Haschka, Rouven E.; Herwartz, Helmut

Endogeneity in pharmaceutical knowledge generation: An instrument-free copula approach for Poisson frontier models

Journal of Economics & Management Strategy

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Haschka, Rouven E.; Herwartz, Helmut (2022) : Endogeneity in pharmaceutical knowledge generation: An instrument-free copula approach for Poisson frontier models, Journal of Economics & Management Strategy, ISSN 1530-9134, Wiley, Hoboken, NJ, Vol. 31, Iss. 4, pp. 942-960,
https://doi.org/10.1111/jems.12491

Received: 4 May 2021 | Revised: 21 December 2021 | Accepted: 13 May 2022

DOI: 10.1111/jems.12491

ORIGINAL ARTICLE

Journal of Economics & Management Strategy WILEY

Endogeneity in pharmaceutical knowledge generation: An instrument-free copula approach for Poisson frontier models

Rouven E. Haschka[1] | **Helmut Herwartz**[2]

[1]Institute of Econometrics and Statistics, University of Cologne, Cologne, Germany

[2]Chair of Statistics and Econometrics, University of Göttingen, Göttingen, Germany

Correspondence
Rouven E. Haschka, Institute of Econometrics and Statistics, University of Cologne, Universitätsstr. 24, D-50923 Cologne, Germany.
Email: rhaschka@uni-koeln.de

Abstract

This study provides an assessment of the R&D–patent relation of European pharmaceutical firms that are not flawed by endogeneity biases. Firms invest in R&D and generate latent knowledge which then manifests in observable patent outcomes through a Poisson model. The process of turning R&D into knowledge is described by a production process subject to inefficiency and endogeneity. To estimate a Poisson stochastic frontier model, the suggested novel copula-based approach directly accounts for the dependence between the endogenous regressors and the inefficiency component. Hence, its implementation does not require any instrumental variables. Simulation results underline that the proposed estimator outperforms conventional instrumental variable estimators. Neglecting endogeneity leads to a substantial underestimation of the R&D elasticity of patents generated in the European pharmaceutical industry.

1 | INTRODUCTION

High innovation output likely influences the business success in the long run, while ongoing investments in R&D are necessary to continuously generate knowledge and remain innovative (Griliches, 1998; Jaffe, 1989; Siebert, 2017). Since innovation-relevant knowledge generated is usually unobserved, it can only be approximated, for instance, by the number of patent grants (Griliches, 1998; Jaffe, 1989; Siebert, 2017). As patent counts are discrete, recent empirical models investigating the R&D–patent relationship rely on count data models, such as Poisson regressions (Fé & Hofler, 2013). Considering inefficiencies within innovation processes has also gained increasing attention in the recent literature (Fu & Yang, 2009; Pieri et al., 2018). Regarding each firm as a "producer" of knowledge, knowledge generation can be described by means of a production process (Daub, 2008; Dittmer & Strätz, 2012; Griliches, 1998; Siebert, 2017). At the same time, this embedding allows for determining firm-specific efficiency within the R&D–patent relationship (for literature reviews, see Abbas et al., 2014; Czarnitzki & Toole, 2011; Holgersson, 2013).

Quantitative assessments of innovative efficiency are often performed in the framework of stochastic frontier analysis (SFA, for methodological reviews, see Amsler et al., 2016; Coelli et al., 2005; Kumbhakar & Lovell, 2003). When modeling stylized innovation processes, a particular model specification issue arises if firms have some a priori information on eventually inefficient knowledge generation, and they adjust the level of R&D expenditures accordingly

(Siebert, 2017). However, such dependence is unobserved to the analyst and has become one of the most common forms of endogeneity in the context of innovation efficiency modeling, which makes drawing valid causal conclusions virtually impossible.[1] Maximum-likelihood-based instrumental variable (ML-IV) estimation methods have recently been proposed to handle endogeneity in SFA models (Amsler et al., 2016; Kutlu, 2010; Prokhorov et al., 2020; Tran & Tsionas, 2013). However, a general drawback of ML-IV methods is, that they rely upon the availability of consensual outside information to construct the instruments (Windmeijer & Santos Silva, 1997). In search of suitable instrumental information, empirical studies typically condition on "internal" information, such as lagged explanatory variables, which are often weak and fail to fully cope with endogeneity (N. Wang & Hagedoorn, 2014).

