Product recalls, market size and innovation in the pharmaceutical industry

Federico Nutarelli*[1] and Massimo Riccaboni†[2]

[1,2]AXES Group, IMT School for Advanced Studies
[1]Invernizzi Center for Research and Innovation, Bocconi University

Abstract

The idea that research investments respond to market rewards is well established in the literature on markets for innovation (Schmookler, 1966; Acemoglu & Linn, 2004; Bryan & Williams, 2021). Empirical evidence tells us that a change in market size, such as the one measured by demographic shifts, is associated with an increase in the number of new drugs available (Acemoglu & Linn, 2004; Dubois et al., 2015). However, the debate about potential reverse causality is still open (Cerda et al., 2007). In this paper, we analyze the effect of market size on innovation as measured by active clinical trials. The idea is to exploit product recalls as an innovative instrument, which is tested to be sharp, strong, and unexpected. This work analyses the relationship between US market size and innovation at ATC-3 level through an original dataset and the two-step IV methodology proposed by Wooldridge et al. (2019). The results reveal a robust and significantly positive response of the number of active trials to market size.

Keywords: Industrial Organisation; Pharmaceuticals; Recalls; Innovation
JEL Codes: O31, J10, J20, I11, L11

1 Introduction

Exploring the actual relationship between market rewards and innovation has been widely investigated in innovation economics (Scherer (1982), Schmookler (2013), Klepper and Malerba (2010)). This has opened the possibility to research public demand in a stimulating innovation, such as in the case of orphan drugs. Schmookler's "demand-pull" hypothesis, which implies that innovation is a function of market demand, has been challenged over the years. Even in the 1990s, Kleinknecht and Verspagen (1990) noticed that the direction of causality between market size and innovation appears to be far from obvious. In particular, the authors

*federico.nutarelli@imtlucca.it
†massimo.riccaboni@imtlucca.it

suggested the presence of a simultaneous relationship between demand and innovation, but did not manage to control for it. More recently, Stoneman (2010) and Ball et al. (2018b) developed more rigorous ways to detect this type of endogeneity.
Acemoglu and Linn (2004) developed a strategy to overcome the endogeneity bias at the market level. Specifically, they exploited changes in the market size for different drug categories driven by US demographic trends (Acemoglu and Linn (2004)). After the contribution of Acemoglu and Linn (2004), the focus moved from ascertaining the presence of the reverse causality of market size and innovation to detecting the best instrument for market size. Indeed, the instrument adopted in Acemoglu and Linn (2004) was later criticised by Cerda (2007) as being itself endogenous. As detailed in Cerda (2007), while pharmaceutical innovation increases the age of patients, the fact that the average age increases implies that more patients need innovative products. This scenario presents, again, the problem of reverse causality. Indeed, while demographic trends affect market size, which have an impact of innovation, the latter influences, in turn, impact on demographic trends.
To the best of our knowledge, this gap in the literature is still unfilled.
Inspired by the studies of Acemoglu and Linn (2004), most authors have concentrated their efforts on the pharmaceutical industry because it constitutes an ideal case study: in this sector the consumers' needs are diverse and almost constant over time, which allows the market to be separated into independent sub-markets based on such needs (Bertoni et al. (2010)). Furthermore, investments in innovation are vital for the industry's existence. As one of the major outputs of the pharmaceutical industry, innovation is also relatively more easily measurable. In the pharmaceutical industry, market size is defined based on the Anatomical Therapeutic Chemical (ATC) classification system; that is, a drug classification system classifying *"the active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties"* (Organization et al. (2009)).
The present work captures the relationships between market size and innovation at the ATC-3 level by instrumenting market size with recalls (see below) of drugs operated by the Food and Drug Administration (FDA). The present work contributes to the literature in three ways.
First, we adopt an innovative measure of innovation; that is, the overall number of trials at the ATC-3 level instead of the cumulative R&D expenditures or New Molecular Entities (NMEs). This necessary modification is able to overcome the limitations of the other two most adopted measures. R&D expenditures are linked both to a firm's long-term profit decisions (Cohen, 2010) and, more critically, to their size. Stoneman (2010), among others, suggested that smaller entrants might be more inclined to invest in R&D expenditures than their bigger veteran competitors. Hence, innovation as cumulative R&D expenditures of the firms composing the market might be related to market size, as measured by the firm's cumulative sales composing the market. The topic concerning NMEs is more delicate, and was also adopted in Acemoglu and Linn (2004). NMEs are innovative products that contain active moieties (i.e., significant parts of the molecule). These molecule parts were not previously approved by the FDA. Therefore, they can be innovative products that have never been exploited in clinical practice. Alternatively, they can also be related to previously approved products. Although a complete definition, the definition of NME does not fully capture, in our opinion, the will to innovate by firms inside the market. The reason hinges around the stage of drug approvals at which NME, with respect to the measure of innovation employed

in the present work, are approved by the FDA.
Pharmaceutical drug approval is a long process. Firms first pass a pre-clinical phase, a stage of research that starts before that clinical trials (testing in humans) can begin. During the pre-clinical phase, public agencies collect critical data on the feasibility of the trial, iterative testing, and the safety of the tested drug. The clinical drug development stage has three phases. In Phase 1, the company conducts clinical trials on healthy individuals. This process helps to determine the drug's basic properties and safety profile in humans. According to DiMasi et al. (2003), *"typically, the drug remains in this stage for one to two years"*. Once the drug is dispensed to volunteers drawn from the targeted population, Phase 2 begins. This phase consists, basically, of performing the trial on a larger group of individuals with respect to the previous phase. Finally, Phase 3 compares a new drug to a standard-of-care drug. Consequently, NMEs are FDA-approved entities that have overcome pre-clinical trials. The number of trials, which is adopted in the present work, also considers the pre-clinical phase. In other words, the latter measure also takes into account potentially unsuccessful trials (i.e., trials not passed to the clinical phase), which equally characterise a firm's innovative drive. The second contribution is given by the adoption of a more refined classification of the relevant market segments (i.e., ATC-3 classes). The available data on the ATC-3 level of classification well captures the structure of sub-markets, which are usually constructed artificially or disregarded by the literature. The literature, in fact, mostly uses ATC-2 and ATC-1 levels. The ATC-3 level is mainly employed for the sake of comparison with other broader levels.
Moreover, the ATC-3 level is employed by antitrust agencies. See Section 3 for further details.
We also make a methodological improvement. This work adopts an IV approach to deal with the endogeneity problem of market size. This enhancement compared to past research consists of the instrumentation of market size with recalls to overcome the endogeneity issue (which has already been detailed). The idea is to exploit sharp and unexpected recalls. The task of characterising recalls as being sharp and unexpected requires a general definition of recalls. We believe that recalls are exogenous. The idea is that, at the market level, firms cannot anticipate a recall being issued to a competitor. In Section 3, we provide several arguments in favor of the sharpness of recalls not being necessarily issued in "more risky" ATC markets.
The FDA refers to a recall as *"the most effective way to protect the public from a defective or potentially harmful product. A recall is a voluntary action taken by a company to remove a defective drug product from the market. Drug recalls are conducted either on a company's initiative or by FDA request"* (FDA U.S. Food Drug (2019)). In a recall, the FDA's role is to oversee a company's strategy, assess the recall's adequacy, and classify the recall. According to their severity, the FDA classifies recalls into Class I (more severe), and Class II and Class III (least severe). Medicines may be recalled for several reasons, ranging from health hazards to potential contamination, adverse reaction, mislabelling, and poor manufacturing. However, recalls should not be confused with withdrawals. Unlike the FDA's definition, the literature often refers to withdrawals as post-marketing recalls imposed by the FDA on firms due to their high severity and risk to human health. Therefore, recalls can be expected and voluntarily made by firms if minor or sharp and unexpected if most severe and forced by the FDA. In other words, according to the definition often adopted in the literature, withdrawals are post-marketing recalls and they are often operated on after a severe Class I

recall. Following Onakpoya et al. (2016), though there are discrepancies among countries, 72% of medicines in the recall procedure due to adverse effects ended up in a withdrawal [1]. To be consistent with our recall data, looking at major recalls, the majority of Class I recalls containing the word "death" within their root cause resulted in a withdrawal.
By the very definition of drug recalls, we expect that there will be a drop in sales consequent to a drug recall in a market. To clarify this mechanism, one can refer to Merck's popular recall of VIOXX in 2004. VIOXX was withdrawn from the market due to an increased risk of serious cardiovascular events. The recall caught both the market and the firm unprepared. After the announcement of the recall of VIOXX in September 2004, shares of Merck and its sales dropped. This drop has been publicised by the mass media (Terence N. (2004), Bowe C. (2005) among others) and is well-recognised by academics (see, e.g. Tong et al. (2009) among others).
In the present work, we try to assess such sharp and unexpected recalls (" major recalls" from now on). The definition of major recalls that we have adopted throughout this work has been recovered by filtering the causes of Class I recalls. We filtered recalls according to the relevance of the cause, its severity in terms of potential danger against human life, and the FDA's actions. Specifically, our definition of major recalls, withdraws, Class I recalls contains critical keywords among their causes, such as "contamination," "death/s," "overdose," "symptoms," "particulate matters," and "adverse reaction."
We did not employ Class II recalls because, in our opinion, they constitute a weaker instrument than major recalls. Nonetheless, we include the analysis using all types of recalls as a robustness check in Tab.7. The first column of Table 7 in the robustness checks shows how, adopting "minor" recalls (mainly Class II recalls whose motivations regard packaging or labeling issues) as an instrument for market size results in a poor instrumentation strategy.

2 Literature Review

The importance of market size in explaining the rate of innovation for many years has been acknowledged in the literature. In 1942, Schumpeter indicated that larger firms are more innovative than smaller ones. In the early 1960s, the focus shifted more broadly on the possible effects of demand on market size (see, e.g., Scherer (1982)). It was not yet clear whether the reverse causality of demand and innovation played a relevant role. Scherer (1982), for instance, argued that causality ran primarily from sales to innovation. However, this study has been criticised in several aspects. The definition of demand was still too broad and was not conclusive about the unique sign of the relationship between demand and innovation; that is, reverse causality (see, e.g., Mowery and Rosenberg (1979)). At that time, the research did not focus specifically on the pharmaceutical sector nor did it consider the aggregate market level (see, e.g., Pakes and Schankerman (1984)).
More recently, Kleinknecht and Verspagen (1990) denounced a clear reverse causality of demand and innovation, thus invalidating the prior studies. Soon after, Geroski and Walters (1995) empirically verified these conclusions and showed how innovations increase demand by creating their demand.

[1] This estimation was manually computed following the list provided in Onakpoya et al. (2016).

It was clear that heterogeneous shifts of demand played a prominent role in determining technological development (see, e.g., Malerba (2007)). Between 1980 and 1990, and most recently in 2002, several studies showed (for instance) how innovation reacted elastically to energy prices.

Nowadays, a large part of the research on the relationship between market size and innovation focuses on the pharmaceutical industry, where innovation represents a pushing power. The literature mainly takes into account the following two levels of aggregation: firm-level and market-level. Previous research efforts have been devoted to identifying the impact of firm size on R&D investments and output. Nevertheless, this question is still an open topic of debate (see Mellahi and Wilkinson (2010), Kolluru and Mukhopadhaya (2017) among others). Specifically, controversial results have emerged due to the difficulty in fully excluding unobservable endogeneity sources varying with time. These unobservables might derive from strategic decisions taken within the firms, which, in turn, might be related to their size. For example, small pharmaceutical firms are likely to take riskier decisions than large established firms (Hall and Rosenberg (2010)). Moving to market aggregation easily avoids these concerns. Unobservables related to market size can in principal be considered as intrinsic characteristics of markets and, consequently, fixed in time. Thus, fixed effect techniques allow researchers to control for unobservable heterogeneity, thus purging the idiosyncratic endogeneity of market size. Therefore, the market seemed to be a more suitable level, and most authors shifted to the latter level of aggregation.

The literature on the pharmaceutical sector is vast and part of its variability is due to the measures of innovation adopted. Some authors have adopted accounting data focusing on R&D. While R&D is robust under perfect capital markets, it becomes inconclusive with imperfect markets where current investment choices reflect the future choices. Within market imperfection, current revenues (market size) are a reasonable proxy for future market size. This might at first cause endogeneity problems due to the correlation of current revenues with unobservables (e.g., the risk propensity of the firm's management). Moreover, given that present R&D may be responding both to present and future sales opportunities, the coefficient of market size might incorporate two effects that are difficult to separate. In light of this issue, authors have included lagged proxies of the market size (see, e.g., Giaccotto et al. (2005), who estimated that a 1% increase in price leads to a 0.58% increase in R&D spending). Other problems related to R&D measure are reported in Hall and Rosenberg (2010). To give an idea, several authors (see e.g. Pammolli et al. (2011) among others) showed how although pharmaceutical firms made substantial investments in R&D, they did not produce innovation [2].

Other measures of innovation include clinical trials (see Kyle and McGahan (2012) among others) and changes in Medicare part D. Medicare part D is an optional US programme to help beneficiaries of the Medicare national health insurance to pay for self-administered prescription drugs. The use of Medicare part-D as an innovation measure might affect both present and future market size (Blume-Kohout and Sood (2013), Dubois et al. (2015)). As suggested in Blume-Kohout and Sood (2013), *"Medicare Part D could have affected firms' R&D expenditures both to its expansion of expected future markets for products still in pipeline, and also via two supply side mechanisms."* (see Blume-Kohout and Sood (2013) for further

[2]Thus undermining the validity of R&D investment as a good proxy for innovation.

details). Thus, Medicare part-D might have stimulated current and, critically, future sales through R&D expenditures. Scholars have found that innovation has a positive response to shocks in market size. Again, the problem remained the possible co-occurrence in innovation's response to both current and expected cash flows generated by market size shocks. It was therefore impossible to distinguish between current and future effects within the estimated coefficient of market size. In the present work, we mitigate this issue by taking all active projects having generated innovation and failed projects in all of the clinical and pre-clinical phases. Thus, our innovation measure only affects current market size. The future market size is partially influenced by currently active (and failed) projects. The latter hypothesis has been tested in the robustness checks (where the lag of innovation has been inserted as a covariate).

Although innovation has been quantified by the number of relevant medical journal articles (Lichtenberg (2006)), they do not account for innovations undertaken exclusively in industries. Further measurements include the number of new drugs launched, including generic drugs (Acemoglu and Linn (2004), Dubois et al. (2015)) in the form of NMEs, New Chemical Entities (NCEs), or approvals of new medicines by the FDA.

Similarly, many measures of market size have been embraced.

Acemoglu and Linn (2004) made the first significant contribution to the relation between market size and innovation in the pharmaceutical industry. Their idea relies on adopting demographic shifts to instrument market size, while controlling for observables or unobservables arising from reverse causality. In particular, Acemoglu and Linn (2004) exploit variations in the expenditure share of different US age cohorts for different therapeutic classes from 1970 to 2000. They discover that a 1% increase in the shares of expenditure would lead to a rise of 4% in the number of new medicines, which is a far higher elasticity than the average elasticity that is found in the remaining literature (Dubois et al. (2015)). Cerda (2007) provided further insights on the results found in Acemoglu and Linn (2004). Employing US demographic data, Cerda (2007), showed that there are essential feedback effects that were not considered in Acemoglu and Linn (2004). For example, new drugs might affect the market size through their impact on the mortality rate. Indeed, innovative medicines are likely to cure more diseases, raising the population's average age and, hence, the number of older people needing such cures. Demand shifts accordingly again raise the issue of reverse causality.

