributed along its length (Salop, 1979). Let  $F_j$  be the number of identical firms active in the market of disease  $j$ , supplying different drugs that are not vertically differentiated. Each firm supplies one drug, so that the number of firms active in the market equals the number

of products. Quality is also assumed to be homogeneous, meaning that patient choices depend only on their location in relation to suppliers. Each patient buys from the closest supplier, so that each firm serves  $n_j/F_j$  patients. For the sake of simplicity, we assume  $I$  exogenous<sup>16</sup> and  $p_j^d = 1$ . The profit function can be written as:

$$E\Pi_j = M_j(1+z)\frac{n_j}{F_j} - (1-\gamma)I. \quad (25)$$

Below, we study the impact of *pull* and *push* incentives on the equilibrium number of firms  $\bar{F}_j$ , and in particular how this impact changes with  $n_j$ .

### A.1 Pull incentive

Assuming that firms are free to enter the market, the usual zero profit condition applies:<sup>17</sup>

$$E\Pi_j = M_j(1+z)\frac{n_j}{F_j} - (1-\gamma)I = 0. \quad (26)$$

By applying the implicit function theorem, we find:

$$\frac{\partial \bar{F}_j}{\partial z} = \frac{\bar{F}_j}{1+z} > 0. \quad (27)$$

As expected, *pull* incentives increase the equilibrium number of firms. Regarding the most interesting question from our perspective, i.e. how this impact changes with  $n_j$ , we find that:

$$\frac{\partial^2 \bar{F}_j}{\partial z \partial n} = \left( \frac{1}{1+z} \right) \frac{\partial \bar{F}_j}{\partial n_j}. \quad (28)$$

Once more applying the implicit function theorem to Eq. 26, we obtain:

$$\frac{\partial \bar{F}_j}{\partial n_j} = \frac{\bar{F}_j}{n_j}, \quad (29)$$

which can be replaced in Eq. 28 to obtain:

$$\frac{\partial^2 \bar{F}_j}{\partial z \partial n} = \frac{\bar{F}_j}{n_j(1+z)} > 0. \quad (30)$$

---

<sup>16</sup>This is equivalent to introducing the hypothesis that investment allows to enhance drug quality, for which a minimum threshold is set, assumed binding for orphan diseases (see Yin, 2008).

<sup>17</sup>To be more precise,  $\bar{F}_j$  should be defined as the largest integer such that the expected profit is non negative. Given the aim of the analysis, and without loss of generality, we treat  $\bar{F}_j$  as a continuous variable.

Hence, in a model with several firms, an increase in the level of *pull* incentives has a greater impact on the number of firms / products for less rare diseases.

## A.2 *Push* incentives

Referring to Eq. 26, the same methods are applied as for the *pull* incentive to determine:

$$\frac{\partial \bar{F}_j}{\partial \gamma} = \frac{I \bar{F}_j^2}{M_j n_j (1+z)} > 0. \quad (31)$$

In this case too, the impact on the equilibrium number of firms / products of the incentive is positive. Concerning the heterogeneous effect across diseases with different levels of prevalence, we find that:

$$\frac{\partial^2 \bar{F}_j}{\partial \gamma \partial n} = \frac{2n_j I \bar{F}_j \frac{\partial \bar{F}_j}{\partial n} - I \bar{F}_j^2}{M_j n_j^2 (1+z)}. \quad (32)$$

Replacing  $\frac{\partial \bar{F}_j}{\partial n}$  from Eq. 29 gives:

$$\frac{\partial^2 \bar{F}_j}{\partial \gamma \partial n} = \frac{\bar{F}_j^2 I}{M_j n_j^2 (1+z)} > 0. \quad (33)$$

According to Eq. 33, with a *push* incentive too, the impact on the equilibrium number of firms / products is greater for less rare diseases.

## B Calibration details

The calibration process is divided into two main parts. The first is the creation of a sample of diseases with different values of prevalence. A number of diseases is considered equal to the number of diseases for which the information on prevalence is not missing in our dataset (2,553, see Table 3). Although prevalence is expressed in relative terms, what matters to a firm when shaping its optimal investment policy is the absolute size of the market. The thresholds separating classes of prevalence are transformed by assuming a total population size of 965.1 million, the sum of the populations (reference year: 2017) in the three main geographical areas where orphan legislation was introduced. These values (see Column 3 in Table 5) and the corresponding percentiles (Column 4) are used as a benchmark for the calibration of the distribution of prevalence among diseases in our simulated sample. Next, an iterative process is used to identify the combination of mean and standard deviation of the Log-Normal distribution that minimizes the sum of the distance, in percentage terms, between the value of prevalence corresponding to the switching percentiles (i.e. 74.17, 82.19 and 94.01; see Column (4) of Table 5) in the original

Class (1)	Threshold		Prevalence: threshold percentile		# ODD: cumulative distribution	
	Relative (2)	Absolute (3)	Original (4)	Calibrated (5)	Original (6)	Calibrated (7)
<i>C1</i>	1/1,000,000	965.1	74.17	72.42	15.72	15.96
<i>C2</i>	9/1,000,000	8,685.9	82.19	87.17	45.49	42.38
<i>C3</i>	9/100,000	86,859	94.01	95.15	60.54	62.89
<i>C4</i>	5/10,000	482,550	100.00	100.00	100.00	100.00

Table 5: Summary of information used to calibrate the model.