Seeing that suitable—that is, "external"—instrument information in frontier models is often scant, unavailable, or weak, this study proposes a frontier model for Poisson distributed outcomes that we conceptualize as an extension of the IV-free joint regression model using copulas introduced by Park and Gupta (2012). Given that copula techniques have been successfully adopted for Gaussian SFA's (Tran & Tsionas, 2015), we extend it towards a Poisson SFA model. Although joint estimation models have been intensively studied (Becker et al., 2021; Haschka, 2022; Tran & Tsionas, 2021) and used in various empirical settings (Datta et al., 2017; Haschka et al., 2021; Manchanda et al., 2004; Zhang et al., 2017), their general applicability in Poisson SFA models has not yet been addressed. The method is based on a copula function to directly model dependencies of endogenous regressors and inefficiency and thus does not require any external information. As an additional model feature, we adopt scaling properties (H.-J. Wang & Schmidt, 2002), which enable a structured perspective on potential determinants of inefficiency. We conduct Monte Carlo experiments and assess the performance of the proposed estimator in comparison with stylized ML as in Fé and Hofler (2013), and IV estimation with "weak" and "strong" instrumentation (N. Wang & Hagedoorn, 2014). Simulation-based evidence suggests that the copula estimator (i) provides unbiased estimates if model regressors are correlated with unobserved inefficiency, (ii) copes with reverse causality, and (iii) delivers quantitative assessments of the correlation between inputs and inefficiencies.

The empirical analysis aims at an unbiased understanding of knowledge generation in a cross section of 137 unaffiliated European pharmaceutical firms. The focus on the pharmaceutical industry follows similar rationales as the analysis of N. Wang and Hagedoorn (2014) and Fé and Hofler (2013). On the one hand, the empirical evidence on knowledge generation has highlighted important sector-specific characteristics. Hence, approaching knowledge generation conditional on the pharmaceutical industry does not suffer from unobserved cross-industry heterogeneities as a potential channel of endogeneity. Moreover, the cross-sectional perspective immunizes the analysis against interindustry variation in other factors such as technological opportunities and the effectiveness of patents as a means of appropriating returns from R&D. On the other hand, the patenting rate in the pharmaceutical industry is among the highest in high-technology sectors (Chen & Chang, 2010; Danguy et al., 2019; Griliches, 1998). Significant differences in gross margins have been primarily attributed to the superior records of pharmaceutical companies in protecting their innovations by patents. For instance, pharmaceutical gross margins are nearly twice as large as those in the semiconductor industry (Chen & Chang, 2010). Moreover, it is worth highlighting that a major fraction of costs is incurred in the R&D stage, with the "R proportion" holding outstanding relevance for this sector. In this regard, endogeneity biases that originate in the personal attributes of the drug developers hold specific relevance for the analysis. For instance, experienced researchers could be expected to reduce trial-and-error experiments in laboratories. In a similar vein, a research staff familiar with patenting processes and authorities has the potential to position a firm among the top performers within the pharmaceutical industry. Finally, given that drug development touches sensitive societal fields, such as public health and health policy, both the composition and experience of the research staff could hold particular importance for effective knowledge generation in the pharmaceutical industry in comparison with other sectors. Going beyond endogeneity concerns, it is worth noting that for this industry research taking place in universities—that is, pharmaceutical faculties—could act as a major source of knowledge transfer and the acquisition of key business skills (Haschka & Herwartz, 2020; Powell, 1998). Hence, the description of knowledge generation subject to intraindustry competition is likely to benefit from the potential to scale inefficiencies in SFA models, as in H.-J. Wang and Schmidt (2002).