Above all, the recent literature on the topic improves the methodological part (see e.g. Lichtenberg (2006), Civan and Maloney (2009), Dubois et al. (2015), Rake (2017) and others). Dubois et al. (2015)'s novelty pertains the usage of global pharmaceutical data at the ATC-1 and 2 levels. Dubois et al. (2015) still employs demographic shifts as an instrument for market size. Civan and Maloney (2009) found that *"the higher the prices of existing drugs in a therapeutic category, the larger the number of drugs in the development pipeline in that therapeutic category"*. It is important to note that Civan and Maloney (2009)'s work is more focused in estimating the elasticity of drug development to the market price of drugs. Their results are conducted using R&D as a measure of innovation and they suffer from the endogeneity of prices in the elasticity equations adopted to reach the results. Finally, Rake (2017) adopts a unique database and a Poisson Quasi Maximum Likelihood approach to reach his results. His contribution is also related to the usage of NMEs and New Drug Approvals (NDA) as measures of innovation.

Authors have found that, on average, a 1% increase in the market size measure increases

innovation of 0.4% to 0.7%.
Previous papers have mainly acted at disease level or, at most, at ATC-1 or ATC-2 levels of aggregation (see e.g.Dubois et al. (2015)). To the best of our knowledge, no studies have focused on the more interesting ATC-3 level, at which antitrust authorities work. According to our work, there are several advantages of using drug classes rather than disease classes. First, because firms directly make the request to undergo a New Clinical Trial (NCT), NME, or NDA, devoting too much attention to the demand-side might neglect the supply-side dynamics, which induce firms to undergo an NDA, NCT or NME. In particular, the aggregate sales of drug classes align to the supply-side dynamics, while sales based on disease classes (i.e., aggregated sales of products purchased by patients) are more related to the demand side.
In other words, while firms might follow demand-side stimuli to undergo an NDA (or an NCT), they (above all) look up at the competitors (i.e., products of other companies in the same ATC class). This applies to commercial trials when the sponsor is a pharmaceutical industry and is not academy/research related. Thus, aggregating sales into disease classes forces them to be clustered in a demand-driven group, which is less informative on the supply-side dynamics (i.e., behavior of competitors), which leads the firm to undergo an NME, NDA or NCT. When estimating the effect of market size grouped into disease-classes (demand-driven) and innovation (supply-driven), one might incur in an underestimation of the latter effect.
Furthermore, by taking disease classes, one includes in the definition of innovation several different chemical and therapeutic typologies of drugs, ranging from topical to systemic drugs, and from vaccines to ointment. This lack of distinction might lead to endogeneity through several channels, such as people's expectations. The patients may be wary of some drugs, which affects the probability of having a larger market size for the product's typology under question. Other endogeneity sources may regard the possibility of a correlation between regressors and the error term (which includes "drug-type"). For instance, regulations may be product type-specific (e.g., the regulations of the WHO vaccines do not apply to other drug types). Knowledge stocks may also be problematic controls, which could again depend on the product type. Moreover, knowledge stocks might increase by developing innovative medicines in classes where only a particular type of medicine has been developed until that moment. An example may be found provided in dermatology, where academics produce papers on adopting topical medicines for systemic usage because some of the systemic medicines have undesired side effects.
Finally, the length of a clinical trial varies depending on the type of medicine under study, which may cause lagged effects of market size if disease class is employed.
To the best of our knowledge, among the innovation measures, no work has exploited INDs and early-stage clinical trials (i.e., pre-clinical and Phase I) together with Phase II and Phase III trials.
The two more recent estimates of the relationship between market size and innovation have been provided in Rake (2017) and Dubois et al. (2015). The latter used NCE to measure innovation and defined market size as a measure of expected revenue. Their dataset included information about sales for 14 different countries. Specifically, Dubois et al. (2015) measured market size as the total revenue over the entire life cycle of a branded drug. Dubois et al. (2015) performed a control function approach and recovered an estimate of the relationship

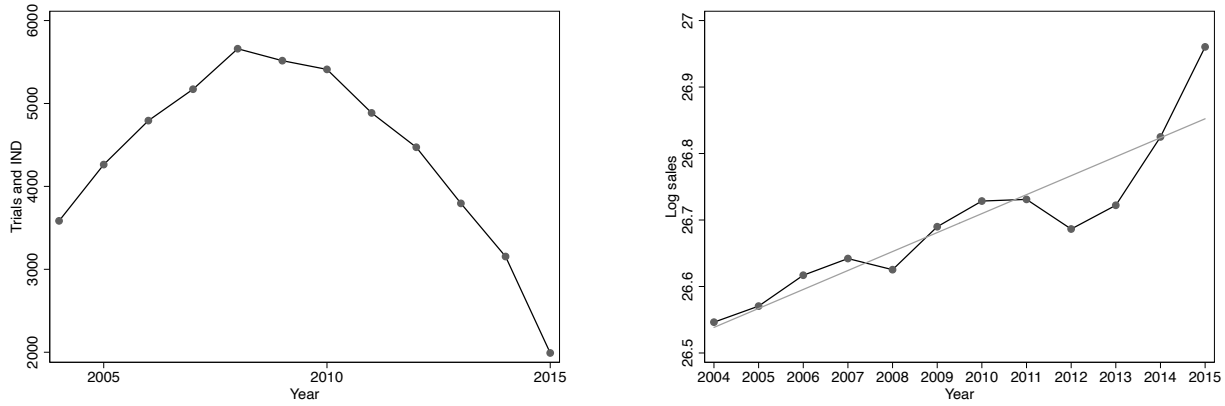
between market size and innovation for each therapeutic class at level 1. The average elasticity of innovation to the market size in Dubois et al. (2015) was about 23%, which is relatively lower than the average estimates. A possible explanation for this may be found in Blume-Kohout and Sood (2013), which states that several of the countries chosen for the analysis regulate prescription drug prices and that regulations may change rapidly over time. Thus, given the lower expected profit per consumer and more significant uncertainty about future profits and prices, a firm's R&D decisions are likely to be less responsive to a unit change in expected revenues for all of these countries combined versus the exact unit change in the US market.
Finally, Rake (2017) adopted several measures of innovation from NCE to clinical trials in Phase II and Phase III. Rake (2017) found no evidence of reverse causality when adopting NCE. One of his efforts was to account for changes in the industry's R&D process, from "random screening" to "guided drug development" (Rake (2017)). He modeled technological opportunities and inserted them as regressors in the analysis, finding a positive relationship with Phases II and III trials. His results are in line with Cerda (2007) and Acemoglu and Linn (2004).

Tab. Appx.1 in the Appendix provides a schematic literature review of previous estimates of the relationship between innovation and market size.

3 Data

The sales data that we have employed comes from the Evaluate dataset. The controls have been extrapolated from Evaluate, the PHarmaceutical Industry Database (PHID), and the FDA. Specifically, some of the regressors derive from an elaboration of the variables that are present in the PHID database.
Sales data for the US pharmaceutical market ranges from 2004 to 2015. Sales data were initially available at the product and molecule level, and have successively been aggregated at the ATC-3 level. In the ATC classification system, drugs are classified at five levels (i.e., ATC-1, ATC-2, ATC-3, ATC-4, ATC-5): the higher the level, the more detailed the classification. Acemoglu and Linn (2004) employed ATC-1 and ATC-2 categorisations to define market size. In particular, they constructed market size as the sum of the average expenditure share of drugs in an ATC-1 (ATC-2) category across all ATC-1 (ATC-2) categories.
The data that are at our disposal allow us to catch the diverse strata of products inside broader classes (ATC-1 and ATC-2) in terms of both demand and supply dynamics. Medicines classified inside an ATC-1 or an ATC-2 level can satisfy patients with completely diverse needs because they are designed to cure various diseases. Meanwhile, a firm investing in the same ATC-2 sector might invest in more ATC-3 sectors. In the case of ATC-1 or ATC-2 adoption, the latter missing information may lead to the construction of uninformative innovation and market size variables. These controls might not consider the firm's specialisation in a sub-sector rather than in another sub-sector belonging to the same ATC-2 or ATC-1 class.
In our previous work, we evaluated other levels of analyses (i.e., firm, product, and ATC-firm aggregations; Nutarelli (2021)) but we have opted for the ATC-3 level in this work because of the importance of ATC-3 level being employed by antitrust agencies. To provide some examples, we mention Provost et al.(2019), Markham (2020), Vaishnav (2011), Hawk et

al.(2000), Cheng (2008), and other cases, mostly pertaining to M&A (e.g., Case M.8889 - TEVA / PGT OTC ASSETS of 2018).
We have avoided adopting the ATC-4 level because at this granularity products belonging to a specific ATC-4 class might not differ substantially from others belonging to another ATC-4 class. This might lead to between-group dependencies (e.g., innovations in an ATC-4 at level 4 may also affect a close ATC-4 class) which could cause any inference to be invalid. Furthermore, at the ATC-4 level, compensation may also intervene between groups. This would invalidate the strength of the instrumental variable recalls.
The available data also contain the product's launch date and ATC code. In this paper, we have focused on worldwide sales of US companies.
Data on NCTs for 2004 to 2015 at the product level come from the ClinicalTrials.gov website, while data on commercial Investigational New Drugs (IND) at product level derive from a Pharmaceutical Industry Database, which is maintained at IMT Lucca.
Clinical trials are research studies that are performed on people to evaluate a medical, surgical, or behavioral intervention. An IND in clinical trials is the means by which a pharmaceutical company obtains permission to start human clinical trials and to ship an experimental drug across state lines before a marketing application for the drug has been approved.
Clinical trials move from Phase I to Phase IV. Fig.1 displays the yearly number of trials and commercial INDs as obtained by the mentioned sources.
It also shows the expected positive trend of the pharmaceutical industry's sales in time.



(a) Number of trials and INDs per year

(b) Trend of sales by year

Figure 1: Overview of the trends for sales and trials

A considerable drop in trials and IND occurred after 2013, as is evident from Fig.1 (a). The reason for this drop is that, in general, clinical trials innovate drugs, approaches, and interventions. However, approaches and interventions are excluded from the count of trials to focus strictly on innovation coming from industrial sources.

Recall data have been manually collected from different sources, including the FDA website, openFDA, various articles, and web sources (e.g., Onakpoya et al. (2016); WHOCC website, PubMed, Siramshetty et al. (2016) and others).

In total, 7.19% of the firm sample (i.e., 697 firms in total) have issued a Class II recall. Among the firms that have issued a recall, 51 firms underwent a recall of Class I, 27 of which issued a single recall of Class I, and just three firms issued more than nine Class I recalls.
The estimates that are provided must be read in the light of the database's possible limitations regarding the presence/absence of firms and products inside it.
In addition, the recalls of pure compounders were only partially included [3] and, when included, were attributed to the unique manufacturer/distributor in the database. Finally, recalls coming from repackaging firms were not attributed uniquely to the repackager (e.g., Aidapak) but the labeller specified in the National Drug Code (NDC).
Due to the need to exclude the recalls coming from repackagers and compounders, it is not easy to establish a unique and unambiguous pattern of recalls over the years. The situation is further complicated wherever different resources employ different methodologies to count the recalls. An example of the cited uncertainty in sources can be found when comparing FDA Enforcement Statistics (2015) and Laguna Hospital (2019). Specifically, FDA Enforcement Statistics (2015) assert that the number of recalled products had remained reasonably constant, except for 2010 and 2013 when the number went down by approximately 35%. This statement contrasts with the reports from Laguna Hospital (2019), as follows: *"a spike in the number of drugs recalled occurred in 2013. There were nearly 60 recalls in that year alone. However, 2017, with 71 recalls, saw nearly the highest number of recalls since 2009. Only 2011 and 2009 surpassed it at 74, and 75 recalls, respectively"*.
Tab. 1, provides a list of the primary sources and the average number of recalls across them. The following is an attempt to overcome any dissimilarities in the data origins.

Year	CNN	Regulatory Focus	Hall et al. (2016)	FDA Enforcement Reports	Laguna Hospital (2019)	AVERAGE
2004		68				68
2005		140				140
2006	384	109				243
2007	391	56				189
2008	426	128		176		244
2009	1742	85		1660		890
2010		135		389		262
2011		236		1279	75	530
2012		381	499	1518		799
2013		1031	1283	848	60	805
2014		640	1344	893		959
2015				1584		1584

Table 1: Consulted sources with reported number of recalls

To overcome any dissimilarities in the data origins, we chose the average as the benchmark to compare with the collected recalls. Fig.2 illustrates in more detail the comparison between the benchmark recalls represented by the average recalls among all sources and the collected recalls. Aidapak's recalls of 2011 are considered to be "outliers" and, for this reason, are not included among the collected recalls at this stage:

[3] Due to the unavailability of data. We verified that the representativeness of recalls is preserved Nutarelli (2021).

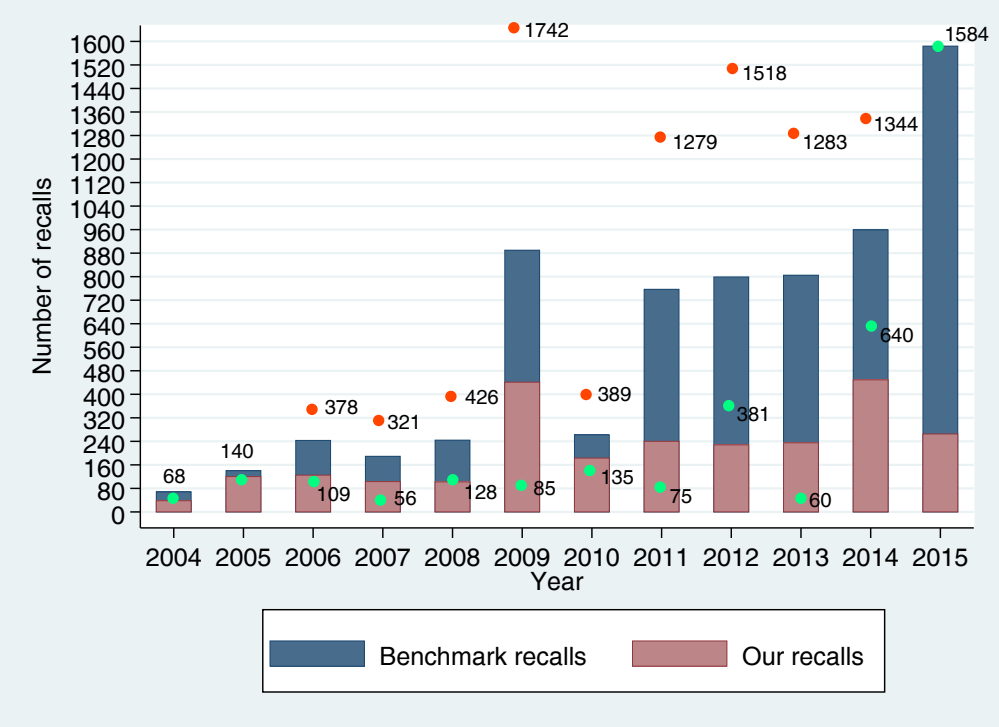


Figure 2: This histogram compares our number of recalls against the number of recalls used as a benchmark (average of sources). We include the minimum and the maximum number of recalls retrieved by the different sources. Mint colored points represent the minimum amount of recalls retrieved among all of the sources at our disposal. Red points represent the maximum number of recalls among all of the sources. A single mint point has been put whenever a single source was present for a year (i.e., 2004, 2005, and 2015).

Fig. 2 underlines a disproportion in terms of the number of recalls starting from 2009, with respect to the benchmark. This deficiency pertains to the counting methodology and the structure of the database (see earlier).

Although the global trend is approximately reproduced, 2011, 2013, and 2015 represent problematic years. The dissimilarity of 2011 concerning the benchmark can easily be explained. Indeed, when excluding Aidapak's recalls from the count of the collected recalls, the latter dropped. Furthermore, 2013 and 2015 have far fewer recalls than expected because more than 60% of the recalls in 2013 and nearly 75% of the recalls in 2015 were represented by compounding firms.

The recalls trend of the benchmark seems to be well reproduced. However, when recalls of pure compounders are excluded from the benchmark number and the sample of collected recalls, Fig. 3 shows an accordance in trends. Aidapak's recalls are included here. It is important to note here that Aidapak is a repackager and not a compounder. In addition, we want to show that 2011 does not constitute a problematic year once Aidapak's recalls are considered.