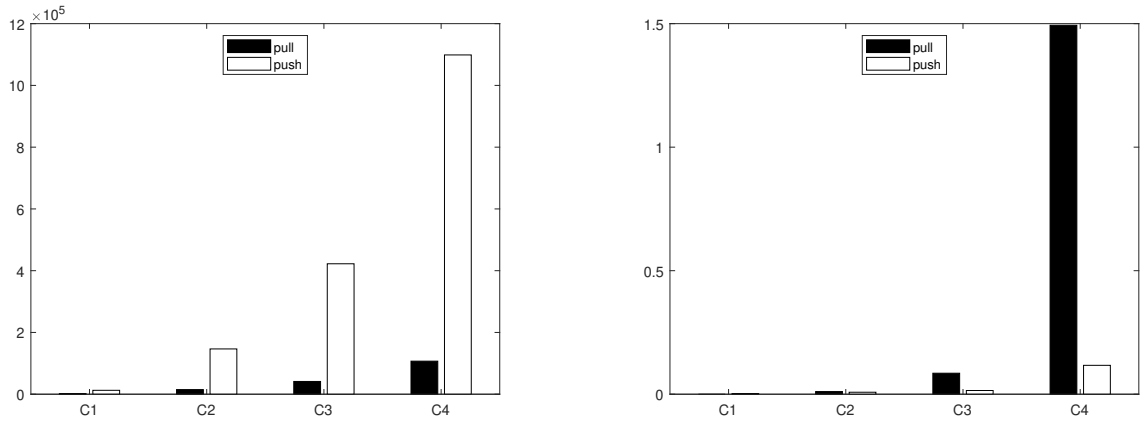
Parameter	Description	Value
mean prevalence	mean of the Log-Normal distribution	4.41
s.d. prevalence	s.d. of the Log-Normal distribution	4.2
$M_j$	expected net revenue per patient	300
$\epsilon$	parameter of $p_j^d(I)$ function	1.4e-05
$E[\Delta]$	mean of the idiosyncratic term	-80e+06
s.d. $[\Delta]$	s.d. of the idiosyncratic term	49e+06
# diseases	number of diseases	2553
# firms	number of firms	20

Table 6: Calibrated parameter values.

distribution (Column 3) and its calibrated counterpart. Finally, given that the generated sample of diseases includes some that could not be considered orphan because their prevalence exceeds the highest threshold (482,550), we drop them (1.92% of the originally generated sample). The match between the calibrated sample of diseases in Column (5) and the original sample looks sufficiently close.

The next step in calibration should ideally be taken to set the parameters to obtain a reliable distribution of ODDs across classes of prevalence in a situation without incentives. The fact that we have no data for the period before the introduction of incentives rules this out. To overcome this problem, the benchmark chosen was the distribution of ODDs across classes of prevalence cumulated over the 10 years after 1983, i.e. the period when only the US had orphan legislation. We are implicitly assuming that the impact of the incentives was still limited in this period compared to subsequent years. Column (6) of Table 5 reports the cumulative distribution of ODD for the four classes of prevalence. The remaining parameters in the model are set in order to obtain, for the simulation of the model with no incentives (step 1), a distribution of ODD across classes of prevalence which is reasonably similar to this benchmark. Table 6 reports the parameter values used in our preferred calibration. Column (7) of Table 5 shows the calibrated cumulative distribution of ODDs for the situation without incentives.

As stated in Section 5, the model is solved: (1) without incentives, (2) with only a *pull*



(a) Change in optimal investment level per disease by class of prevalence ( $z = 0.2, \gamma = 0.745$ ). (b) Change in the number of firms that invest per disease by class of prevalence ( $z = 0.2, \gamma = 0.745$ ).

Figure 6: Change in optimal investment level and number of investors due to the incentives.

incentive ( $z = 0.2, \gamma = 0$ ), (3) iterating the solution to find the value of a *push* incentive leading to the same number of ODDs as in the second step, with  $z = 0$ . This condition is satisfied with  $\gamma = 0.745$ .

The most significant output, i.e. the distribution of the additional number of ODDs across classes of prevalence for the two types of incentive, is shown in the main text (Figure 1). Figure 6 shows how the two decisions considered in our theoretical analysis, i.e. investment intensity and whether or not to invest, contribute to this result. Figure 6(a) shows that, given our parameter values, the impact of both types of incentives on the optimal investment level, which was ambiguous in the theory, increases with the level of prevalence. For all classes of prevalence, the increase is greater for *push* incentives, due to the lower expected value of the marginal cost of the investment involved. On the other hand, the distribution of additional firms that decide to invest is greatly in favor of less rare diseases when the incentive is *pull* (Figure 6(b)). This is in line with the theoretical result that only for a *pull* incentive is a direct mechanism at work (Proposition 1). Therefore, the difference in terms of probability of investing (Figure 6(b)) seems to be driving the clear advantage for the class of less rare diseases in terms of additional ODDs shown in Figure 1.