Our empirical results enable two major conclusions. First, we diagnose with high significance regressor endogeneity challenging the estimation of the marginal causal effect of R&D expenditures for steering patenting outcomes in the European pharmaceutical industry. Specifically, this effect is underestimated by almost 50% under exogeneity assumptions. Given empirical evidence of R&D-inefficiency dependence, it should be acknowledged that firms adjust their R&D inputs according to their own inefficiency. As this dependence is unobserved to analysts or competitors, active management should crucially take account of feedback between R&D and inefficiency for the

purposes of market screening and unraveling competitive efficiency levels. Second, pointing to important knowledge spillovers, pharmaceutical firms in proximity to a university are on average 56% more efficient than firms without such a neighbor. Accordingly, the acquisition of knowledge from neighboring universities largely improves innovativeness and complements in-house innovative activities.

In Section 2, we establish an intraindustry model of patent generation at the firm level, outline the Poisson SFA model with inefficiency scaling, and describe the copula-based estimation approach. In Section 3, we examine the finite sample performance of alternative estimators by means of a Monte Carlo study. Section 4 provides a detailed comparative analysis of patent outputs of the European pharmaceutical industry, before Section 5 concludes. Detailed simulation results and a set of complementary simulations are provided in the Supporting Information.

2 | LATENT KNOWLEDGE GENERATION AND IV-FREE ESTIMATION

The efficient management of knowledge generation is crucial for long-term business success and requires an unbiased assessment of the translation from R&D expenditures into knowledge manifestations, such as patents. In this section, we first sketch a production model to describe the transition of innovation inputs into knowledge subject to inefficiencies that emerge from an intraindustry perspective covering a cross section of competing firms. We establish an empirical model that further advances the approach taken by Fé and Hofler (2013) and apply a production function approach in which patents are the realization of a latent knowledge generation process. Second, we turn to an encountering of potential sources of endogeneity as pertinent threats to an unbiased analytic assessment of causal effect structures. In the third instance, we develop an explicit SFA count data model and suggest an IV-free approach to model estimation.

2.1 | Approximating the firm-level innovation process

An extensive body of literature has shown that innovation processes are more complex than stylized translations of R&D into knowledge (for a systematic review, see Hall et al., 2005). On the one hand, knowledge generated is unobserved and thus remains latent by nature. When using the number of patents as an approximation, it is necessary to take into account the discrete nature of patent outcomes (Hall et al., 2005). On the other hand, it has become pertinent to allow for inefficiencies within the process turning R&D into knowledge (Griliches, 1998; Siebert, 2017). Against this background, we model innovation processes by means of Poisson frontier models to explicitly consider these particularities (Fé & Hofler, 2013). Stylized characteristics of industrial R&D activity suggest that it is reasonable to assume that patent counts follow a Poisson distribution (for pioneering work, see Hausman et al., 1984; Jaffe, 1986; Pakes & Griliches, 1984). Especially in the pharmaceutical industry, patents can be thought of as measuring the number of successful outcomes from a large but unobserved number of projects within a firm's R&D lab, each of which has a small probability of success. Furthermore, intraindustry companies are subject to similar restrictions within the market, for example, to the likelihood that the patent office responds to an inquiry from a patent applier. Nevertheless, the success in patenting likely depends on the experience of researchers or personal relations in dealing with the patent office. Accordingly, firms with less experienced researchers in this regard are less successful in patenting, giving rise to inefficiencies in patent generation.

To make these considerations explicit, consider the following model for $i = 1, 2, ..., N$ firms:

$$\text{patents}_i \sim \text{Pois}(\text{latent knowledge}_i) \text{with} \tag{1}$$

$$\text{latent knowledge}_i = f(X_i; \beta) \times TE_i. \tag{2}$$

In (2), $\text{latent knowledge}_i$ denotes the knowledge generated from a vector of innovation inputs X_i subject to a production function $f(X_i; \beta)$, for example, of Cobb–Douglas type (Haschka & Herwartz, 2020; Siebert, 2017), and technical efficiency $TE_i \in [0, 1]$. Both knowledge and technical efficiency are known to the firm, but unobserved by the analyst. Since patents can be seen as an observable manifestation of knowledge, we assume that $\text{latent knowledge}_i$ turns into patents_i through a Poisson distribution, that is, $\mathbb{E}[\text{patents}_i | X_i, TE_i] = \text{Pois}(\text{latent knowledge}_i)$. Further, the

Poisson model comes with the advantage that latent knowledge$_i$ does not need to be observed and the number of patent applications does not need to be known.