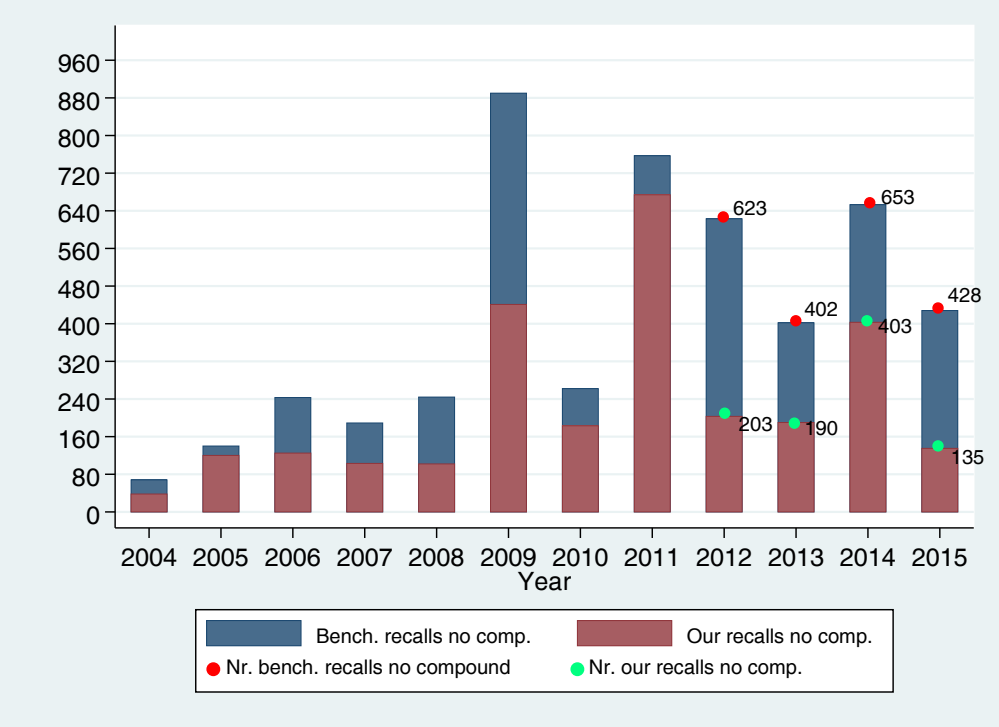


Figure 3: This histogram compares our number of recalls against the number of recalls used as benchmark without compounding recalls. We included the minimum and the maximum number of recalls retrieved by the different sources. Mint colored points represent the minimum amount of recalls retrieved among all of the sources at our disposal. Red points represent the maximum number of recalls among all of the sources.The situation is almost unchanged with respect to Fig. 2 until 2011. From 2011 on, the recalls that are collected in our dataset follow the benchmark if the compounder's recalls are excluded more precisely.

As a further check of the exogeneity of recalls, we constructed a box plot that displays the average number of trials (and their dispersion) in both ATC markets having undergone a recall and not having undergone a recall by year. The latter exercise helps us to understand that major recalls do not necessarily intervene in more innovative markets. Indeed, Fig. 4 shows that the yearly number of trials of ATC markets undergoing major recalls almost coincides with the average number of trials in all other markets.

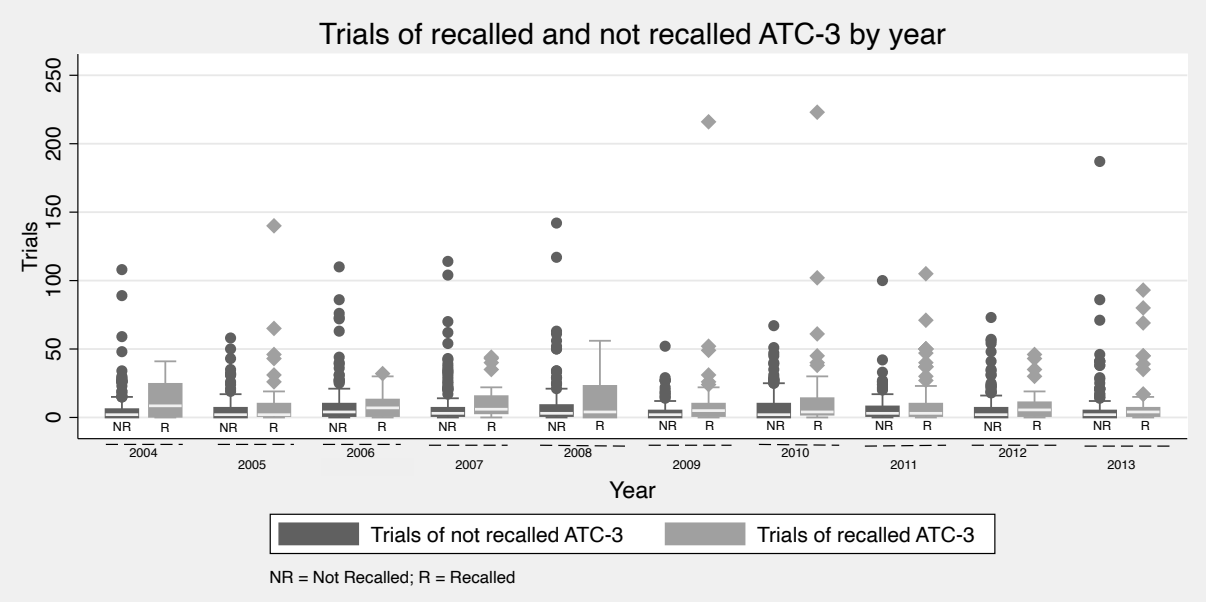


Figure 4: The box plots show how, on average, recalls do not necessarily happen in more innovative markets. The average number of trials amounts to 123 for ATC-3 markets having undergone a recall and to 118 for ATC-3 markets not having undergone a recall.
This graph reports a yearly analysis of the average number of trials in recalled and not recalled ATC-3 groups. The average number of trials is similar in both the ATC-3 markets, having undergone at least a recall (R) and ATC-3 markets not having undergone any recall (NR).

To conclude, Fig. 5 provides further evidence for the possible relation between innovation and recalls, which was raised in the introduction.
In particular, from Fig. 5 it is evident that the percentage change (increase or decrease) in the number of activated trials [4] after the year of recall is minimal and in line with the average change in the two years before the recall. Note that (in this context) we mentioned the activated trials in that, by construction, the difference between the active trials in t and the active trials in the previous period determines the number of activated trials in t.

[4]Constructed as $ln(Trials_t) - ln(Trials_{t-1})$.

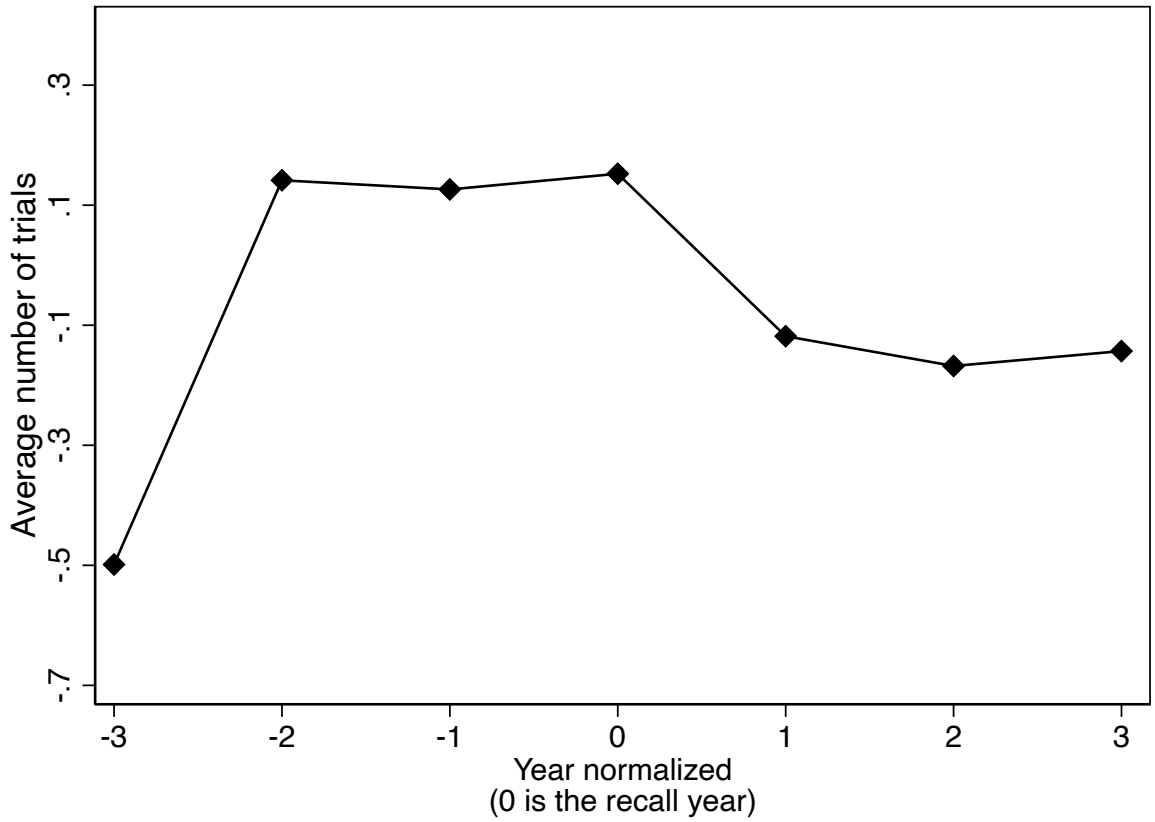


Figure 5: Change in the number of trials as measured by $ln(Trials_t) - ln(Trials_{t-1})$.

4 Methodology

The main theoretical framework is the same as that adopted in Acemoglu and Linn (2004). In particular, Acemoglu and Linn (2004) model innovation, which is the dependent variable of the present model, as being proportional to market size. We refer the reader to Acemoglu and Linn (2004) for further details.

The measure of innovation is the number of clinical trials in all phases for the ATC-3 category i. The measure of market size is the sum of the product's sales for the i^{th} market. We added further potential determinants, time effects, and category effects to the analysis. The theoretical model that is returned is the well-known estimation Poisson model:

$$E[N_{it}|\mu_i, \zeta_t, X_{it}, M_{it}] = \exp(\beta_1 \cdot \log M_{it} + \beta_2 \cdot X_{it} + \mu_i + \zeta_t) \quad \forall i = 1, \dots N, t = \dots T \qquad (1)$$

where E is the expectations operator; M_{it} represents endogenous market size; X_{it} captures age (e.g., the average age of products in category i weighted for the product's size), diversification and innovation patterns (e.g., the scientific production); μ_i denotes the ATC fixed effects and ζ_t denotes the time fixed effects. The estimation of (1) would lead to biased estimates for two reasons: first, the non-linearity in (1) makes it impossible to estimate the fixed effects consistently; and second, market size is endogenous.

To deal with both problems, a novel control function (CF) IV approach, as described in Lin and Wooldridge (2019), has been adopted. With respect to the literature, the present method allows us to simultaneously deal (i.e., testing and estimating) with two potential sources of endogeneity: the first source is due to correlation of covariates with time-constant, unobserved heterogeneity; and the second is due to the correlation of covariates with time-varying idiosyncratic errors. Furthermore, this can be easily extended to non-linear scenarios with fixed effects.

Specifically, κ_{it} denotes the idiosyncratic shock and c_i denotes the individual heterogeneity, and therefore the unobserved effects non-linear model allowing for both idiosyncratic endogeneity and heterogeneity endogeneity might look as follows:

$$E[N_{it}|M_{it}, z_{it}, c_i, \kappa_{it}] = c_i \exp(x_{it}\beta_1 + \kappa_{it}) \qquad (2)$$

where $x_{it} = (M_{it}, z_{it})$. z_{it} would typically include a full set of time effects, and M_{it} is the endogenous variable. All of the exogenous variables, which include the vector z_{it}, can be correlated with the heterogeneity (i.e., no random effects). There is also a set of excluded exogenous R_{it2} that serves as an instrument for the potentially endogenous variable. In the present work, R_{it2} is represented by recalls. Lin and Wooldridge (2019) noted that, without idiosyncratic endogeneity, an appealing estimator would be a fixed-effects Poisson estimator, which, when viewed as a QMLE, would only require a strict exogeneity assumption to ensure consistency with respect to the idiosyncratic shocks. This assumption is exploited as a null hypothesis for testing idiosyncratic endogeneity against the alternative of full dependence of the error term of the specification of M_{it} and κ_{it}. The alternative is to exploit the reduced form equation for the endogenous variable

$$M_{it} = z_{it}\Pi + c_{i2} + u_{it2} \quad \forall t = 1, \dots T \qquad (3)$$

where because the z_{it} is strictly exogenous, the correlation between κ_{it} and functions of u_{it2} is tested. Lin and Wooldridge (2019) developed a simple procedure that allows us to test for

idiosyncratic endogeneity and produce consistent estimates in the co-presence of non-linearity, fixed effects and both types of endogeneity. The algorithm follows the steps below:

1. Estimate the reduced form for the endogenous through fixed effects and obtain the fixed effects residuals $\ddot{u}_{it2} = \ddot{M}_{it} - \ddot{z}_{it}\hat{\Pi}$

2. Use fixed effects Poisson on the mean function

$$E[N_{it}|M_{it}, z_{it}, c_i, \ddot{u}_{it2}] = c_i \exp(x_{it}\beta_1 + \ddot{u}_{it2}\rho)$$

   use the robust Wald test of $H_0 : \rho = 0$

The described procedure allows us to consistently estimate the fixed effects in the presence of non-linearity. The fixed effects Poisson enables the elimination of the ATC-level fixed effects, hence performing a conditional Maximum Likelihood (ML) consistent estimation. We refer the reader to Cameron and Trivedi (2013) for further details.
A crucial characteristic of Poisson-FE models is that they require the dependent variable to be nonzero for at least one time period. The lower the proportion of zeroes in the dependent, the better the model works. The last condition has been fulfilled by dropping those ATC categories that do not meet it, constituting approximately 10% of the total ATC-3 in the sample.
We estimated several instances that are common to literature to check for either delayed effects of trials or the presence of a bias if market size were considered exogenous (or fixed effects omitted).

Throughout, the problem of endogeneity in market size has been exposed as being intrinsic to market size. Hence, instrumentation of the endogenous M_{it} is required. Market size is instrumented through normalised recalls. The normalisation is on the number of products present in the market i at time t. Calling m the major recalls, the normalised recalls are denoted as follows:

$$\tilde{m} = \frac{m}{\#prod.} \cdot 100.$$

As aforementioned, normalisation is necessary to avoid another source of endogeneity. Indeed, ATC markets with more products are more likely to undergo a recall by definition. Omitting this control would partly invalidate the estimates. The belief is that markets which undergo major recalls will tend to experiment with a sudden negative shock in sales. The relevance of the instrument is tested in Section 5.
The instrument is not directly related to the dependent variable. The central argument that might directly connect normalised recalls to trials is that the lack left by recalls is filled with innovations. Hence, sectors that are more prone to undergo a recall should also be the most innovative. In the literature, there seems to be contrasting evidence about this topic. Although the argument would imply a positive impact of recalls on innovation, the recent events seem to contradict these findings. Indeed, although there has been an increasing number of recalls from 2004 to 2015 (see, e.g., Fig.2), the innovation crisis of the pharmaceutical industry is a widely known and recognised phenomenon in the literature (see e.g.Pammolli et al. (2011), Price and Nicholson (2014) among others). It might be argued that the contrasting effects leading to the drop of innovation have overtaken the

stimulation of innovation by recalls, thus favoring the decreasing pharmaceutical innovation trend. Therefore, the positive effect of recalls on innovation could still be present but may be hidden. Empirical research has conducted few analyses to explore the relationship between innovation and recalls, or withdrawal in general. The most severe recalls have received considerable media coverage, which has allowed researchers to collect data on market reaction to such bad events (see, e.g., Pérez-Rodríguez and Valcarcel (2012)). Authors working on this stream of literature have concluded that the impact of recalls and withdrawals on market innovation has a high variability: some recalls have had considerable effects, while others have had none at all. There does not seem to be a systematic way to identify the recalls whose announcement impacted innovation among major recalls. Market reactions depend mostly on the period during which the recall took place and eventual delays in the FDA's communication of the recall. Generally, however, the market does not systematically overreact to such shocks, which invalidates any dependence between recalls and innovation.