## C Additional information on the sample of diseases

Starting from the initial sample of 9,530 records available from the Orphanet database, we excluded 2,208 which do not refer to a specific disease, but to aggregations called “group of phenomes” (e.g. “rare pulmonary diseases”). Diseases emerging in the antenatal period or



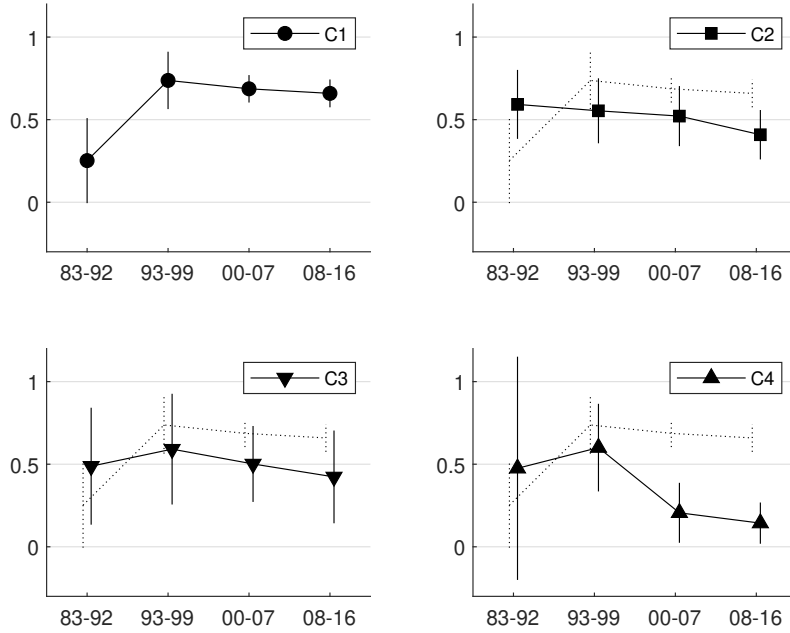


Figure 7: Predicted probability that  $\mathcal{I}_j = 0$  (Gumbel; Eq. 21): average values for the classes with known prevalence in each period (with 95% confidence interval; class  $C1$  reported in all graphs as a reference).

causing death before birth (323 diseases) were excluded. 568 diseases referring to surgical procedures were also dropped, and 192 items with obsolete nomenclature were update and moved accordingly.

Concerning the FDA database of ODDs, this includes 3,996 designations granted between 1983 to 2016. Of these, 408 records referring to products for surgery, prevention, transplant, diagnostics and imaging procedures were excluded, and 199 records were dropped because information on the treated disease cannot be retrieved from Orphanet.

Where the FDA designated indication refers to a “group of phenomes”, we rely on the hierarchical classification of orphan diseases provided by Orphanet to link the ODD with all relevant diseases.<sup>18</sup> In particular, one ODD is attributed to each disease, i.e. a non-fractional count is adopted. In the robustness checks (see Section D), results are shown for fractional counting.