2.2 | Endogeneity in knowledge generation models

A particular empirical challenge arises if firms have some a priori information on eventually inefficient knowledge generation and adjust the level of R&D expenditures accordingly (Siebert, 2017). In (2), $\mathrm{Cov}[X_i, TE_i] \neq 0$ introduces endogeneity to the model. However, unbiased assessments are pivotal to unravel efficiency differences for the strategic positioning of firms to improve the processes of knowledge generation and patenting.

Regarding the origin of endogeneity in knowledge generation models, one might consider three nonexclusive channels. First, patent generation is typically seen as a reflection of R&D expenditures. However, patents themselves might also lead to R&D activities (Correa, 2004; Scherer, 2001). For instance, Czarnitzki and Toole (2011) argue that patents reduce uncertainty and thereby influence R&D decisions. Accordingly, endogeneity might result from patterns of reverse causality (Arora et al., 2003; Cincera, 1997). Technology shocks that affect investment decisions provide a second origin of endogeneity. For instance, Alexopoulos (2011) has shown that technological shocks are positively linked to inputs of knowledge production, such as R&D. Since such shocks are unobserved to analysts, they show up in model terms assessing productive efficiency. As a result, technology shocks induce correlation between model residuals and explanatory input variables (Alexopoulos, 2011; Schilling, 2015). Third, the experience of researchers or personal relations in dealing with the patent office are also likely to shape efficiency. Seeing a major fraction of costs being incurred in the R&D stage with outstanding relevance of the "R proportion," this channel appears to be particularly important in the pharmaceutical industry.

Obtaining unbiased estimates under endogeneity invokes further complications if productive output is quantified in terms of a count variable, since discrete distributions are often more complex to handle in comparison with stylized continuous output measures (Fahrmeir & Lang, 2001). For instance, Fé and Hofler (2013) adopt a Poisson frontier model to estimate a knowledge generation function for a number of patents awarded to pharmaceutical firms conditional on supposedly exogenous R&D expenditures. Cincera (1997) estimates the effect of R&D expenditures on discrete patent outcomes by means of lagged explanatory variables as instruments. However, after testing instrument validity, the author finds that any (weak) exogeneity hypothesis for R&D is rejected with conventional significance. Nonetheless, seeing strong arguments for the prevalence of endogeneity (Cincera, 1997; Correa, 2004; Scherer, 2001), endogeneity could have biased the understanding of the R&D–patent relation obtained within the framework of standard Poisson frontier regressions.

2.3 | Copula-based Poisson frontier modeling

Figure 1 presents a flowchart of the proposed model. When conducting market or competitor analyses, (i) the R&D-inefficiency dependence (indicated by the double arrow) must be taken into account and (ii) approximating observed patent outcomes by means of a count data model (e.g., Poisson regressions) is crucial, since otherwise both estimated R&D elasticities and firm-specific inefficiency measures are biased. Determining competitors' efficiency holds particular relevance for the strategic positioning of the firm, since inefficient firms are likely to perform weaker regarding long-run business success.