In summary, direct connection sources between innovation and recalls are mainly due to fixed and time effects. FDA delays cannot be easily controlled. The FDA developed precise guidance and protocols for recall communication and announcement for the period considered in the present work. Hence, delays constitute a minor issue because the FDA regulates them. For the sake of completeness, thanks to the FOIA agreement signed, openFDA, and FDA enforcement report, it has been possible to verify the occurrence of delays. These sources have allowed us to access the time gaps between recall initiation, recall classification, and recall termination. The communication of the recall is part of the initiation process. Above all, in the case of severe recalls, it must be prompt. The average time between the initiation and the termination for Class I and Class II recalls has been around 23 months. A delay in communication might happen in the first initiation phase. The average time that the initiation phase took for any Class I and Class II recall was approximately four months. For our sample of major recalls, the average time of the initiation phase has been approximately 2 to 3 months, which is in line with prompt communication criteria. This evidence enforces the limited impact of delays on the analysis.

By dropping out unobserved heterogeneity and including time dummies in the primary specification, we can control for possible direct connections between recalls and innovation. Thus, the mentioned operations ensure that there is only an indirect effect of recalls through sales.

Further arguments in favor of the indirect effect of recalls on innovation follow.

In particular, the recalls that are taken into account are severe recalls of marketed products. The time gap between trial phases and the marketing of a drug usually takes between 8 to 14 years. This significant time gap is relevant to guess and understand a competitor's possible reactions to a drug recall in the same sector where a firm is operating. We believe that a competitor that underwent a recall in the sector in which both firms operate does not increase nor decrease the risk of innovation in the short run. Indeed, marketed products undergo major recalls long after they have been commercialised. In addition, the market requires an extended period to fully recover from the lack of sales following the recall of a drug. Hence, there is no need to invest in clinical trials to take advantage of such a shortage in the short run. As a further check of the latter conjecture, we build a time-to-event analysis in the Fig.Appx.1. Fig.Appx.1 takes into account all types of recall and it clearly shows how, although having undergone a recall decreases the survival probability of a drug, the

probability that a recalled drug survives two years is still consistent. The median survival time is five years. Thus, after a recall, the drop in sales is likely to remain unfilled for years. Indeed, had firms found innovative replacements for recalled drug d, which allowed them to recover the shortages left by the recall of d, then there would be no reason to keep selling drug d for years. Therefore, recalled products leave a long-term lack in terms of sales within the ATC-3 market to which they belong.

A further argument against the coverage of lacks left by recalls through innovative products is that these shortages might be filled by drugs that are already present in the market, whose trials started before, soon after, or at the same time as the trials leading to the recalled drug. This eventuality is reasonable because suspended or terminated studies are excluded from the sample, which means that the remaining clinical trials sponsored by concurrent firms are likely to arrive on the market with products belonging to the same therapeutic class. Competition between wholesalers within an ATC might be revealed in early stages once it is evident that a firm will develop an innovative cure. The development of alternative drugs is encouraged from the early trial phases, when there are still opportunities to arrive first in the market. Medicines substituting recalled drugs in the same ATC might be developed soon after the recalled medicine in a "first to arrive" competition, rather than a "fill the gaps of recalls" logic. The latter may also happen because the demand for patented medicines of the type of the recalled drug was likely to be more consistent when the trial for the recalled drug started. In the eventuality that demand propagates at the recall's time, either already existing generics or new ones (note that trials of generics is less time consuming because they only need to ensure bio-comparability) might intervene and fill the gap.

It is worth noting that the potential positive relationship of recalls and innovation exploiting lacks in the market passes indirectly through market size. Indeed, the emergence of new trials within a market after a recall depends on the demand that the product in question generated in the market. If a recalled product had no underlying demand, then it is reasonable to expect that no company will begin an expensive trial only to fill the lack left by the recalled product. Therefore, the response of innovation seems to depend not directly on the recall but on the underlying magnitude of the recalled product's demand (i.e., on market size).

Finally, another possible critique undermining the instrument's validity is that a recall of product i might have provoked the recall of trials concerning similar products. This domino effect hangs on the causes of the recall. Indeed, if the recall concerns only the specific product being withdrawn from the market, then implications for other companies' products are unlikely. For instance, it is possible that after the recall of the COX-2 inhibitor, Vioxx, due to cardiovascular side effects, all firms having ongoing trials on the same target suspended or withdrew trials relating to COX-2 inhibitors. To the best of our knowledge, no effort has been made to explore this possibility in the drugs market. The only work approaching the critique is Ball et al. (2018a). However, the author focuses on the medical device industry, which has different legislation for recalls than the drug market. Indeed, the recall of a device is a common practice that is usually made by firms who wish to make a repair or an update. The device is usually promptly placed back onto the market. The way we that managed this circumstance is threefold. First, we considered only active trials, thus excluding suspended and withdrawn trials (i.e., also those suspended due to the recall of other drugs). Second, we removed trials of companies undergoing a recall. Ultimately, as far as it has been possible to

link the reason for the severe recalls [5], we dropped trials adopting a similar active principle. The latter instance happened in a few cases.

[5] Mainly in the case of adverse events caused by an active principle adopted in the drug to the scope of a trial.

5 Results

This results section is divided into two main subsections. First, the impact of recalls on the endogenous market size is analyzed as measured by total sales of ATC i. Our aim is to provide convincing arguments in favor of the relevance of the adopted instrument.
Second, the results of the impact of the instrumented market size on innovation are presented.

5.1 The impact of recalls on sales

5.1.1 Summary statistics

This section will report the summary statistics for the sample. Tab.2 contains average values and standard deviations of the relevant variables for the full sample and two separate sub-samples for observations associated (or not) to recalls. This table includes information at the ATC-3 level and refers to major recalls. Tab.2 details all of the relevant controls that were employed to construct Tab.7.

Table 2: Summary statistics at the ATC-3 level for the full sample, the subset of ATC-3 having undergone a recall in the period considered, and the subset that has not undergone a recall. Database at ATC-3 level is balanced.

		ATC-3			
Variable		Full Sample	Subs. recalls	Subs. no recalls	**Description**
Sales (log)	Overall mean	19.405	20.794	19.054	Log of sales at ATC-3 level.
	Overall Std. Dev.	2.233	1.475	2.256	
	Between Std. Dev.	2.152	1.441	2.163	
	Within Std. Dev.	.614	.378	.661	
Outflow rate ($\frac{K_{t+1}}{P_{-1}}$)	Overall mean	.086	.056	.093	This is defined as the number of lost products in an ATC-3 (K_{t+1} in regressions) over the total number of products in $t-1$ (P_{-1} in regressions).
	Overall Std. Dev.	.257	.064	.285	
	Between Std. Dev.	.109	.036	.120	
	Within Std. Dev.	.233	.053	.259	
Avg. age of firms within ATC	Overall mean	35.907	33.275	36.573	This is the average age of the firms competing within an ATC-3. The foundation year of the firms was present in the data.
	Overall Std. Dev.	7.631	5.420	7.960	
	Between Std. Dev.	6.767	4.452	7.093	
	Within Std. Dev.	3.556	3.160	3.650	
Herfindahl–Hirschman Index (hhi)	Overall mean	.431	.268	.434	The hhi measures the competition within a market. It can range from 0 to 1.0, moving from a huge number of very small firms to a single monopolistic producer.
	Overall Std. Dev.	.260	.159	.262	
	Between Std. Dev.	.236	.166	.240	
	Within Std. Dev.	.110	.020	.115	
Share generics by ATC	Overall mean	.746	.725	.752	This represents the percentage of generic products, among all products sold in an ATC-3 market
	Overall Std. Dev.	.255	.214	.264	
	Between Std. Dev.	.238	.210	.245	
	Within Std. Dev.	.092	.052	.099	
Avg. age prod. by ATC	Overall mean	13.159	12.043	13.441	This represents the average age of product within an ATC-3. The age of a product is based on the foundation year of the firm that produced it.
	Overall Std. Dev.	5.333	3.763	5.627	
	Between Std. Dev.	4.909	3.568	5.164	
	Within Std. Dev.	2.109	1.304	2.268	
Scientific knowledge within ATC	Overall mean	6.327	6.787	6.211	The number of papers and scientific publications for an ATC-3 present in PubMed and other sources.
	Overall Std. Dev.	1.718	1.623	1.724	
	Between Std. Dev.	1.705	1.627	1.709	
	Within Std. Dev.	.242	.177	.256	
Number of firms within ATC	Overall mean	21.054	32.802	18.082	Number of firms trading within an ATC-3
	Overall Std. Dev.	20.954	23.442	19.175	
	Between Std. Dev.	20.604	23.069	18.876	
	Within Std. Dev.	4.050	5.367	3.645	

Tab 2 displays the overall, between, and within standard deviation for the main controls,

including sales. The statistics are provided for the total sample, the subset of ATC-3 having undergone at least a recall, and the sub-sample of ATC-3 without recalls. The panel of sales in Tab. 2 displays how, typically, the recalls are found in larger markets than the average. For this reason, recalls have been normalised by the number of products in the ATC market to avoid any possible problems of reverse causality with the market size. The normalised recalls have been denoted as $\tilde{recalls}$ in the following paragraphs.
Moreover, as expected, more competitive markets are more prone to recalls, as displayed by the Herfindahl–Hirschman Index (hhi). There is evidence of differences in terms of competition between ATC-3 groups. Nutarelli (2021) provides further insights about which type of firms and products generally undergo a recall. Specifically, Nutarelli (2021) evidences a general tendency for recalls to be located in large established firms and to regard relatively older products than the average age.
In contrast, with respect to firm and product levels, recalls are located in more dynamic ATCs, where recalled drugs were pioneering in the past. The fixed effects technique accounts for the time-invariant characteristics of ATCs.
The recalls intervene in firms with a high share of generics (see Nutarelli (2021)). This finding might result from a less stringent policy for generic drug approvals than branded drug approvals. A growing concern for generic safety is a well-known problem in the literature (see, e.g., Gallelli et al. (2013)).
In addition, in ATC markets, the outflow rate presents a within variance higher than the between variance. The latter means no difference between ATC-3 groups concerning the outflow rate. As opposed to the firm level, this inversion is expected. While strategic policies of product placement might occur in firms, this is not the case for ATC aggregation, where market laws apply. Thus, on average, even two utterly different ATC markets would display similar outflow rates following only a demand-supply logic.
Two other variables seem to be related to recalls at the ATC-3 level: the first is scientific knowledge within an ATC and the second is the number of firms trading within an ATC. Specifically, the recalls happen in ATC markets which, on average, trade more firms and where scientific knowledge is more advanced than in other markets.
In summary, the recalls regard relatively old drugs that were produced by large established firms. The major recalls occur in relatively dynamic markets whereby, on average, many younger firms operate, who trade relatively young products. A possible reason why markets with the described characteristics more easily undergo recalls is that they are precisely the markets that are monitored by the legislator with special attention.

5.1.2 Analysis of the determinants of drug recalls

This section will report the first-stage results. A Fixed-Effects estimation method is employed. Tab.7 shows the estimates of the first stage at the ATC-3 level. As detailed, a further level of ATC-Firm has been added to test for compensations within ATCs inside firms. For consistency with the best model, the sample was truncated in 2013, also for the first stage. Outcomes with a non-truncated sample display very similar results (see Nutarelli (2021)). The F-statistic amounts to 14.32. The standard errors included in the tables of the present section and the following sections are all robust and are clustered at the ATC-3 level of aggregation.

Table 3: First stage results at different levels.
ATC-3 aggregation represents the main specification.

	(ATC-3-Firm Aggregation) Log sales	**(ATC-3 Aggregation)** Log sales
$\tilde{recalls}$	−0.0053	−0.0283***
	(0.0033)	(0.0056)
$\tilde{recalls}_{t-1}$	−0.0226**	−0.0267***
	(0.0083)	(0.0070)
$\frac{K_{t+1}}{P_{-1}}$		0.1932**
		(0.0628)
average age firm		0.1576
		(0.0921)
average age firm^2		−0.0020
		(0.0013)
$\tilde{hhi}^{(a)}$		1.2405***
		(0.2590)
share generics in ATC		−0.1895
		(0.3373)
papers		−0.0260
		(0.0507)
# firms		0.0077
		(0.0071)
Year Dummies	Yes	Yes
Obs.	48915	1664
Groups	8634	208

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
(a) Relative Herfindahl–Hirschman index

Huber-White robust and clustered at ATC-3 level. (ATC-3-Firm Aggregation) fits an F.E. model at ATC-Firm level (i.e., ATC-3 lines of productions within firms). This level is introduced to check the possibility of compensations between sales of products belonging to the same ATC-3 within a firm (ATC-3 Aggregation) fits an F.E. model at the ATC-3 level. $\tilde{recalls}$ represents normalised recalls. At the ATC-3 level, recalls are normalised for the number of products within an ATC.

We found a significant and negative impact of recalls on the logarithm of sales at the market level. In Nutarelli (2021), it is shown that sales of firms undergoing a recall are unaffected. At the same time, the production lines of medicines belonging to the same ATC-3 encounter a drop in sales due to recalls (ATC-Firm Aggregation in Tab.7). This evidence excludes the possibility of compensations between sales of products belonging to the same ATC inside a firm. Therefore, the negative effect of recalls at the market level is enforced, whose lack is not filled by the same firms with other medicines of the same ATC-3.
The second column of Tab.7 represents the first stage of the principle analysis. As illustrated, the effect of recalls at the ATC-3 level is powerful and significant for current recalls and delayed recalls. After performing a sufficient amount of bootstrap repetitions, we found that the t-statistic is invariant to whether we use recalls or lag recalls to obtain it [6]. This finding corresponds to a Sargan-Hansen test for over-identification in our contest, which implies the absence of over-identifying restrictions (Lin and Wooldridge (2019)).
We believe that the key reason for the strength of this result relies on the level of aggregation. While firms with high-quality managements and inclined to risk can promptly make up for severe recalls, the latter take ATC-3 markets unaware. Consequently, competitors could not anticipate severe recalls against firms producing in the same ATC as theirs, which can only be detected at the market level.
The absence of compensations at the market level has been further tested. In particular, we analyzed the effect of recalls on aggregated sales once the firms' sales having undergone a recall are removed from the sample. The drop in sales seems to disappear once firms having

[6]The t-stat is obtained after 30000 repetitions and amounts to 2.438.

undergone a recall are excluded (see Fig. Appx.2 in Appendix).
The fall of sales observed at the ATC-3 level becomes evident not only from the estimates in Tab.7 but also from the study of abnormal values in Section 5.1.3
Finally, it might be argued that because recalled products are the most innovative, no direct substitute is present in the same market. However, we have the generic name of products (both recalled and not recalled) and the active principle of medicines at our disposal and it has been possible to detect an average of 10 products within the market exploiting the same active principle as the recalled products. Hence, we have also validated the hypothesis that the lacks left by recalls are filled with products already present on the market and that recalled products are not necessarily the most innovative that have no substitutes.