	(1)		(2)		(3)		(4)	
	Gumbel	count	Gumbel	count	Gumbel	count	Gumbel	count
<i>C2</i>	2.022 (1.271)	1.587*** (0.584)	1.931* (0.997)	1.535*** (0.548)	1.801** (0.917)	1.507*** (0.554)	1.945** (0.924)	1.643*** (0.482)
<i>C3</i>	1.392** (0.644)	1.496*** (0.387)	1.374** (0.636)	1.472*** (0.362)	1.363* (0.723)	1.444*** (0.393)	2.061 (1.323)	1.795*** (0.500)
<i>C4</i>	1.318 (1.685)	1.792*** (0.669)	1.200 (1.805)	1.758*** (0.666)	1.317 (1.950)	1.749** (0.782)	2.457** (1.174)	2.280*** (0.577)
<i>C0</i>	2.496*** (0.478)	1.428*** (0.476)	2.450*** (0.538)	1.456*** (0.447)	2.392*** (0.555)	1.580*** (0.385)	2.397*** (0.494)	1.218*** (0.284)
<i>D2</i>	3.172*** (0.657)	2.051*** (0.449)	3.134*** (0.680)	2.040*** (0.469)	3.018*** (0.689)	2.046*** (0.402)	3.482*** (0.582)	2.381*** (0.309)
<i>D3</i>	2.704*** (0.907)	2.140*** (0.322)	2.629*** (0.759)	2.105*** (0.344)	2.520*** (0.648)	2.120*** (0.356)	3.507*** (0.537)	2.675*** (0.283)
<i>D4</i>	2.485*** (0.809)	3.112*** (0.300)	2.426*** (0.692)	3.090*** (0.321)	2.341*** (0.617)	3.119*** (0.345)	2.630*** (0.498)	3.137*** (0.257)
<i>C2</i> × <i>D2</i>	-3.415*** (1.043)	-1.732*** (0.525)	-3.335*** (0.966)	-1.695*** (0.566)	-3.191*** (0.963)	-1.738*** (0.546)	-3.487*** (1.137)	-1.989*** (0.492)
<i>C2</i> × <i>D3</i>	-3.139** (1.303)	-1.325** (0.575)	-3.010*** (1.049)	-1.269** (0.563)	-2.846*** (0.955)	-1.321** (0.560)	-3.746*** (0.966)	-1.436*** (0.497)
<i>C2</i> × <i>D4</i>	-3.568*** (1.329)	-1.347** (0.584)	-3.442*** (1.088)	-1.307** (0.586)	-3.274*** (0.993)	-1.406** (0.591)	-3.258*** (0.832)	-1.285*** (0.447)
<i>C3</i> × <i>D2</i>	-2.549*** (0.733)	-1.286*** (0.407)	-2.478*** (0.724)	-1.253*** (0.461)	-2.324*** (0.751)	-1.251*** (0.421)	-3.698*** (0.881)	-1.826*** (0.432)
<i>C3</i> × <i>D3</i>	-2.626*** (0.698)	-1.011*** (0.335)	-2.552*** (0.639)	-0.972*** (0.347)	-2.412*** (0.656)	-1.055*** (0.372)	-3.675*** (1.066)	-1.480*** (0.486)
<i>C3</i> × <i>D4</i>	-2.855*** (0.781)	-1.114*** (0.354)	-2.800*** (0.722)	-1.089*** (0.361)	-2.718*** (0.712)	-1.266*** (0.396)	-4.011*** (1.305)	-1.401*** (0.531)
<i>C4</i> × <i>D2</i>	-2.416 (1.624)	-1.239* (0.708)	-2.291 (1.784)	-1.193* (0.714)	-2.326 (1.870)	-1.275* (0.748)	-4.332*** (1.270)	-2.148*** (0.617)
<i>C4</i> × <i>D3</i>	-4.327*** (1.636)	-1.438** (0.688)	-4.183** (1.774)	-1.372** (0.686)	-4.187** (1.917)	-1.535* (0.793)	-6.202*** (1.322)	-2.126*** (0.646)
<i>C4</i> × <i>D4</i>	-4.561*** (1.612)	-1.607** (0.718)	-4.475** (1.747)	-1.565** (0.706)	-4.495** (1.881)	-1.803** (0.815)	-6.265*** (1.555)	-2.020*** (0.616)
<i>C0</i> × <i>D2</i>	-3.660*** (0.631)	-1.870*** (0.640)	-3.631*** (0.732)	-1.869*** (0.624)	-3.524*** (0.691)	-1.918*** (0.460)	-3.366*** (0.594)	-1.596*** (0.332)
<i>C0</i> × <i>D3</i>	-3.043*** (0.578)	-1.308*** (0.396)	-2.992*** (0.548)	-1.291*** (0.395)	-2.883*** (0.562)	-1.378*** (0.364)	-2.985*** (0.562)	-1.242*** (0.311)
<i>C0</i> × <i>D4</i>	-2.481*** (0.501)	-1.392*** (0.404)	-2.447*** (0.531)	-1.385*** (0.391)	-2.386*** (0.557)	-1.519*** (0.364)	-2.067*** (0.518)	-1.081*** (0.288)
<i>EarlyD</i>			-0.258 (0.363)	0.218 (0.153)	-0.187 (0.350)	0.230 (0.147)		
$\ln(SP_{j,t-5})$					-0.033 (0.028)	0.077*** (0.018)		
Constant	-1.351 (0.889)	-4.729*** (0.299)	-1.285* (0.720)	-4.745*** (0.301)	-0.996* (0.602)	-4.960*** (0.329)	-1.693*** (0.524)	-4.573*** (0.268)
$\ln(\alpha)$	0.898*** (0.250)		0.893*** (0.202)		0.794*** (0.160)		0.849*** (0.160)	
<i>N</i>	136036		136036		136036		111023	
<i>AIC</i>	56006.45		55958.26		55702.65		50339.77	
<i>BIC</i>	56939.41		56910.86		56674.90		51253.43	

Robust (clustered across pathologies) standard errors in parentheses.

Therapeutic class and genetic dummy variables included in all specifications.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 7: Full estimation results (Table 4)