Further, the consideration of external determinants of inefficiency allows one to disentangle firm-specific unobserved inefficiency from observed environmental influences, such as university spillovers (Siegel et al., 2003). Especially in the pharmaceutical industry, similar research taking place in universities with pharmaceutical faculty

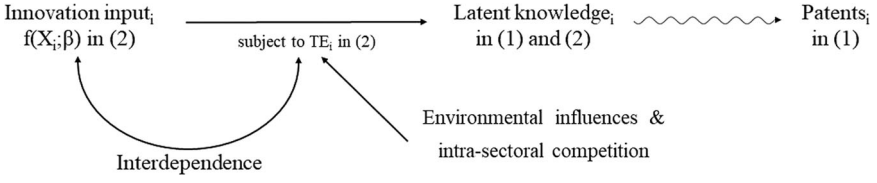

FIGURE 1 Stylized illustration of the proposed model

likely promotes local knowledge that can be absorbed and thus benefits firm performance (Szücs, 2018; Yu-ming, 2009). For instance, knowledge externalities originating in universities likely complement the in-house research of firms located in geographic proximity (Blume & Fromm, 2000; Oshri et al., 2015). In this respect, unraveling sources of (in)efficiency is pivotal when conducting market or competitor analyses.

To estimate the relationship in (1) and (2) summarized in Figure 1, we start from the copula approach of Tran and Tsionas (2015) that copes with endogenous interrelations, and generalize it such that it is applicable to Poisson distributed outcomes. The approach is particularly useful, as it avoids the search for suitable instrumental information. More specifically, we construct the joint distribution of the endogenous regressors and the inefficiency component conditional on the assumption that the outcome variable is Poisson distributed. As an additional model feature, we adopt scaling properties and thereby allow for observable determinants of inefficiency.

2.3.1 | The Poisson stochastic frontier model

For the general model representation, let Y_i denote the discrete output variable, for example, the number of patents granted to firm i. By assumption, Y_i exhibits a Poisson distribution for a sample of N producers, that is,

$$Y_i \sim \text{Pois}(\lambda_i), \quad i = 1, 2, ..., N, \tag{3}$$

where λ_i is the latent mean of the Poisson distribution, which is presumed behind the issuance of observable patent counts of firm i, that is, latent knowledge. The structured SFA model relates the unknown mean to observable covariates, that is, innovation inputs, and a stochastic component that accounts for deviations from fully efficient knowledge generation. Specifically,

$$\lambda_i = \exp\left\{\kappa + \boldsymbol{x}_i'\boldsymbol{\alpha} + \boldsymbol{z}_i'\boldsymbol{\beta} - u_i\right\}, \tag{4}$$

$$u_i = h_i u_i^*, \tag{5}$$

$$u_i^* \sim \mathcal{N}^+\left(\mu, \sigma_{u^*}^2\right). \tag{6}$$

Equation (4) describes the innovative production process with λ_i capturing the unobserved knowledge generated. Exogenous inputs are collected in the vector $\boldsymbol{x}_i$. Extending the model in Fé and Hofler (2013), we allow for endogenous input variables that are correlated with u_i^* and collected in the p-dimensional vector $\boldsymbol{z}_i$. For both categories of input variables, we assume linear influences on λ_i with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Finally, the intercept κ accounts for total factor productivity (Hausman et al., 1984).[2]

Similar to common approaches in SFA modeling, the term u_i describes production inefficiencies, that is, $u_i \in [0, \infty)$. In (5), we extend the basic Poisson SFA by introducing the function h_i to scale the distribution of u_i^*, that is, $u_i = h_i u_i^*$, where the underlying stochastic component u_i^* exhibits a truncated normal distribution (Kumbhakar & Lovell, 2003). By virtue of the scaling index $h_i = \exp(\boldsymbol{s}_i'\boldsymbol{\delta})$, it is possible to relate inefficiencies to (observable) environmental influences, for example, spillover effects, collected in the vector $\boldsymbol{s}_i$ (H.-J. Wang & Schmidt, 2002).[3]