5.1.3 Analysis of abnormal values

This section reports our estimates of the influence of drug recall on sales. The effect of recalls is defined by taking a reference value of the given economic indicator as it would be observed under "normal" dynamics of economic conditions; this is called the "potential" value. We hence define the Abnormal Value (AV) of indicator y associated with unit i in time t as the observed and potential value difference. Thirumalai and Sinha (2011):

$$AV_{it} = y_{it} - E\left(y_{it}\right), \tag{4}$$

The potential value $E(y_{it})$ is estimated by running a Fixed-Effects regression on the following model:

$$y_{it} = \alpha + \beta y_{st} + \gamma X_{it} + \mu_i + \lambda_t + u_{it}, \tag{5}$$

where y_{st} is the aggregated value of y in year t at the sector level. The usual control variables (X) and year dummies are included as regressors. After obtaining estimates of AV_{it} for all i and t, referred to as $\widehat{AV}_{it}$, the time dimension is re-scaled. Specifically, the time dimension is centered on the year when the recall is issued for all units experiencing a recall in the time frame considered. Only these units are kept in the sample. The market-level Abnormal Value $\overline{AV}_t$ associated to recalls is then computed as the simple average of $\widehat{AV}_{it}$ for any $t \in \{-(T-1), ..., (T-1)\}$, as follows:

$$\overline{AV_t} = \sum_{i=1}^{N_t} \widehat{AV}_{it}, \tag{6}$$

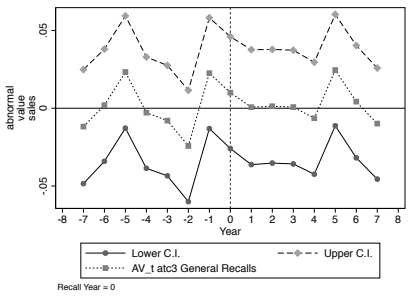
where N_t is the number of units with available data in t among those experiencing one recall. Confidence intervals for $\overline{AV}_t$ are constructed by calculating the variance of $\widehat{AV}_{it}$ as follows:

$$Var\left(\overline{AV}_t\right) = \frac{\sum_{i=1}^{N_t} Var\left(\widehat{AV}_{it}\right)}{N_t^2}, \tag{7}$$

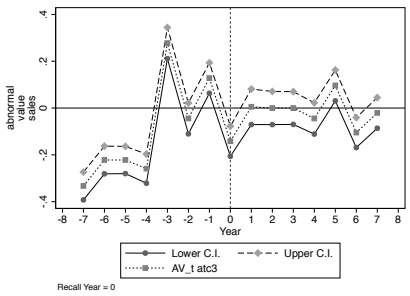
where $Var(\widehat{AV}_{it})$ is the variance of the forecast error derived from estimation of Equation 5. The focus of our analysis is on the growth rate of sales volumes. The exercise is replicated for three classifications of recalls (i.e., standard recall definition, major recalls, and type of recall)

and three levels of analysis (i.e., product, firm, and sector level). The main text only reports the analysis at the ATC-3 level because it is the level at which the first and second stages are conducted. Abnormal values at the firm and product level can be found in the Appendix. Note that in the model for the sector level, y_{st} is replaced with y_{mt} in Equation 5; that is, the value at the whole market level.

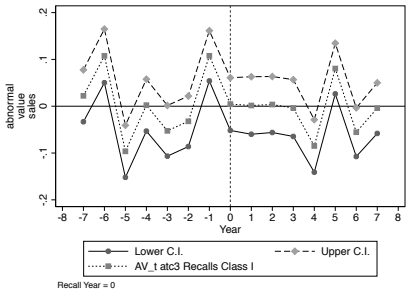
Fig. 5 reports estimates of the effects of recalls on the AV of sales growth.



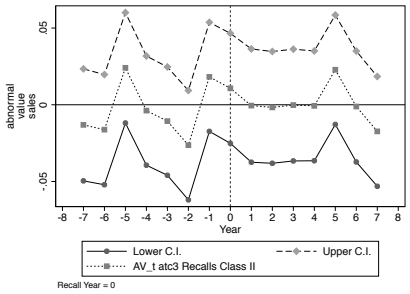
ATC-3: general recalls sales



ATC-3: Maj. recalls sales (preferred specification)



ATC-3: Class I recalls sales



ATC-3: Class II recalls sales

Figure 6: Abnormal values at ATC-3 level of aggregation. Years are normalised so that year 0 represents the year of recall. The four scenarios include the path of sales before and after the recall year, using four different definitions of recalls: major recalls, Class I recalls, general recalls and Class II recalls. As is evident from the diagrams, sales drop at recall year for every type of recall. Major recalls present a more pronounced drop. Moreover, the lowest error bound is reached when using major recalls.

Fig.6 shows the abnormal values for the ATC-3 level. Confidence intervals are constructed at the 95% level. As is evident from Fig.6, after the initial drop at the year of recall, sales quickly recover one or two years after year 0 (see major recalls). The latter observation classifies the instrument employed in our work as a short-run effect. This distinguishes the effect of our instrument from the long-run effect that demographic shocks that were produced in the work of Acemoglu and Linn (2004).
The analysis of abnormal values confirms what was found in previous paragraphs, especially a considerable impact of recalls on sales in the year of the recall. The error bound is lower for the ATC-3 level with major recalls, thus enforcing the expectation of a drop in sales at the time of recall.

5.2 Relationship between innovation and market size

In this section, we will report the results concerning the relationship between market size, M_{it} and innovation, N_{it}. Because the data at our disposal are already converted into dollars of 2015 using the Consumer Price Index (CPI), market size is measured directly as the sum of sales over ATC market i at time t. Innovation is measured with the number of activated trials in ATC i at time t. The time window ranges from 2004 to 2013. The last two years of the samples (i.e., 2014 and 2015) have been cut away because very few trials have been conducted in this period. Consequently, including 2014 and 2015 may have led to biases in the procedure, which exploits Poisson estimates. Indeed, this method does not tolerate a value of 0 for the dependent in most observations.

The panel is strongly balanced, as required by the procedure. Each year has data for 208 therapeutic classes.

The best model is estimated by Eq.(1).

We have introduced several regressors. These comprise supply-side determinants, technological opportunities, and age determinants. We draw some controls directly from the literature, comprising knowledge stock (see, e.g., Cerda (2007), Acemoglu and Linn (2004) among others) as measured by the number of papers referred to the ATC category i. The PubMed database has been consulted. Specifically, we collected the number of scientific works for a given ATC-3 in a given year through Mesh Terms. According to NIH, MeSH terms are official words or phrases that are selected to represent particular biomedical concepts. When labeling an article, the indexers only selected terms from the official MeSH list, and never drew on other spellings or variations. To decide whether or not a paper referred to a specific ATC class, we first associated the a Mesh Term to an ATC category i, primarily exploiting the official synthetic description of ATC. If the latter did not produce a result or did not match evidence from the literature, then a double-check was made using level 3 indications as Mesh terms.[7]. The NCBI Mesh database allowed us to customise the searches. Because the number of papers showed an upward trend, the variable has been detrended by first differentiating its logarithm.

Another critical control drawn from the literature is the share of generics. As noted in Dubois et al. (2015), the ease of entry and substantial financial incentives to use generics will reduce the expected profitability of the innovation. Hence, detecting the degree of penetration of generics within the markets is vital because it might discourage firms from undertaking innovation.

As emphasised in both Acemoglu and Linn (2004) and Dubois et al. (2015), a further source of declining margins of innovation is represented by the increasing number of young entrants within an ATC market. Pharmaceutical competition, in general, might undermine innovation productivity. It is, thus, imperative to measure and control for competition.

Apart from Acemoglu and Linn (2004), the empirical literature does not explicitly model

[7]For instance, category C6B is described as "PULMONARY ARTERIAL HYPERTENSION (PAH) PRODUCTS." Due to the length of the name and possible different abbreviations employed in the Mesh Terms list, Mesh Terms have been searched by looking at different specifications of the description, such as "PAH PRODUCTS," "PULMONARY ARTERIAL HYPERTENSION PRODUCTS." If the latter did not produce a result or the results were not in line with the findings in the literature, then the Mesh indication at level 3, "PAH" was selected

competition (see Dubois et al. (2015)). In the present work, we constructed two measures to control for pharmaceutical competition. The first measure is the Herfindahl index (*hhi* hereafter). The latter is employed as an indicator of competition among firms within an ATC-3 market. Compared to other measures, such as the concentration ratio, the major benefit of the Herfindahl index is that it gives larger firms more weight. The index can range from 0 to 1.0, moving from many tiny firms to a single monopolistic producer. The second measure to control competition is the average age of firms within a market. It controls other aspects of competition compared to *hhi*. While the *hhi* measures the "degree of monopoly" within an ATC, it cannot fully describe the types of firms populating the market. For this reason, we introduced a further control (i.e., the average age of firms), which mainly catches the presence of small biotechnology firms in the market. On the one hand, these firms are known to compete for innovation; and on the other hand, they are known to have fewer financial resources in contrast with established companies (see, e.g., Hall and Rosenberg (2010) among others). Because margins decline with the number of young entrants, we expect a negative sign for the firms' average age.

Tab.4 presents the main results of the analysis. This is technically the second stage of the procedure described in the methodological section. In particular, calling z_{it2} the excluded instruments ($\text{recalls}_{it}, \text{recalls}_{it-1}$) [8], the first stage estimation computes the residuals,$\ddot{u}_{it2}$, of a linear fixed-effect model whose dependent is market size. The second stage incorporates the residuals and estimates a fixed effect Poisson model. Please refer to steps 1. and 2. in the methodological section.

In contrast to the literature, in the present work it is not necessary to construct M_{it} based on demographic shifts because the innovative instrument, recalls, already purges market size from endogeneity. In the following, M_{it} is simply the logarithm of collapsed sales at ATC-3 level (i.e., the product of the number of purchased drugs expressed in standard units to ensure comparability with their price).

Notice that a critical assumption of the model is that excluded exogenous, R_{it} appearing within z_{it}, do not explicitly appear in the equation of trials. For the more refined aggregation level at our disposal, ATC-3, it is plausible to assume that the average elasticity is the same across categories.

[8] We included two instruments because, following Hansen et al. (2008), instrumenting with more valid instruments leads to more accurate estimates.

Table 4: Impact of market size on innovation. Column (P) employs a simple Poisson model not considering fixed effects. Column (NB) adopts a Negative Binomial, due to the presence of overdispersion. Column (CF-IV) is the main specification (control function fixed effect IV Poisson). Column (A-B) and Column (A-B linear) add the lag of the dependent following Acemoglu and Linn (2004). Column (NR) is the main specification eliminating all of the regressors

	(P) Trials	**(NB)** Trials	**(CF-IV)** Trials	**(A-B)** Trials	**(A-B linear)** *log Trials*	**(NR)** Trials
$trials_{t-1}$				−0.00741	0.0732*	
				(0.0005)	(0.0335)	
Log sales	0.1378***	0.122***	0.6362**	0.802**	0.1176***	0.8229**
	(0.0060)	(0.0224)	(0.2149)	(0.266)	(0.0153)	(0.3174)
residuals			−0.8018***	−0.862**		−0.9711**
			(0.2157)	(0.269)		(0.3177)
$\frac{K_{t+1}}{P_{-1}}$	−0.5378***	−0.423*	−0.0926	−0.484***	−0.0504	
	(0.0847)	(0.182)	(0.0909)	(0.147)	(0.0914)	
average age firm	0.2890***	0.178***	−0.1332***	−0.106*	0.0634**	
	(0.0139)	(0.0333)	(0.0377)	(0.0398)	(0.0214)	
average age firm^2	−0.0038***	−0.00234***	0.0021***	0.00178***	−0.0008**	
	(0.0002)	(0.0004)	(0.0005)	(0.0005)	(0.0003)	
$hhi^{(a)}$	0.2245***	0.548**	−0.3199	−0.145	0.1106	
	(0.0446)	(0.200)	(0.2903)	(0.360)	(0.1153)	
share generics in ATC	−0.5571***	−0.526**	−0.3168**	−0.898***	−0.2036	
	(0.0404)	(0.180)	(0.1124)	(0.143)	(0.1068)	
average age product	−0.0658***	−0.0512*	−0.0592**	−0.0928**	−0.0564***	
	(0.0061)	(0.0240)	(0.0190)	(0.0323)	(0.0143)	
average age product^2	0.0010***	0.0005	0.0011	0.0100	0.0010*	
	(0.0002)	(0.0007)	(0.0010)	(1.64)	(0.0004)	
papers	0.5608***	0.175	0.1558*	0.101	−0.0443	
	(0.0728)	(0.237)	(0.0750)	(0.0013)	(0.1477)	
papers^2	−0.6672***	−0.316*	−0.0067	−0.0929	−0.0083	
	(0.1056)	(0.131)	(0.0838)	(0.083)	(0.0883)	
# firms	0.0090***	0.0111***	0.0008	−0.0032	0.0048**	
	(0.0007)	(0.0028)	(0.0035)	(0.0792)	(0.0016)	
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	1664	1664	1664	1664	1664	1872
Groups	208	208	208	208	208	208
Pseudo R^2	0.1346	0.027	.	.	.	.
Overdis.	Yes	.	.	.	.	.
Zero-inflated (Vuong)	.	No	.	.	.	.

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$^{(a)}$ Relative Herfindahl–Hirschman index

Huber-White robust and clustered at ATC-3 level. The dependent variable is the count of active trials in ATC i at time t in (P), (NB), (CF-IV) and (NR). The time interval is 10 years.

Column (P) presents a simple Poisson model with exogenous market size, which explores whether market size's positive effect is robust in the absence of fixed effects and endogeneity controls. Due to the presence of overdispersion, Column (NB) performing a naive [9] Negative Binomial model is added. Column (CF-IV) is our main specification (i.e., a fixed effect Poisson controlling for market size's endogeneity).

The coefficients of interest in Tab.4 are Log sales and residuals: the former represents the market size, and the latter measuring endogeneity of market size. Specifically, a significant coefficient of residuals denotes a correlation between the error term (see the specification in Tab.4; i.e., second stage regression) and functions of the error of the model of the market size (first stage). In other words, residuals control for co-movements of sales and unobservables related to the number of trials. Market size is hence "purged" from the alleged endogenous part. Endogeneity is tested with a Wald test on the residuals' coefficient ρ. If ρ is significantly different from zero, then endogeneity is present. This latter instance occurs in our model,

[9] With exogenous market size and not controlling for fixed effects.

as expected (see Column (CF-IV)). In particular, fully robust standard errors detect a strong idiosyncratic endogeneity. The exploitation of Fixed Effects methodologies allows the unobserved heterogeneity to be correlated with all explanatory variables and the excluded exogenous recalls. The evidence is that, even after allowing the market size to be correlated with the ATC heterogeneity, market size is not exogenous to idiosyncratic shocks.
The coefficient of market size is positive and significant, which is in line with past works. According to our estimate, a 10% increase in market size leads to an increase of almost 6.3 % of active trials. This magnitude conforms with the literature. Indeed, previous research generally finds elasticities to be approximately 0.5, which is consistent with our estimates.
The recent literature has speculated on the possibility that, though clinical trials might respond elastically to market size, the proportion resulting in effective innovation might decline (see e.g.Dubois et al. (2015) among others). Hence, authors might have overestimated the effect of market size on clinical trials because the latter should only be computed on the trials that effectively brought innovation. In this paper, we exploited active trials as a dependent, which partially solves the issue. We believe that active trials constitute a subset of promising trials in terms of innovative contribution. The estimated higher effect than the literature that adopts NMEs or NCEs as a dependent can be well explained by the substantial costs for developing new pharmaceutical entities. Drug development is, in fact, quite expensive—the cost range between $800 million to $2.5 billion (see, e.g. the FDA programme MedWatch). Undertaking clinical trials is, instead, sensibly cheaper, amounting to an average of $20 million to $40 million (see Martin et al. (2017) as well as John Hopkins Bloomberg Health School, 2018). Thus, it is reasonable to suppose that, ceteris paribus, a 10% increase in market size stimulates more trials than NMEs or NCEs (on average). However, exceptions are still present (see Acemoglu and Linn (2004), Duggan and Scott Morton (2010), who estimated an higher elasticity than the one of the present work).
The coefficient of the average age of firms and its square is in line with past observations (see, e.g., Huergo and Jaumandreu (2004) and Balasubramanian and Lee (2008) for specific studies on the topic). This effect evidences how the oldest firms tend to introduce less innovation than entrants in their early years. However, firms above intermediate ages *appear almost as active in process innovations as entering firms, and even more in product innovations* (Huergo and Jaumandreu (2004)).
Moreover, innovation decreases with the share of generics within a market. Thus, the effect theorised in Dubois et al. (2015) of decreasing margins of innovation proportionally to the entrance of generics is revealed to be correct (see also Lanjouw (2005)).
In line with Acemoglu and Linn (2004) and Rake (2017), technological advancements as measured by detrended papers are positively related to innovation. It is therefore reasonable to suppose that more trials emerge in markets where scientific research is prolific.