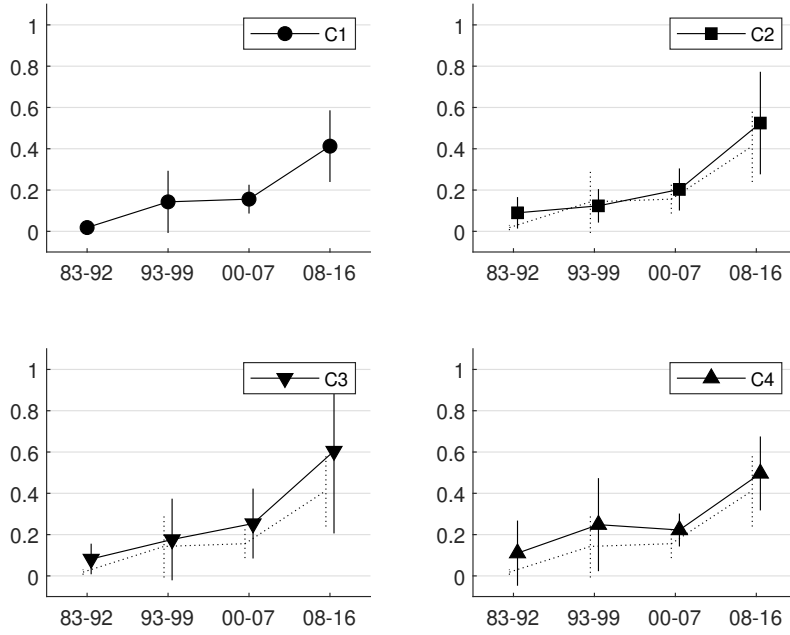


Figure 8: Predicted number of ODDs, conditional on  $\mathcal{I}_j = 1$ , computed as  $\exp(x_{jt}\hat{\beta}_2)$  (count; Eq. 22): average values for the classes with known prevalence in each period (with 95% confidence interval; class  $C1$  reported in all graphs as a reference).

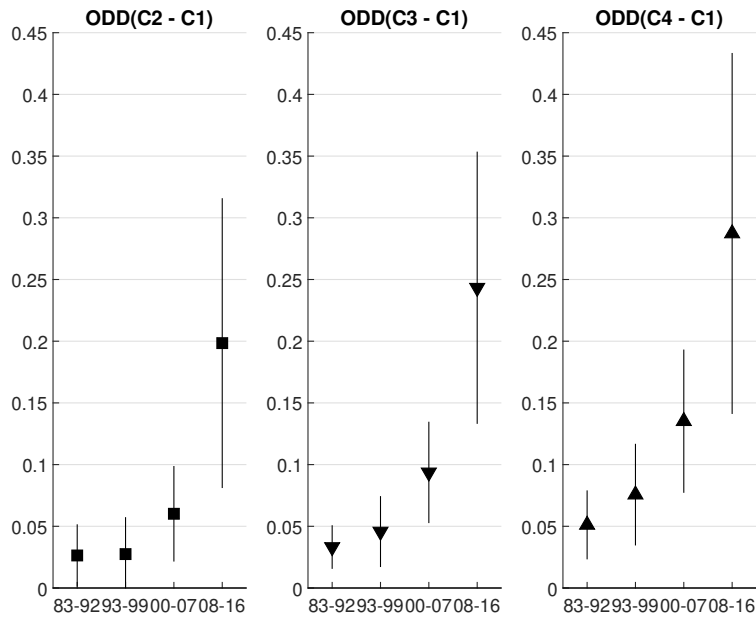


Figure 9: Differences in the expected number of ODDs with respect to  $C1$  (95% confidence interval reported in the graphs).

## D Detailed results and robustness checks

Table 7 reports the full set of estimated coefficients for the regression of Table 4. The results for the interactions of interest are commented in the main text. The coefficients of the dummy related to the time period show that, for diseases in  $C1$ , both the probability of a “certain zero” and the expected number of ODDs (conditional on positive investment) increase over time.

In the Gumbel part of the model, the value of the coefficients for the prevalence dummies measures the difference in  $\mathcal{P}(\mathcal{I}_{jt}^I = 0)$  between each class of prevalence and the reference category  $C1$  in the period 1983-92. These generally lack strong statistical significance, indicating that in the first time period considered in the analysis (83-92), classes of prevalence did not significantly differ in terms of probability of a “certain zero”.<sup>19</sup> However, they differ in the subsequent time periods, as highlighted by a test on the null hypothesis  $Ci + Ci \times Dt = 0$ : these sums are all negative and statistically significant, indicating that in all periods but the first the probability of a “certain zero” is lower for diseases in classes other than  $C1$ .

In the count part of the model, coefficients for  $Ci$  are positive and statistically significant: conditionally on  $\mathcal{I}_{jt}^I = 1$ , less rare diseases experience larger investments (higher number of ODDs) in the first time period. However, the gap among classes of prevalence shrinks over time in percentage terms.

For our main specification (Column 1) Figure 7 plots the predicted probability that  $\mathcal{I}_{jt}^I = 0$  for each class of prevalence, while Figure 8 does so for the predicted number of ODDs, conditional on  $\mathcal{I}_{jt}^I = 1$ . Finally, Figure 9 plots the differences in the predicted number of ODDs for each class of prevalence with respect to  $C1$ , as resulting from the combination of the Gumbel and count part of the model. Differences increase over time for all classes of prevalence, but especially for less rare diseases.

Column (2) of Table 7 takes into account the characteristics of the disease in terms of life expectancy, and includes a dummy variable that identifies diseases causing premature death ( $EarlyD$ ). This variable is not significant either in the Gumbel or in the count part of the model, but its joint effect in the two equations is statistically different from zero (p-value=0.033).