2.3.2 | Maximum-likelihood estimation

Under the assumption that all inputs are exogenous, that is, $\boldsymbol{\beta} = 0$ in (4), the conditional distribution of Y_i given $\boldsymbol{x}_i$, $\boldsymbol{s}_i$, and u_i^* can be derived by multiplying the respective marginals, that is, the probability mass function of Y_i and the probability density function of u_i^*. Accordingly, the likelihood for a sample of N observations is

$$L_{\mathrm{ML}}(\theta|Y,\boldsymbol{x},\boldsymbol{s}) = \prod_{i=1}^{N}\int_0^{\infty}\underbrace{\left[\frac{\lambda_i^{Y_i}\exp(-\lambda_i)}{Y_i!}g\left(u_i^*\right)\right]}_{f^*\left(u_i^*,Y_i\right)}du_i^*, \tag{7}$$

where $\theta = \left(\kappa, \boldsymbol{\alpha}', \boldsymbol{\delta}', \mu, \sigma_{u^*}^2\right)'$ is the vector of model parameters to be estimated and $g\left(u_i^*\right)$ is the density of the half-normal distribution (see Equation 6). Owing to the latency of u_i^*, it has to be integrated out from the joint density for each observation i. Since the likelihood lacks a closed form, the integral has to be evaluated by means of numerical integration procedures.[4] The estimator resulting from the maximization of $\ln L_{\mathrm{ML}}$ mimics the estimator in Fé and Hofler (2013) and is henceforth denoted as $\hat{\theta}_{\mathrm{ML}}$.

2.3.3 | Joint estimation by means of copulas

Now consider the case when endogenous covariates enter the model, that is, $\boldsymbol{\beta} \neq 0$ in (4). Let $f\left(z_{1i}, ..., z_{pi}, u_i^*, Y_i\right)$ denote the joint density of $(z_{1i}, ..., z_{pi})$, u_i^* and Y_i for a single firm i. The joint distribution captures the entire dependence of the endogenous regressors and the inefficiency term (Park & Gupta, 2012). Hence, the endogeneity problem can be solved by maximizing $f\left(z_{1i}, ..., z_{pi}, u_i^*, Y_i\right)$ instead of $f^*\left(u_i^*, Y_i\right)$ in (7). Noticing that the conditional probability mass function of Y_i corresponds to $f\left(z_{1i}, ..., z_{pi}, u_i^*\right)p(Y_i)$, only the joint density $f\left(z_{1i}, ..., z_{pi}, u_i^*\right)$ has to be determined in advance. Following Tran and Tsionas (2015), we suggest a copula function to construct and estimate $f\left(z_{1i}, ..., z_{pi}, u_i^*\right)$ conditional on the observed data. According to Sklar's theorem (Sklar, 1959), the joint density can be given as

$$f\left(z_{1i}, ..., z_{pi}, u_i^*\right) = c\left(F_1(z_{1i}), ..., F_{pi}(z_{pi}), G\left(u_i^*\right)\right)g\left(u_i^*\right)\prod_{j=1}^{p} f_j(z_{ji}), \tag{8}$$

such that $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_0^{\infty} f\left(z_{1i}, ..., z_{pi}, u_i^*\right)dz_{1i}\cdots dz_{pi}\,du_i^* = 1$. In (8), capital and small letters denote marginal cumulative distributions and density functions, respectively, and c is the copula density, which links its components by means of the probability integral transform (Li & Racine, 2007; Nelsen, 2006). When replacing $g\left(u_i^*\right)$ in (7) by $f\left(z_{1i}, ..., z_{pi}, u_i^*\right)$ from (8), the likelihood accounts for endogenous interrelations between $(z_{1i}, ..., z_{pi})$ and u_i^*, since the dependence structure is completely characterized by means of the joint density. As a result, the copula-based likelihood is

$$\begin{aligned} L_{\mathrm{Cop}}(\theta|Y,\boldsymbol{x},\boldsymbol{z},\boldsymbol{s}) &= \prod_{i=1}^{N}\int_0^{\infty}\left[\frac{\lambda_i^{Y_i}\exp(-\lambda_i)}{Y_i!}c\left(F_1(z_{1i}), ..., F_p(z_{pi}), G\left(u_i^*\right)\right)g\left(u_i^*\right)\prod_{j=1}^{p}f_j(z_{ji})\right]du_i^* \\ &\propto \prod_{i=1}^{N}\int_0^{\infty}\left[\frac{\lambda_i^{Y_i}\exp(-\lambda_i)}{Y_i!}c\left(F_1(z_{1i}), ..., F_p(z_{pi}), G\left(u_i^*\right)\right)g\left(u_i^*\right)\right]du_i^*. \end{aligned} \tag{9}$$