The discrepancies in the magnitude of the coefficients between the main specification (Column (CF-IV)) and Columns (P) and (NB) of Tab.4 can be explained in several ways. In Columns (P) and (NB) of Tab.4, the correlation over time of units is not controlled. Consequently, it is assumed that units are independent over the cross-sectional dimension and over the time dimension, which is quite a strong constriction in a longitudinal setting. This assumption means that the same individual (market) observed at two different times, t_0 and t_1, is considered to independent from themselves. In other words, individual (market) i

at time t_0 is another individual (market) than individual (market) i at time t_1. The main implication of this presumption is that unobserved the time-independent heterogeneities of individuals do not affect other individuals. However, we know that the same individual when observed two different times is considered to be "two distinct individuals." Thus, in the model of Columns (P) and (NB), it is ultimately assumed that unobserved shocks of an individual (market) i at time t do not influence individual (market) i at time $t+k$. In other words, we are mixing between and within individual effects. Between effects are obtained once the time component is averaged out from the variables. Between-effect settings exploit differences between units, which in our case are independent by definition (we take ATC-3 markets, see previous Sections), not taking into account time variations. Therefore, the market size variance (time-demeaned) will be higher in a between-effect setting because it considers the average market size difference between independent ATC-3 markets. Furthermore, given the opposite time trends of trials and market size (see Fig. 1) in a between-effect setting, the between effects of market size on innovation will be deflated. Indeed, the innovation trend decreases from a specific time onwards, while the market size trend increases. However, given that time variations are not controlled in a between-effect setting, an inverse proportionality of market size and innovation emerges. Mixing between and within individual effects will, hence, result in overall lower coefficients of market size of Columns (P) and (NB) when compared to that of Column (CF-IV).
Ultimately, the downwardly biased coefficients of Columns (P) and (NB) suggest that the unobserved heterogeneity is negatively correlated to trials.
To provide an example, consider the possibility that an ATC experienced a sizeable positive shock (more trials) in 2010. For some reason, the mentioned shock is not modeled nor measured. All else being equal, the apparent fixed effect for that ATC in the period 2004 to 2013 will appear to be higher. However, from the literature, we know that the margins on each product will be lower when more products are available for treating a particular clinical condition (see Bresnahan and Reiss (1991) among others). Therefore, the unobserved positive shock for ATC i^{th} would lower the margins of all competitor products in the same market, which will push down the sales for the same market. This negative correlation between the market size regressor and the error term deflates the estimate for market size. And vice versa, in Column (CF-IV), time dependency is controlled and deflation is eliminated. Therefore, the coefficient of market size result is higher than in Columns (P) and (NB).
Columns (P) and (NB) do not control the reverse causality of market size on innovation. Not considering the reverse causality of market size contributes to upward biasing of the market size's coefficient (see, e.g., Acemoglu and Linn (2004)). Therefore, there are two contrasting effects in Columns (P) and (NB): the upward effect due to the reverse causality endogeneity and the downward bias given by the unobserved heterogeneity endogeneity. These two effects do not compensate, and the negative heterogeneity bias prevails over the reverse causality endogeneity bias.

Robustness checks Columns (A-B)-(NR) of Tab.4 investigate the robustness of the effect of market size on innovation. Three additional models are added to the preferred specification. In particular, Column (A-B) reproduces the exercise of Acemoglu and Linn (2004) to control

for possibly varying over time technological flows (see below) by adding lagged trials among the regressors. Because the estimating equation in Column (A-B) is nonlinear, we perform this instrumentation strategy by adding the residuals of the first stage. Column (A-B linear) is the same as Column (A-B), where the dependent is log linearised and residuals are ignored. Column (A-B linear) ignores both the presence of non-linearities and endogeneity.
Adding lags of the dependent is a valuable exercise. Indeed, following Acemoglu and Linn (2004), the primary threat to the identification strategy of innovation is represented by changes in the flow rate of innovation for every dollar spent for research on a drug (note that permanent differences in innovation are already dropped through the ATC fixed effects). Differences in the flow rate of innovation suggest that technological progress is more difficult in some lines than others. The parameter denoting innovation flow is part of the theoretical specification of innovation drawn from Acemoglu and Linn (2004). Following Acemoglu and Linn (2004), if the flow rate of innovation varies over time, then it is also likely to be serially correlated. Adding the lag of log innovation to the preferred specification is a simple way to check for the importance of these concerns. The lagged trials are instrumented with their lags through a system GMM one-step procedure. The p-value of the Hansen test of overidentification of model in Column (A-B linear) is 0.175, which mainly falls between the tolerance levels of 0.1 and 0.25 indicated in Roodman (2009). The Arellano-Bond test is investigated in Tab.5.

	z-score	p-value
Arellano-Bond test for AR(1) in first differences:	z = -10.47	Pr > z = 0.000
Arellano-Bond test for AR(2) in first differences:	z = 0.88	Pr > z = 0.377
Arellano-Bond test for AR(3) in first differences:	z = -1.46	Pr > z = 0.145
Arellano-Bond test for AR(4) in first differences:	z = 0.33	Pr > z = 0.740

Table 5: Arellano-Bond test for autocorrelation of first differenced residuals of GMM

When the idiosyncratic errors are independently and identically distributed (i.i.d.), the first-differenced errors are first-order serially correlated. Thus, as expected, Tab.5 presents strong evidence against the null hypothesis of zero autocorrelation in the first-differenced errors at order 1. As suggested in Roodman (2009), *"in the context of an Arellano-Bond GMM regression, which is run on first differences, AR(1) is to be expected, and therefore the Arellano-Bond AR(1) test result is usually ignored in that context"*. Moreover, this output presents no significant evidence of serial correlation in the first-differenced errors at orders 2, 3, and 4.
In Column (A-B linear), market size is considered to be exogenous, although fixed effects are controlled. The model in Column (A-B linear) is linear. To ensure comparability among models, trials have been transformed to a logarithmic scale. Column (A-B linear) is, in other words, an essential control because, although it controls for fixed effects, it ignores the presence of potential non-linearity (misspecification) and endogeneity, thus proposing the

hypothesis of serial correlation.
Finally, Column (NR) presents the model without further control as estimated by the preferred specification's control function approach. The idea beyond Column (NR) is to check whether not controlling for regressors compromises the main specification estimates.

The outcomes of Columns (A-B)-(NR) of Tab.4 confirm the estimates of the main specification concerning the positive effect of market size on innovation.
Columns (A-B)-(NR) in Tab.4 all display a positive effect of market size on innovation.
Specifically, Column (A-B) confirms the results of Acemoglu and Linn (2004) in finding no evidence of serial autocorrelation. In particular, the coefficient of lag trials is negative and non-significant, as in Acemoglu and Linn (2004). Possible explanations are already in Acemoglu and Linn (2004) and are, therefore, not discussed in the present work. In Column (A-B linear) of Tab.4, the positive coefficient of lagged trials is significant at the 5% tolerance level. This evidence is almost in line with Acemoglu and Linn (2004) when no instrumentation is performed. Under this scenario, the lagged dependant's coefficient was found to be positive and not significant, which was also found in Acemoglu and Linn (2004). Market size is again strongly and positively related to innovation, with a coefficient having the lowest magnitude of the specifications analyzed until now. Indeed, some of the variability might be caught by the lagged dependent. Moreover, possible misspecification bias due to the not correction of nonlinearity might intervene.
Notice that the effect of the market size in Column (A-B linear) of Tab.4 display similarities to Columns (P) and (NB) of the same table, which does not control for endogeneity. Furthermore, the effect of market size is larger in the models correcting for endogeneity. Therefore, in general, the lack of control for temporal dependence may matter very little for estimation because it is also consistent with the fact the autocorrelation coefficient is very weak. It is, hence reasonable to suppose that the lower magnitude of the coefficient of size in Columns (P), (NB) and (A-B linear) of Tab.4 is primarily a consequence of considering market size as exogenous. It is possible to provide further checks by controlling for possible overidentification of the instrumented lagged dependent variable. To do so, Tab.6 Column (1) reports the two-step robust system GMM estimates of Column (A-B linear) of Tab.4, which, instead, performed a one-step system GMM.

Table 6: Coefficients of market size and lag dependent when a two-step GMM is employed. Column (2) includes suspended and withdrawn trials in the dependent

	(1) *log Trials*	**(2)** *log Trials*
$trials_{t-1}$	0.0592	0.380*
	(0.0426)	(0.189)
Log sales	0.1208***	−0.533**
	(0.159)	(0.179)
Year Dummies	Yes	Yes
Obs.	1664	1664
Groups	208	208

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered at ATC-3 level.(1) is the two-step GMM version of Column (A-B linear) of Tab.4. Only the critical coefficients are included. (2) is equal to (1) where also suspended and withdrawn trials are included in the dependent. Only the critical coefficients are included. Both equations are linearised to enable a simple comparison with Column (A-B linear) Tab.4.

The coefficient of the market size in Tab.6 is higher compared to that in Column (A-B linear) of Tab. 4. Furthermore, the lagged dependent variable is not significant, which is in line with Acemoglu and Linn (2004). The same applies if the count of trials is employed as a dependent (see Nutarelli (2021)).
The sign of the estimates of market size does not change compared to the preferred model.
A final robustness check has been made by including all of the trials (i.e., active trials, and suspended and withdrawn trials). The number of classes employed is the same, although the number of trials increased by 0.57% in total. This is performed in Tab.6 column (2). As displayed, our estimation does not confirm the hypothesis in Dubois et al. (2015), showing a lower coefficient instead in terms of both magnitude and significance level. The hypothesis is that including non-active trials, which are less responsive to market size, biases the estimate toward randomness. For instance, firms with all suspended trials are unaffected by governmental price regulations that reduce the price of treatment. Simultaneously, increases in market size could be less effective on these companies, which have already reduced costs due to inactive trials. Therefore, the presence of endogeneity is confirmed.

Further robustness checks have been performed by changing the market size's proxy to align to Acemoglu and Linn (2004), moving to another database to collect sales data (Evaluate sales are employed), and employing all of the recalls at our disposal to instrument market size. In particular, Tab.7 shows the outcomes of the analysis adopting Class II and Class I recalls as an instrument for market size. Tab.8 measures market size through the number of patients within an ATC-3, in accordance with Acemoglu and Linn (2004).
Because the number of patients is highly correlated with sales and it is employed as a natural alternative to sales, we have adopted recalls as an instrument for the number of patients.
The F-test amounts to 12 for the analysis with Evaluate and to 4 for the analysis with the number of patients.

Table 7: Column (First-stage all recalls) and Column (Second-stage all recalls) represent first and second stage results using all of the recalls at our disposal. Data are aggregated at the ATC-3 level. The impact of market size on innovation is tested using Evaluate database in Columns (First-stage Evaluate) and (Second-stage Evaluate) Furthermore, Column (First-stage Minor Recalls) shows the poor strength of minor recalls (i.e., recalls whose motivation pertains labeling and packaging).

	(First-stage Minor Recalls) Log sales	**(First-stage all recalls)** Log sales	**(Second-stage all recalls)** Trials	**(First-stage Evaluate)** Log sales	**(Second-stage Evaluate)** Trials
$\tilde{recalls}$	0.00138	0.0009		-0.260^{***}	
	(0.00302)	(0.00345)		(0.0059)	
$\tilde{recalls}_{t-1}$	-0.0017	-0.0223^{***}		-0.0180	
	(0.00308)	(0.00678)		(0.0105)	
Log sales			0.636^{*}		0.710^{**}
			(0.255)		(0.275)
Residuals			-0.802^{***}		-0.802^{***}
			(0.216)		(0.275)
$\frac{K_{t+1}}{P_{-1}}$	0.187^{***}	0.191^{**}	-0.0926	-0.173	-0.147
	(0.0617)	(0.0621)	(0.0909)	(0.154)	(0.276)
average age firm	0.151	0.153	-0.133^{***}	0.0470	0.224^{***}
	(0.0916)	(0.0914)	(0.0377)	(0.0678)	(0.0045)
average age firm^2	-0.0019	-0.00195	0.00214^{***}	-0.0004	-0.0814^{***}
	(0.00127)	(0.00127)	(0.000488)	(0.0008)	(0.0316)
hhi	1.243^{***}	1.239^{***}	-0.320	1.721^{***}	-0.60
	(0.259)	(0.259)	(0.290)	(0.419)	(0.495)
share generics in ATC	-0.155	-0.181	-0.317^{**}	-0.372	0.00640
.	(0.341)	(0.339)	(0.112)	(0.328)	(0.157)
average age prod.	0.0533	0.0539	-0.0592^{**}	-0.0330	-0.0732^{**}
	(0.0587)	(0.0586)	(0.0190)	(0.0504)	(0.0225)
average age prod.^{2}	-0.003	-0.0037	0.00109	0.00138	0.0009
	(0.00216)	(0.00216)	(0.000966)	(0.00206)	(0.0009)
papers	-0.0264	-0.0267	0.156^{*}	-0.139	0.265^{**}
	(0.0510)	(0.0514)	((0.0750)	(0.0803)	(0.0917)
# firms	0.0077	0.00734	0.000825	-0.0106	0.0224^{***}
	(0.00699)	(0.00701)	(0.00346)	(0.0095)	(0.0045)
Year Dummies	Yes	Yes	Yes	Yes	Yes
Obs.	1664	1664	1664	1136	1056
Groups	208	208	208	142	132

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered at ATC-3 level.

In the first column, Tab. 7 replicates the main analysis while adopting minor recalls rather than major recalls. In contrast to major recalls, minor recalls are represented by the available recalls that are due to labeling or packaging issues. This exercise provides an intuition about the differences in the motivations of the recalls, showing how minor recalls are, in fact, a weak instrument of market size (neither minor recalls in t nor lagged recalls are significant in the first stage). This weakness is also due to the fact that minor recalls are mainly voluntary, and hence can be easily anticipated by firms. This enforces the usage of major recalls as an instrument for market size (as explained in the main analysis).

Moreover, Tab. 7 reports the first and second stages of employing Class I and Class II recalls as an instrument for market size in the second and third columns. The remaining columns are devoted to the results obtained using the Evaluate database to collect market sales. The results of the primary analysis are confirmed in both exercises.

Employing all recalls decreases both the magnitude and the significance of the coefficients of sales. Moreover, only the lag of recalls is a good instrument at the market level. These two effects are expected because minor recalls may attenuate the drop in sales consequent to a recall. Indeed, within Class II, temporary recalls (e.g., recalls due to a labeling error) are also comprehended, which may not be unexpected to the firm (most of them are voluntary). For this reason, they might not be taken into account by the company's management. Therefore, losses in terms of sales are well compensated. Furthermore, minor recalls are not publicised and cannot damage the image of the company or the market in which they happen.

Hence, adding minor recalls overtakes the strong and negative impact of current recalls, and

consequently affects the estimates of the market size in the second stage. However, because Class II recalls often regard minor but persistent issues[10], a cumulative effect intervenes, and lagged recalls remain an excellent instrument.
The outcomes of the principal analysis remain robust when data on sales are collected from the Evaluate database.
Tab.8 reports the second stage results of the analysis with the number of patients as a measure of market size. The first stage results are investigated in the Appendix.