Column (3) includes the stock of publications at time  $t - 5$  (in log) to proxy the level of scientific knowledge related to disease  $j$ : inputs from science play a relevant role in stimulating R&D efforts. Indeed, results highlight that a larger stock of publications increases the number of ODDs in the count part (it also reduces the probability of a “certain zero”, although not

---

<sup>18</sup>For example, some drugs were designated for the treatment of hypereosinophilic syndrome, which is classified as a “group of phenomes” in Orphanet and comprises different diseases on the Orphanet list (i.e., idiopathic hypereosinophilic syndrome, primary hypereosinophilic syndrome, and secondary hypereosinophilic syndrome).

<sup>19</sup>In all models it is not possible to reject the null hypothesis that all coefficients for the class of prevalence are equal to zero at the 1% level.

significantly).

The next two subsections present further robustness checks that are organized along two dimensions. First, different ways of measuring the dependent variable (Table 8) are considered. Then, the sample is modified and additional control variables are introduced (Table 9).

## D.1 Counting the number of ODDs

Column (1) of Table 8 excludes from the count of ODDs the designations received after the drug obtained marketing approval for other indications.<sup>20</sup> In this case, the innovation can be considered less substantial. When excluding these designations from the count, results are qualitatively similar to those shown in Table 4.

In Column (2) we replace fractional with non-fractional counting. This means that, when an ODD is assigned to a “group of phenomes”, a number of ODDs equal to the reciprocal of the number of diseases in that group is assigned to each of them. Convergence issues arise, and we are forced to omit those diseases with (yet) unknown prevalence from the analysis. Estimated coefficients change, and statistical significance is no longer observed for the interaction terms in the count part of the model. However, results in the Gumbel part are confirmed, with a larger decrease in the probability of a “certain zero” for  $C4$ . In terms of the dynamics of the expected number of ODDs, the increasing effort directed towards less rare diseases (in class of prevalence  $C2$ ,  $C3$ , and  $C4$ ) compared to more rare diseases (class  $C1$ ) is confirmed.

Finally, Column (3) includes only ODDs assigned to private companies (96% of the ODDs in our sample), excluding ODDs assigned to universities, hospitals and medical centers, non-profit organizations and patient associations. Our main results are again unaffected.

## D.2 Sample issues and control variables

In the count part of the model presented in Column (1) of Table 9, additional controls are added in the form of interactions between therapeutic class dummies ( $TC_j$ ) and period dummies.<sup>21</sup> These interaction terms aim to capture the effect of technological reforms at the therapeutic class level. Technological breakthroughs fostering the level of innovative effort in a specific therapeutic class might bias our results if correlated with the level of prevalence.

Column (2) includes a proxy for the per patient net revenue at the industry level ( $mg$ ).<sup>22</sup> In

---

<sup>20</sup>The relevant information was retrieved from the list of orphan-designated products with at least one marketing approval for a common disease provided by the FDA and the Drugs@FDA database.

<sup>21</sup>The interaction terms are included only in the count part of the model as this specification has a lower Bayesian Information Criterion (BIC) compared to models where the interaction terms are also (or solely) included in the Gumbel part of the model.

<sup>22</sup>On the basis of the theoretical model, disease-specific per patient net revenues ( $m_j$ ) should affect incentives to undertake R&D investments. Unfortunately, we are not aware of reliable proxies for net revenues, as well as