Before proceeding with the implementation, we briefly describe model identification (for a detailed outline of identification in copula-based joint estimation models, see Haschka, 2022). Recall that the copula function disentangles joint variation among endogenous regressors from remaining unexplained variation due to u^*. Accordingly, the model is identified as long as the marginal distributions of the endogenous regressors differ from the half-normal distribution assumed for u^*. Subsequently, the conditional expectation of u^* given $\boldsymbol{z}$ becomes a nonlinear function and this nonlinearity allows us to identify the linear regression coefficients in (4) through the joint distribution. However, if z_j is indeed half normal, the model is not identified because the copula can no longer distinguish the variations as a result of endogenous regressors from the variation caused by u^* (Tran & Tsionas, 2015). Consequently, the identification problem has important implications for empirical applications. If one assumes a half-normal distribution for the inefficiency term, it has to be examined whether the endogenous regressors exhibit half-normal distributions before application. As the copula further

requires sufficient variation in regions where the conditional expectation is nonlinear, binary endogenous regressors also cannot be modeled.[5]

2.3.4 | Implementation of the copula-based estimator

Regarding the copula c, we use the Gaussian copula, which is applicable in many settings and has several desirable theoretical properties (Danaher & Smith, 2011). For instance, the Gaussian copula is more flexible for multidimensional modeling than Archimedean copulas (for further arguments in favor of the Gaussian copula, see Park & Gupta, 2012; Tran & Tsionas, 2015).[6] The density of the Gaussian copula $c\left(F_1(z_{1i}), ..., F_p(z_{pi}), G\left(u_i^*\right); \Sigma\right)$ in (9) is given by

$$c(\cdot) = (\det(\Sigma))^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\Phi^{-1}(F_1(z_{1i})), ..., \Phi^{-1}(F_p(z_{pi})), \Phi^{-1}\left(G\left(u_i^*\right)\right)\right)'\right.$$
$$\left.\frac{1}{2} \times (\Sigma^{-1} - I_{p+1})\left(\Phi^{-1}(F_1(z_{1i})), ..., \Phi^{-1}(F_p(z_{pi})), \Phi^{-1}\left(G\left(u_i^*\right)\right)\right)\right\}, \tag{10}$$

where Φ^{-1} is the quantile function of the standard normal distribution and Σ is the correlation matrix formalizing the dependence among the respective components in c. As a particular merit of the Gaussian copula, the off-diagonal entries in Σ coincide with the Pearson correlations. Hence, the copula delivers quantitative assessments of the correlation between endogenous regressors and inefficiency, which is a valuable diagnostic when modeling economic production patterns.

The parametric forms of $G\left(u_i^*\right)$ and $g\left(u_i^*\right)$ follow from the assumption in (6). The marginal densities $f_j(z_{ji})$ in (9) do not contain any parameter of interest and can be dropped from the likelihood, since they enter as normalizing constants. Building on Tran and Tsionas (2015), we replace $F_1(z_{1i}), ..., F_p(z_{pi})$ by their respective empirical counterparts before maximizing the likelihood. Given observed samples of z_{ji}, $j = 1, ..., p$; $i = 1, ..., N$, we use the empirical cumulative distribution function of z_j, that is, $\tilde{F}_j = \frac{1}{N+1}\sum_{i=1}^{N} \mathbb{1}(z_{ji} \leq z_{0j})$.[7] Seeing that the empirical cumulative distribution function is a step function by definition, the model is at odds with the assumptions behind Sklar's theorem (Sklar, 1959), which states that the copula in (10) is only uniquely defined if its margins are continuous. However, we show in the Supporting Information that such a violation is not problematic, as the model can be identified regardless of the noncontinuity of $\tilde{F}_j$, even in small samples. Finally, $\ln L_{\text{Cop}}$ (see Equation 9) is maximized with respect to $\theta = \left(\kappa, \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\delta}', \mu, \sigma_{u^*}^2, \text{vechl}[\Sigma]'\right)'$, where vechl[$\Sigma$] stacks the lower diagonal parameters of Σ in a column vector comprising p unknown parameters.[8] To test for the presence of regressor endogeneity, either model selection criteria such as Akaike's Information Criteria or Bayesian Information Criteria or likelihood-based tests can be used. Regarding the latter, the correlation coefficients in vechl[Σ] capture the endogeneity of the regressors in $\boldsymbol{z}$. Under regularity conditions, the ML estimators of model parameters are asymptotically normally distributed. Thus, the Wald statistic for testing the null hypothesis of the exogeneity of regressor z_p, that is, $H_0 : \rho_{z_p} = 0$ is $\hat{\rho}_{z_p} / \text{SD}(\hat{\rho}_{z_p})$. The likelihood-ratio and score tests for endogeneity can also be derived in a straightforward manner under the copula model. In the following, the resulting estimator is denoted $\hat{\theta}_{\text{Cop}}$. In summary, the proposed approach is IV-free and copes with interrelationships between input selection and productive inefficiency without any additional distributional assumptions (Tran & Tsionas, 2015).