Table 8: Impact of market size on innovation using number of patients as a proxy of market size (only second-stage)

	(1) Trials
Log patients	3.274***
	(0.648)
residuals	−3.291***
	(0.647)
$\frac{K_{t+1}}{P_{-1}}$	1.476***
	(0.377)
average age firm	0.589***
	(0.154)
average age firm^2	−0.00478**
	(0.0016)
hhi	0.145
	(0.260)
share generics in ATC	−0.648**
	(0.244)
average age product	−0.278***
	(0.0514)
average age product^2	0.0011***
	(0.0022)
papers	0.546***
	(0.147)
papers^2	0.193
	(0.114)
# firms	0.0895***
	(0.0144)
Year Dummies	Yes
Obs.	1056
Groups	132

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered at ATC-3 level. (1) employs MEPS database and matches the ATC-3 present in our database. Market size is measured through the number of patients within ATC-3.

Adopting the number of patients as a proxy for market size confirms the first and second stage results compared to the principal specification. However, the outcomes are weaker in terms of significance than our main specification. Therefore, recalls do not seem a vital instrument for the number of patients. On the one hand, a more significant number of patients within an ATC-3 class might increase the probability of an adverse event than in a scarcely populated ATC-3 class, which would increase the probability of a major recall. On the other hand, there is no reason to believe that an adverse event would happen in a more populated class, which might refer to commonly employed (and therefore well-tested) medicines.

[10]Minor recalls often pertain to the manufacturing of the product accessories. Their cause ranges from label mix-up, particulate matter in specific lots, and packaging issues. Though minor recalls do not directly threaten the health of the patients, they are difficult to correct in the short-term by firms.

Moreover, a recall in a class causes a decrease in the number of patients adopting pharmaceuticals in the questioned ATC-3 class, thus compensating for the possible positive effect implied by a higher probability of adverse events. In the second stage, Tab.8 enforces the results found in Acemoglu and Linn (2004) where a coefficient of market size between 3 and 4 was found. Therefore, the significance of residuals confirms the presence of endogeneity.

6 Conclusions

Recent research has stressed the importance of market size in determining the innovation rate in the pharmaceutical industry. At the same time, after Cerda (2007)'s critique, instrumenting with demographic shifts remains a weak, although valid, strategy. Moreover, recent contributions have stressed the importance of modeling competition and technological opportunities adequately (see Rake (2017), Dubois et al. (2015)). For example, many scholars have pointed to the importance of advances in molecular biology and related fields for the industry's technological opportunities and innovative capabilities (Rake (2017)). Finally, the literature lacks aggregation analyses that would allow us to easily draw policy implications. For this reason, the present work employs ATC-3, which is the aggregation level used by the antitrust authorities.

The empirical estimates are conducted on a unique database that is integrated with additional sources. The variety of our sources enabled us to collect and adequately classify data on trials and drug recalls in ATC-3 categories. The methodology employed is innovative (Lin and Wooldridge (2019)) and, in contrast from previous techniques, it permits us to control for both idiosyncratic endogeneity and heterogeneity endogeneity. This technique has two stages. A simple Wald test on the residuals' coefficient in the second stage allows us to verify the presence of idiosyncratic endogeneity.

Recalls are an innovative instrument that has been employed for the first time in this paper. Recalls have been collected by consulting various sources, including the FDA Enforcement reports, openFDA, and a database deriving from FOIA agreements with FDA. Recalls are representative. Major recalls have been selected to meet the criteria of sharpness, indirect effect on the dependent variable (innovation), and exogeneity. The first stage displayed a significant negative impact of recalls on market size, thus validating the instrument. To the best of our knowledge, the effort constitutes an empirical novelty in the literature that mainly focuses on optimal management of recalls and provides a theoretical argument of the negative impact of recalls at the firm level. However, few papers have focused on the impact of drug recalls at the market level.

Data on clinical trials have been drawn from the Clinicaltrials.gov website, from the preclinical phase to Phase IV. They have been integrated with data on INDs from a privately owned database that is maintained at IMT School for Advanced Studies. To overcome issues deriving from the potential more robust response of market size to trials as a whole rather than on essential trials (i.e., bringing most probably to an innovation), only activated trials have been selected. This exercise also provides a valid answer to the argument that the recall of a product might imply the suspension of drug trials within its same family. Indeed, suspended and withdrawn trials have been excluded from the analysis. Nonetheless, as a

robustness check, estimates are computed, including the latter in the analysis. This effort confirms the presence of idiosyncratic endogeneity and the positive sign of the estimates. However, the magnitude and the significance level decrease.
Our preferred estimates align with literature, displaying an increase in the innovation of 6.3% after an increase in market size of 10%. Most recent studies by Dubois et al. (2015) display a lower coefficient of 0.23%. The authors specify how a comparison with other works exploiting different measures of innovation remains a difficult task. They further explain that their use of global data rather than US data for the estimations might have led to less responsiveness.
Our results are robust to several specifications. The coefficient of independent variables is in line with expectation, as well as the scarce effect of lagged trials, which was already tested in Acemoglu and Linn (2004). Further checks confirm a positive and significant effect of market size on innovation, even when fixed effects are not controlled, and the market size is considered exogenous. This latter verification partially validates (for what concerns the sign and the significance) the recent findings of Rake (2017), who did not find evidence of reverse causality. However, the coefficient's magnitude decreases sensibly when compared to the preferred specification, showing a significant bias.
Our estimates remain robust even when no control is inserted in the analysis.
This work provides exciting policy implications for innovation's stimuli and it sheds some light on the impact of recalls at the market level. In particular, governments should be aware of the fact that innovation is mainly an economic phenomenon. Companies innovate mainly to gain a financial return. Given the positive relationship between market size and innovation, authorities and policymakers should not penalise economic players too much. To guarantee their citizens' future welfare, they should promote research and invest in new technologies smartly by managing competition from generics.
Moreover, recalls do not just have an impact at the firm level but they also impact at the market level. Specifically, they provoke adverse market shocks, thus affecting economic stability and welfare. Authorities should therefore apply more stringent rules to avoid severe recalls. At the same time, they should consider that an intensification of Class II and Class III recalls due to the presence of more players might be physiological.
Future research might employ more up-to-date data, and also include recalls of compounders and repackaging firms.

Acknowledgements: We are grateful to Crisis Lab. for the PHID database. We thank Prof. Jeff Wooldridge for the kind support in applying and interpreting his innovative methodology. We would also like to thank the FDA for providing data on recalls and clearing any doubt on recall timing and procedures. We are grateful to Evaluate data for allowing very up-to-date and detailed checks on clinical trial data. Furthermore, we are thankful to Springer AdInsight for enabling a detailed classification at the ATC-3 level of rare clinical trials. Finally, we thank the Young Economists of Tuscan Institutions (YETI), the AXES unit of the IMT Lucca for Advanced Studies, Dr. Armando Rungi, Dr. Francesco Serti, Dr.Laura Magazzini of Sant'Anna School for Advanced Studies (Pisa) and the ASSA 2022 participants for the valuable comments on the work.

Appendices

A Literature review table

The following table summarises the findings from the literature on the relationship between market size and innovation in the pharmaceutical industry. In particular, the focus is on relevant works coming after Acemoglu and Linn (2004). The reason for this choice is that Acemoglu and Linn (2004) represents a milestone in investigating the relation between market size and innovation in pharmaceuticals. It overcomes issues emerging in previous studies (e.g., and above all, the one of endogeneity) and is taken as a reference point by authors willing to further delve into this stream of literature.
Furthermore, the literature review only reports the relationship between market size and innovation in the pharmaceutical industry. Indeed, different industries have different definitions of recalls.

Table Appx.1: This table reports relevant papers about the relation of innovation and market size after Acemoglu and Linn (2004). Details on the entries are reported as footnotes as ^{d_k}. Critiques related on the entries are reported as footnotes as ^{c_k}.

Paper	Data and sample	Unit of observation	Measurement of innovation	Estimation method	Report estimate of size	Proxy of market size
Acemoglu and Linn (2004)	US; March CPS, 1965–2000; March CPS, 1965–2000; FDA; OECD	Broad ATC-2 classes	NME ^{c_2}	QML	> 0 ^{d_1}	Demographic measures
Cerda (2007)	US; FOIA request; gov. funds on R&D $^{d_{2,i}}$; & U.S statistical abstract$^{d_{2,ii}}$; 1968-1997	15 drug categories $^{d_{2,iii}}$	NME	FE, GLS, IV, Tobit	>0 $^{d_{2,iv}}$	Demographic measures
Rake (2017)	US; R&D; FDA; OECD; ClinicalTrials.gov^{d_3}; 1974-2008	Disease classes	NDA; NME; Phase II and Phase III trials	QMLE (Poisson, 1995)	0.3444 (NME); 0.3521 (NDA)	Demographic measures
Dubois et al. (2015)	14 countries ^{d_4,c_1}; 1997-2007;IMS, WHO	Chemical entity; Dummies for ATC-1 and ATC-2	NCE (elasticity) ^{c_3}	OLS,2SLS,CF approach (Wooldr.,2002)	0.23 (average across ATC classes)	Deaths and GDP
Blume-Kohout and Sood (2013)	US; 1998-2010; Pharmaprojects $^{d_{5,i}}$; MEPS; OECD; NIH	49 therapeutic classes	R&D $^{d_{5,ii}}$	Negative Bin.; Poisson	0.26; 0.41; 0.51	Demographic shifts

List of Abbreviations:

Abbreviation	*Definition*
ATC	Anatomical Therapeutic Chemical class
CF	Control Function
CPS	Current Population Survey
FDA	Food and Drug Administration
FE	Fixed Effects
GLS	Generalised Least Squares
IMS	Intercontinental Marketing Services
IV	Instrumental Variable
MEPS	Medical Expenditure Panel Survey
NME	New Molecular Entities
NCE	New Chemical Entities
NDA	New Drug Approval
QML	Quasi Maximum Likelihood
QMLE	Quasi Maximum Likelihood Estimate
2SLS	Two Stage Least Squares
WHO	World Health Organisation

Further details:

^{d_1} The estimates suggest that a 1 percent increase in the potential market size for a drug category leads to a 6 percent increase $^{d_{2,i}}$ Data on government funds used on the R&D process of the pharmaceutical sector; $^{d_{2,ii}}$ population data for market size; $^{d_{2,iii,iv}}$ > 0 means that the exogenous increase in market size is initially associated with approximately 0.08 more drugs introduced in the market. These new drugs reduce the mortality rates of individuals aged 65 and older by 0.8 percent. This decrease in mortality rate leads to increases in market size (more demand), producing an additional increase of drugs equal to 0.096

^{d_3} Both Cerda and Rake consulted the 19^{th} edition of the Drug Information Handbook published by Lexi-Comp and the American Pharmaceutical Association (Lacy et al., 2010). This handbook is comparable to a pharmaceutical dictionary providing a list of active ingredients in the drug, the medical conditions the drug is used for, and further information such as adverse effects. The work takes into account only those medical conditions which can be found on the FDA-approved label. Hence, unlabeled and investigation uses are not present. For the period 1974 to 2008, the FDA approved 599 unique NMEs and 1665 unique NDAs. These approvals refer to the 208 diseases or medical indications analyzed in this study. However, an NME or NDA may be used as therapy for several medical indications. In this case, an NME or NDA is counted as innovation for all the medical indications for which it is approved.

^{d_4} Data come from IMS and include all product sales in 14 countries (Australia, Brazil, Canada, China, France, Germany, Italy, Japan, Mexico, Korea, Spain, Turkey, United Kingdom, USA). Dubois et al. have data on the ATC-4 (they report 607 different classes), the main active ingredient of the drug (they report 6216 different active ingredients), the name of the firm producing the drug, whether it has been licensed, the patent start date, and the format of the drug (the work reports 471 different formats). Products in the same ATC-4 by definition have the same indication and mechanism of action. The authors do not consider OTC drugs. Quantities are given in standard units: one standard unit corresponding to the smallest typical dose of a product form, as defined by IMS Health.

$^{d_{5,i}}$ Pharmaprojects trend data "snapshot"; $^{d_{5,ii}}$ (focus on R&D): focus only on one instance of innovation as explained in Hall and Rosenberg (2010). Authors specify the adoption of clinical trials (from pre-clinical Phase to Phase III) not taken from ClinicalTrials.gov (see below).

Further related evidence:

For a drug class with average Medicare market share (41%, in 2004 to 2005), Duggan and Scott Morton's result translates to an 11% increase in revenues following Medicare Part D. Their Phase I estimates correspond, for a drug class with average Medicare market share, to a 26% increase for 2004–2005, a 33% increase post-implementation in 2006 to 2007, and a lagged 51% increase in 2008 to 2010. These estimates imply an elasticity of Phase I clinical trials of 2.4 to 4.7 compared to the market size, bracketing Acemoglu and Linn's estimated elasticity of 3.5 for approved NMEs. However, when considering all clinical trials combined—including Phase III trials for supplemental indications the estimated elasticity of clinical trials with respect to market size is somewhat lower than Acemoglu and Linn's estimated elasticity of 6 for all new drug approvals, but certainly still more prominent than the Dubois et al. (2015) estimate of about 0.25. To summarise *"The results indicate that the increase in outpatient prescription drug coverage provided through Medicare Part D has had a significant impact on pharmaceutical R&D"*

Critiques:

^{c_1} Blume-Kohout and Sood (2013) states that several of the countries chosen regulate prescription drug prices, and regulations may change rapidly over time. Thus, given the lower expected profit per consumer and greater uncertainty about future profits and prices, firms' R&D decisions are likely to be less responsive to a unit change in expected revenues for all these countries combined versus the same unit change in the US market (Sood et al., 2009).
^{c_2} Blume-Kohout and Sood (2013) measured firms' innovative activities via clinical trials, whereas Dubois et al. (2015) and Acemoglu and Linn (2004) evaluate the responsiveness of approved and marketed drugs to changes in market size. Use of demographic shifts.
^{c_3} Dubois et al. (2015): the authors recognise to Blume-Kohout and Sood (2013) the fact of having exploited an innovative measure of Market Share (policy change in Medicare Part D). Use of demographic shifts.

List of controls:

Acemoglu and Linn (2004) Potential Supply-Side Determinants of Innovation (changes in scientific incentives); Proxies for pre-existing time trends across sectors; lag dependent var; life-years lost; public funding; pre-existing trends; major category trends; health insurance market size; (see page 1077 to 1080 for further details on variables).
Cerda (2007): Gov. expenditure (Medicare and social security); Gov. research efforts (grants on research); year dummies; some demographic information such as prevalence rates of disease i on males (fraction of males/white/married attending hospital due to i), blacks, whites, and married individuals, as well as the average age of individuals affected by disease i.
Rake (2017) The empirical analysis draws upon the literature concerning the "demand-pull" versus "technology-push" debate and takes into account demand- and supply-side factors as the explanatory variables for pharmaceutical innovation. Regressors used comprise knowledge stock (consisting of the scientific publications (Pub_{it}) related to medical indication i and published in year t (BioPharmInsight database); Regulatory stringency (average time between the submission of a new drug approval to the FDA and its final approval); pre-sample mean of new pharmaceuticals; mortality rate per medical indication in 1983 to account for differences in the pre-sample prevalence of medical indication; pre-sample technological opportunities are constructed as the average annual growth rate of the knowledge stock from 1979 to 1983.
Blume-Kohout and Sood (2013) prescription drugs; funding grants for each disease class.