	(1) excl.appr.		(2) fractional		(3) firm only	
	Gumbel	count	Gumbel	count	Gumbel	count
<i>C</i> 2	1.871* (1.068)	1.544** (0.619)	6.060*** (1.849)	1.820*** (0.678)	2.338** (1.151)	1.449** (0.607)
<i>C</i> 3	1.318** (0.643)	1.483*** (0.369)	5.470*** (1.769)	2.421*** (0.529)	1.435** (0.598)	1.411*** (0.310)
<i>C</i> 4	1.587 (1.820)	1.667** (0.752)	4.014** (1.592)	2.833*** (0.529)	1.040 (1.764)	1.656*** (0.530)
<i>C</i> 0	2.672*** (0.508)	1.554*** (0.401)			2.628*** (0.461)	1.346*** (0.294)
<i>D</i> 2	3.220*** (0.687)	2.148*** (0.360)	3.977* (2.285)	0.346 (0.573)	3.686*** (0.593)	1.982*** (0.303)
<i>D</i> 3	2.702*** (0.587)	2.161*** (0.340)	3.922** (1.761)	0.491 (0.565)	2.997*** (0.483)	2.169*** (0.262)
<i>D</i> 4	2.384*** (0.589)	3.082*** (0.329)	4.307*** (1.640)	1.427*** (0.507)	2.746*** (0.429)	3.133*** (0.255)
<i>C</i> 2 × <i>D</i> 2	-3.254*** (1.003)	-1.738*** (0.564)	-11.953*** (3.713)	-0.349 (0.652)	-3.825*** (0.922)	-1.590*** (0.457)
<i>C</i> 2 × <i>D</i> 3	-2.979*** (1.017)	-1.304** (0.610)	-11.903*** (2.899)	-0.045 (0.739)	-3.510** (1.515)	-1.175* (0.643)
<i>C</i> 2 × <i>D</i> 4	-3.427*** (1.146)	-1.339* (0.693)	-25.855*** (3.020)	0.246 (0.678)	-3.725*** (1.080)	-1.148** (0.563)
<i>C</i> 3 × <i>D</i> 2	-2.483*** (0.786)	-1.345*** (0.391)	-2.889 (2.531)	0.073 (0.610)	-3.144*** (0.864)	-1.320*** (0.365)
<i>C</i> 3 × <i>D</i> 3	-2.499*** (0.593)	-1.023*** (0.361)	-4.210** (1.787)	0.303 (0.590)	-2.876*** (0.616)	-0.961*** (0.303)
<i>C</i> 3 × <i>D</i> 4	-2.748*** (0.653)	-1.090*** (0.374)	-4.732** (2.160)	0.438 (0.553)	-3.091*** (0.582)	-1.040*** (0.321)
<i>C</i> 4 × <i>D</i> 2	-2.539 (1.772)	-1.137 (0.722)	-3.208 (2.550)	0.049 (0.622)	-2.499 (1.738)	-1.091** (0.531)
<i>C</i> 4 × <i>D</i> 3	-4.642** (1.804)	-1.408* (0.771)	-4.036** (1.980)	0.510 (0.615)	-4.514** (1.800)	-1.300** (0.514)
<i>C</i> 4 × <i>D</i> 4	-4.950*** (1.792)	-1.513* (0.791)	-4.800** (2.117)	0.306 (0.563)	-4.800*** (1.840)	-1.448*** (0.538)
<i>C</i> 0 × <i>D</i> 2	-3.838*** (0.636)	-2.036*** (0.428)			-4.105*** (0.589)	-1.765*** (0.339)
<i>C</i> 0 × <i>D</i> 3	-3.225*** (0.504)	-1.440*** (0.380)			-3.191*** (0.495)	-1.246*** (0.297)
<i>C</i> 0 × <i>D</i> 4	-2.588*** (0.511)	-1.480*** (0.353)			-2.544*** (0.451)	-1.317*** (0.289)
Constant	-1.248** (0.611)	-4.779*** (0.334)	-5.018*** (1.899)	-5.950*** (0.531)	-1.402** (0.600)	-4.667*** (0.319)
ln( $\alpha$ )	0.922*** (0.187)		0.197 (0.169)		1.038*** (0.128)	
<i>N</i>	136036		66257		136036	
<i>AIC</i>	54366.23		11832.12		54258.32	
<i>BIC</i>	55299.19		12623.93		55191.29	

Robust (clustered by pathology) standard errors in parentheses.

Therapeutic class and genetic dummy variables included in all specifications.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 8: Results – Robustness checks on how the number of ODDs is measured.