B Time-to-event analysis

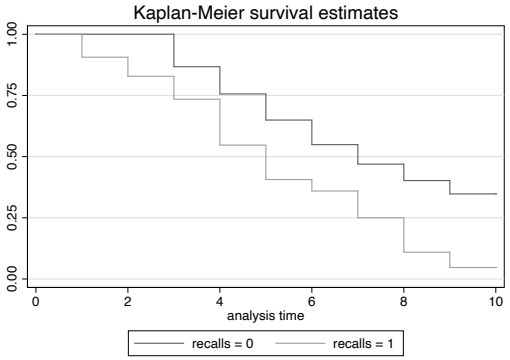


Figure Appx.1: Kaplan-Meyer time-to-event analysis. The x-axis represents the number of years until death. Recalled products (recalls =1) are already scaled down at 2 years of survival time with respect to not recalled products (recalls =0). The median survival time for recalled medicines is about 4 years, while the one for not-recalled medicines is about 7 years. The survival function of recalled products persists in its falling below the survival function of not recalled drugs. This means that recalls affect sales for a long period of time. In other words, within the market of the recalled product there will be a lack of potential sales left by the recalled products. Missed sales are hence not a temporary event. This demonstrates the length of the lack that should be covered to fill the gap provoked by the product's recall.

C Abnormal values (firm and product levels)

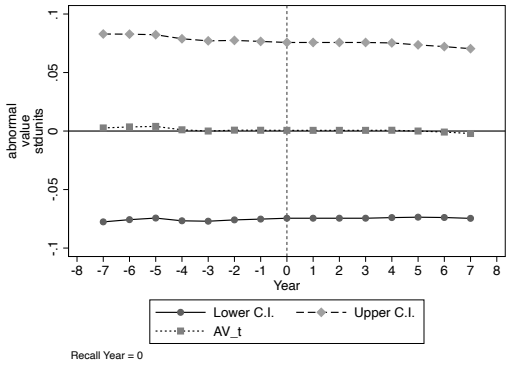
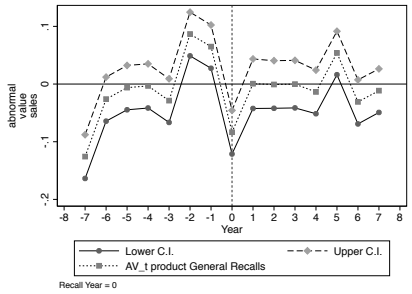
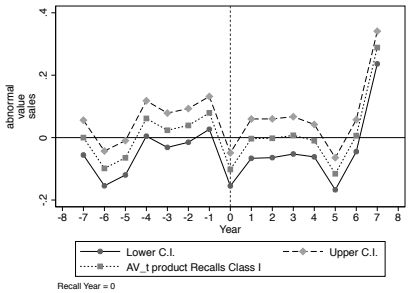


Figure Appx.2: Effect of recalls on market sales once firms having undergone a major recall are cancelled out. The absence of any effect (i.e., increases of sales due to recalls of competitors) at recall time for products other than the ones of the recalled firm witnesses the absence of compensations both at time 0 or soon after the recall.

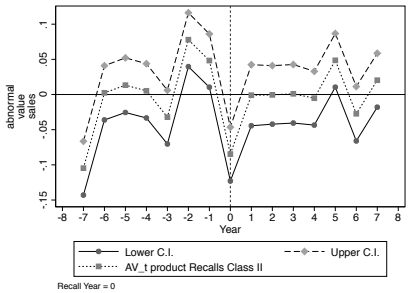
The following figures represent abnormal values at firm and product level for different typologies of recall (according to their gravity).



Product: general recalls sales

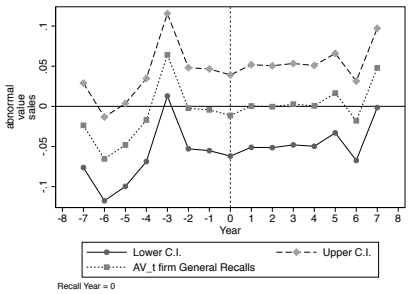


Product: Class I recalls sales

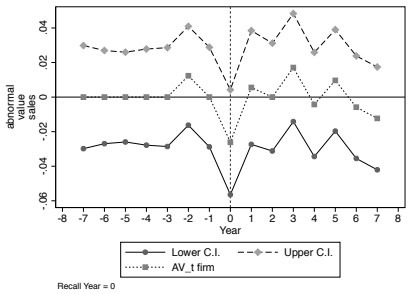


Product: Class II recalls sales

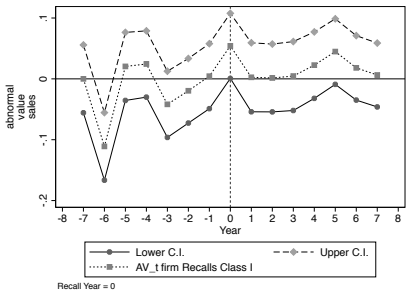
Figure Appx.3: Abnormal values for product aggregation. Years are normalised. Year 0 represents the year of recall. The three scenarios include the path of sales before and after the recall year, using three different definitions of recalls: Class I recalls, general recalls and Class II recalls. As shown in the pictures, sales at product level drop at recall year.



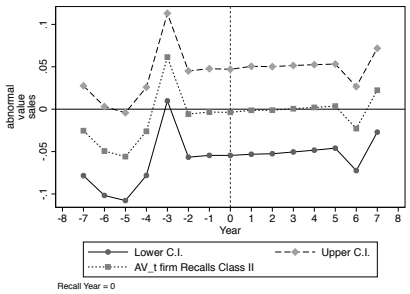
Firm: general recalls sales



Firm: Maj.recalls sales



Firm: Class I recalls sales



Firm: Class II recalls sales

Figure Appx.4: Abnormal values at firm aggregation. Years are normalised. Year 0 represents the year of recall. The four scenarios include the path of sales before and after the recall year, using four different definitions of recalls: major recalls, Class I recalls, general recalls and Class II recalls. A part from major recalls, catching firms unaware, the other types of recalls do not affect firms sales. This might be due to compensation of sales within firms.

The effect of recalls is evident for all aggregations but firm level, where the effect is not evident (future development). Possible hypotheses are detailed in the main text.

D First stage rob. checks

This section displays the significant coefficients of the first stage employing the number of patients as a measure for market size.

Table Appx.3: First stage of the robustness check using the number of patients as measure of market size

	(1)
	# patients
$\tilde{recalls}$	0.0519
	(0.0333)
$\tilde{recalls}_{t-1}$	−0.262*
	(0.149)
Year Dummies	Yes
Obs.	1056
Groups	132

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Huber-White robust and clustered at ATC-3 level standard errors are in parentheses. (1) First stage results when MEPS database is employed and market size is measured with the number of patients. Second stage results are in Tab. 8.

References

D. Acemoglu and J. Linn. Market size in innovation: Theory and evidence from the pharmaceutical industry. *The Quarterly Journal of Economics*, 119(3):1049–1090, 2004.

Z. Alsharkas. Firm size, competition, financing and innovation. *International Journal of Management and Economics*, 44(1):51–73, 2014.

R. Bala, P. Bhardwaj, and P. K. Chintagunta. Pharmaceutical product recalls: Category effects and competitor response. *Marketing Science*, 36(6):931–943, 2017.

N. Balasubramanian and J. Lee. Firm age and innovation. *Industrial and Corporate Change*, 17(5):1019–1047, 2008.

G. Ball, J. T. Macher, and A. D. Stern. Recalls, innovation, and competitor response: Evidence from medical device firms. 2018a.

G. P. Ball, R. Shah, and K. D. Wowak. Product competition, managerial discretion, and manufacturing recalls in the us pharmaceutical industry. *Journal of Operations Management*, 58:59–72, 2018b.

J. M. Bertoni, J. P. Arlette, H. H. Fernandez, C. Fitzer-Attas, K. Frei, M. N. Hassan, S. H. Isaacson, M. F. Lew, E. Molho, W. G. Ondo, et al. Increased melanoma risk in parkinson disease: a prospective clinicopathological study. *Archives of Neurology*, 67(3):347–352, 2010.

M. E. Blume-Kohout and N. Sood. Market size and innovation: Effects of medicare part d on pharmaceutical research and development. *Journal of Public Economics*, 97:327–336, 2013.

Bowe C. Merck quarterly profits hit by vioxx recall, 2005. URL `https://www.ft.com/content/b507ea92-b25b-11d9-bcc6-00000e2511c8`. [Online; accessed 15-April-2021].

T. F. Bresnahan and P. C. Reiss. Entry and competition in concentrated markets. *Journal of Political Economy*, 99(5):977–1009, 1991.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*, volume 53. Cambridge university press, 2013.

R. A. Cerda. Endogenous innovations in the pharmaceutical industry. *Journal of Evolutionary Economics*, 17 (4):473–515, 2007.

J. Cheng. An antitrust analysis of product hopping in the pharmaceutical industry. *Columbia Law Review*, 108:1471, 2008.

A. Civan and M. T. Maloney. The effect of price on pharmaceutical r&d. *The BE Journal of Economic Analysis & Policy*, 9(1), 2009.

J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185, 2003.

P. Dubois, O. De Mouzon, F. Scott-Morton, and P. Seabright. Market size and pharmaceutical innovation. *The RAND Journal of Economics*, 46(4):844–871, 2015.

M. Duggan and F. Scott Morton. The effect of medicare part d on pharmaceutical prices and utilization. *American Economic Review*, 100(1):590–607, 2010.

M. et al. Pharmaceutical antitrust law in european union. Dechert LLP, 2019.

FDA U.S. Food Drug. Cfr - code of federal regulations title 21 (subpart "a-general provisions", sec. 7.3 "definitions") no. 21cfr7.3, 2019.
https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=7.3.

L. Gallelli, C. Palleria, A. De Vuono, L. Mumoli, P. Vasapollo, B. Piro, and E. Russo. Safety and efficacy of generic drugs with respect to brand formulation. *Journal of Pharmacology & Pharmacotherapeutics*, 4 (Suppl1):S110, 2013.

P. A. Geroski and C. F. Walters. Innovative activity over the business cycle. *The Economic Journal*, 105 (431):916–928, 1995.

C. Giaccotto, R. E. Santerre, and J. A. Vernon. Drug prices and research and development investment behavior in the pharmaceutical industry. *The Journal of Law and Economics*, 48(1):195–214, 2005.

B. H. Hall and N. Rosenberg. *Handbook of the Economics of Innovation*, volume 1. Elsevier, 2010.

K. Hall, T. Stewart, J. Chang, and M. K. Freeman. Characteristics of fda drug recalls: A 30-month analysis. *American Journal of Health-System Pharmacy*, 73(4):235–240, 2016.

C. Hansen, J. Hausman, and W. Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008.

I. Hashi and N. Stojčić. The impact of innovation activities on firm performance using a multi-stage model: Evidence from the community innovation survey 4. *Research Policy*, 42(2):353–366, 2013.

V. J. Hawk, B.E. and H. H.L. Recent developments in eu merger control. *Antitrust*, (15):24, 2000.

E. Huergo and J. Jaumandreu. How does probability of innovation change with firm age? *Small Business Economics*, 22(3-4):193–207, 2004.

B. Jovanovic. Product recalls and firm reputation. Technical report, National Bureau of Economic Research, 2020.

A. Kleinknecht and B. Verspagen. Demand and innovation: Schmookler re-examined. *Research Policy*, 19(4): 387–394, 1990.

S. Klepper and F. Malerba. Demand, innovation and industrial dynamics: an introduction. *Industrial and Corporate Change*, 19(5):1515–1520, 2010.

S. Kolluru and P. Mukhopadhaya. Empirical studies on innovation performance in the manufacturing and service sectors since 1995: A systematic review. *Economic Papers: A Journal of Applied Economics and Policy*, 36(2):223–248, 2017.

M. K. Kyle and A. M. McGahan. Investments in pharmaceuticals before and after trips. *Review of Economics and Statistics*, 94(4):1157–1172, 2012.

J. O. Lanjouw. Patents, price controls, and access to new drugs: how policy affects global market entry. Technical report, National Bureau of Economic Research, 2005.

F. R. Lichtenberg. Pharmaceutical innovation as a process of creative destruction. *Knowledge Accumulation and Industry Evolution: The Case of Pharma-Biotech*, page 61, 2006.

W. Lin and J. M. Wooldridge. Testing and correcting for endogeneity in nonlinear unobserved effects models. In *Panel Data Econometrics*, pages 21–43. Elsevier, 2019.

C.-j. Luan, C. Tien, and Y.-c. Chi. Downsizing to the wrong size? a study of the impact of downsizing on firm performance during an economic downturn. *The International Journal of Human Resource Management*, 24(7):1519–1535, 2013.

F. Malerba. Innovation and the evolution of industries. In *Innovation, Industrial Dynamics and Structural Transformation*, pages 7–27. Springer, 2007.

A. Markham. Lurbinectedin: first approval. *Drugs*, pages 1–9, 2020.

L. Martin, M. Hutchens, C. Hawkins, and A. Radnov. How much do clinical trials cost?, 2017.

K. Mellahi and A. Wilkinson. A study of the association between level of slack reduction following downsizing and innovation output. *Journal of Management Studies*, 47(3):483–508, 2010.

D. Mowery and N. Rosenberg. The influence of market demand upon innovation: A critical review of some recent empirical studies. *Research Policy*, 8(2):102–153, 1979.

F. Nutarelli. At the intersection between machine learning and econometrics: theory and applications ((unpublished doctoral dissertation). *IMT Lucca for Advanced Studies*, 2021.

I. J. Onakpoya, C. J. Heneghan, and J. K. Aronson. Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis. *Critical Reviews in Toxicology*, 46(6):477–489, 2016.

W. H. Organization et al. The anatomical therapeutic chemical classification system with defined daily doses-atc/ddd. 2009.

A. Pakes and M. Schankerman. The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In *R&D, Patents, and Productivity*, pages 73–88. University of Chicago Press, 1984.

F. Pammolli, L. Magazzini, and M. Riccaboni. The productivity crisis in pharmaceutical r&d. *Nature Reviews Drug Discovery*, 10(6):428–438, 2011.

J. V. Pérez-Rodríguez and B. G. Valcarcel. Do product innovation and news about the r&d process produce large price changes and overreaction? the case of pharmaceutical stock prices. *Applied Economics*, 44(17): 2217–2229, 2012.

W. Price and I. Nicholson. Making do in making drugs: Innovation policy and pharmaceutical manufacturing. *BCL Rev.*, 55:491, 2014.

B. Rake. Determinants of pharmaceutical innovation: the role of technological opportunities revisited. *Journal of Evolutionary Economics*, 27(4):691–727, 2017.

D. Roodman. How to do xtabond2: An introduction to difference and system gmm in stata. *The Stata Journal*, 9(1):86–136, 2009.

F. M. Scherer. Demand-pull and technological invention: Schmookler revisted. *The Journal of Industrial Economics*, pages 225–237, 1982.

J. Schmookler. *Invention and Economic Growth*. Harvard University Press, 2013.

V. B. Siramshetty, J. Nickel, C. Omieczynski, B.-O. Gohlke, M. N. Drwal, and R. Preissner. Withdrawn—a resource for withdrawn and discontinued drugs. *Nucleic Acids Research*, 44(D1):D1080–D1086, 2016.

G. N. Stock, N. P. Greis, and W. A. Fischer. Firm size and dynamic technological innovation. *Technovation*, 22(9):537–549, 2002.

P. Stoneman. *Soft Innovation: Economics, Product Aesthetics, and the Creative Industries*. Oxford University Press, 2010.

G. Symeonidis. Innovation, firm size and market structure: Schumpeterian hypotheses and some new themes. 1996.

Terence N. Merck pulls vioxx painkiller from market, and stock plunges, 2004. URL `https://www.nytimes.com/2004/09/30/business/merck-pulls-vioxx-painkiller-from-market-and-stock-plunges.html`. [Online; accessed 15-April-2021].

S. Thirumalai and K. Sinha. Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences. *Management Science*, 57:376–392, 2011.

C. H. Tong, L.-I. Tong, and J. E. Tong. The vioxx recall case and comments. *Competitiveness Review: An International Business Journal*, 2009.

A. Vaishnav. Product market definition in pharmaceutical antitrust cases: Evaluating cross-price elasticity of demand. *Columbia Business Law Review*, page 586, 2011.

Ž. Vujović. A case study of the application of weka software to solve the problem of liver inflamation. 2021.