	(1)		(2)		(3)		(4)	
	$TC_j \times Dp_j$		$mg_t$		all obs.		$SP_{83-5} > 0$	
	Gumbel	count	Gumbel	count	Gumbel	count	Gumbel	count
$C2$	0.834 (0.656)	1.521** (0.613)	1.693* (1.009)	1.484*** (0.559)	2.848*** (0.653)	1.439*** (0.458)	2.909*** (0.567)	2.300*** (0.454)
$C3$	0.210 (0.567)	1.439*** (0.461)	1.127* (0.609)	1.409*** (0.353)	1.690*** (0.503)	1.171*** (0.225)	0.775 (0.579)	1.588*** (0.396)
$C4$	0.559 (0.789)	1.668*** (0.616)	1.024 (1.440)	1.715*** (0.567)	0.836 (1.569)	1.324*** (0.445)	1.418 (1.132)	1.904*** (0.544)
$C0$	0.471 (0.489)	0.668 (0.426)	0.260 (0.635)	1.322*** (0.393)	2.695*** (0.381)	1.118*** (0.245)	3.715*** (1.205)	2.315*** (0.780)
$D2$	0.722 (0.559)	0.487 (0.434)	3.318*** (0.627)	1.611*** (0.424)	3.824*** (0.366)	1.904*** (0.244)	2.624*** (0.837)	1.888*** (0.643)
$D3$	0.518 (0.621)	0.675 (0.468)	3.135*** (0.742)	1.413*** (0.339)	3.658*** (0.329)	2.176*** (0.233)	3.417*** (0.806)	2.724*** (0.475)
$D4$	0.366 (0.552)	1.475*** (0.449)	3.735*** (0.680)	1.706*** (0.338)	3.054*** (0.353)	2.952*** (0.231)	3.026*** (0.867)	3.481*** (0.477)
$C2 \times D2$	-2.056** (0.864)	-1.516*** (0.564)	-3.127*** (0.959)	-1.652*** (0.521)	-3.819*** (0.738)	-1.502*** (0.450)	-3.327*** (1.228)	-1.713* (0.942)
$C2 \times D3$	-1.928** (0.793)	-1.206* (0.628)	-2.780*** (1.064)	-1.219** (0.559)	-4.309*** (0.913)	-1.408*** (0.520)	-3.656*** (0.910)	-1.925*** (0.583)
$C2 \times D4$	-2.439*** (0.724)	-1.284** (0.620)	-3.339*** (1.067)	-1.244** (0.569)	-4.021*** (0.857)	-1.040* (0.543)	-4.183*** (0.771)	-1.919*** (0.621)
$C3 \times D2$	-1.238 (0.764)	-1.060** (0.446)	-2.234*** (0.666)	-1.189*** (0.393)	-2.992*** (0.494)	-0.965*** (0.253)	-2.375*** (0.895)	-1.342** (0.636)
$C3 \times D3$	-1.378** (0.700)	-1.062** (0.474)	-2.328*** (0.611)	-0.915*** (0.332)	-3.246*** (0.468)	-0.864*** (0.251)	-3.004*** (0.839)	-1.521*** (0.517)
$C3 \times D4$	-1.675** (0.812)	-1.059** (0.494)	-2.636*** (0.674)	-1.028*** (0.344)	-3.148*** (0.483)	-0.810*** (0.283)	-2.798*** (0.751)	-1.343** (0.551)
$C4 \times D2$	-1.112 (0.842)	-0.707 (0.572)	-2.099 (1.446)	-1.160* (0.602)	-2.467* (1.468)	-0.969** (0.457)	-1.846 (1.134)	-1.057 (0.695)
$C4 \times D3$	-3.642*** (0.925)	-1.359** (0.620)	-3.968*** (1.440)	-1.354** (0.578)	-3.590*** (1.343)	-1.208*** (0.427)	-3.617*** (1.021)	-1.633*** (0.589)
$C4 \times D4$	-3.846*** (0.831)	-1.435** (0.620)	-4.335*** (1.379)	-1.507** (0.599)	-2.470* (1.482)	-0.949** (0.476)	-3.221*** (1.030)	-1.586*** (0.576)
$C0 \times D2$	-1.411*** (0.547)	-0.947** (0.419)	-3.730*** (0.611)	-1.782*** (0.518)	-3.877*** (0.427)	-1.509*** (0.293)	-3.381*** (0.847)	-1.986*** (0.653)
$C0 \times D3$	-1.006* (0.587)	-0.647 (0.431)	-3.372*** (0.565)	-1.197*** (0.345)	-3.561*** (0.397)	-1.171*** (0.284)	-4.012*** (0.853)	-2.024*** (0.519)
$C0 \times D4$	-0.482 (0.517)	-0.608 (0.417)	-3.598*** (0.550)	-1.262*** (0.345)	-2.726*** (0.428)	-1.053*** (0.278)	-3.391*** (0.959)	-1.895*** (0.531)
$mg_t$			-1.759*** (0.325)	1.511*** (0.180)				
$mg_t \times C0$			1.655*** (0.333)					
Constant	0.748 (0.611)	-3.355*** (0.477)	0.949 (0.890)	-6.367*** (0.404)	-1.035** (0.416)	-4.025*** (0.207)	-0.616 (0.669)	-4.299*** (0.375)
$\ln(\alpha)$	0.780*** (0.171)		0.825*** (0.209)		0.788*** (0.163)		0.581** (0.231)	
$N$	136036		133348		212092		91392	
$AIC$	55485.69		55477.80		71869.02		37279.63	
$BIC$	57214.13		56438.27		72844.17		38174.81	

Robust (clustered by pathology) standard errors in parentheses.

Therapeutic class and genetic dummy variables included in all specifications.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 9: Results – Robustness checks: sample issues and additional control variables.

particular, it includes the ratio between the producer price index of manufacturing pharmaceuticals and medicines, and the price index for private fixed investment in intellectual property products for firms operating in pharmaceutical and medicine manufacturing (as a proxy for R&D expenditures).<sup>23</sup> The ratio between the two indexes grew substantially over the observation period. Also included is an interaction term between  $C0$  (missing prevalence) and the ratio in the Gumbel part of the model.<sup>24</sup>

Column (3) includes the full set of diseases at all time periods, without the selection based on the stock of publications.

Finally, estimation in Column (4) is made on the balanced panel of diseases known at the beginning of the observation period (i.e., with a positive value of  $SP_{t-5}$  in year 1983). By using a balanced set of observations in Column (3) and (4), we aim to investigate whether these results are driven by sample composition.

All in all, the robustness checks carried out in this section confirm the main results in Table 4 (and Table 7).

---

price indices or dynamics in R&D costs, at the disease-level.

<sup>23</sup>Both indices were downloaded from Federal Reserve Economic Data. See: <https://fred.stlouisfed.org>. Data for the producer price index in 1983 are not available, so one observation for each disease was lost.

<sup>24</sup>This is the only interaction between class of prevalence and industry margin that is statistically significant either in the Gumbel or count part of the model. Note that, according to the theory, an increase in  $M_j$  works as an increase in  $z$ , meaning that the size of the impact depends on  $n_j$ .