3 | SIMULATION STUDY

To examine the finite sample properties of the proposed copula estimator ($\hat{\theta}_{\text{Cop}}$), we conduct some Monte Carlo experiments. The IV-free estimator promises unbiased estimation under correlation between observed production inputs and unobserved stochastic inefficiency. however, in the case of exogeneity, it might suffer from inefficient parameter assessment in comparison with stylized ML estimation ($\hat{\theta}_{\text{ML}}$). Alternatively, ML-IV estimators could hold interest to handle endogeneity in Poisson frontier models (Drivas et al., 2014). In sum, our comparative investigation

involves four estimators, namely, the copula estimator, the common ML estimator, and two variants of ML-IV estimation, denoted $\hat{\theta}_{\text{IV1}}$ and $\hat{\theta}_{\text{IV2}}$ below, which differ in the choice of the external information employed. We next explain the generation of Poisson distributed stochastic outcomes and encounter the alternative estimators entering our comparative analysis. We evaluate distributional properties of rival parameter estimates by means of stylized box plots in the main text of this section. More detailed simulation results on the average performance of alternative estimators, as well as discussions on inferential results are provided in the Supporting Information and briefly discussed.

3.1 | Data generation

Discrete outcomes $Y_i \sim \text{Pois}(\lambda_i)$, $i = 1, 2, ..., N$, are drawn after structuring the Poisson mean λ_i (see Equation 6) as

$$\lambda_i = \exp\left(\kappa + \alpha x_i + \beta z_i - h_i u_i^*\right), \tag{11}$$

where κ is an intercept parameter, $x_i \sim N(0, 1)$ is an exogenous input variable and potential endogeneity is channeled through a correlation between the second input variable z_i and productive inefficiency u_i^*. To be explicit about the origin of potential endogeneity, it holds that

$$z_i = \gamma w_i + \eta_i \quad \text{and} \quad \text{Corr}\left[u_i^*, \eta_i\right] = \varphi, \tag{12}$$

where $w_i \sim N(0, 1)$ is strictly exogenous and might be considered as "external" information (if $\gamma \neq 0$) to enable ML-IV estimation (for a similar approach, see Drivas et al., 2014). Apparently, the joint generation of u_i^* and η_i is crucial for the emergence of endogeneity in the form of a simultaneous input adaptation to technology shocks. Specifically, we draw u_i^* and η_i jointly from a normal copula with correlation $\text{Corr}\left[u_i^*, \eta_i\right] = \varphi$ and a half-normal marginal distribution for u_i^*, that is, $u_i^* \sim \mathcal{N}^+\left(0, \sigma_{u^*}^2\right)$. The choice of the marginal distribution of η_i allows us to distinguish two frameworks of data generation and model estimation. On the one hand, we consider a benchmark scenario for which the underlying true distributions are in line with density specifications required for the purposes of ML-IV estimation. On the other hand, we consider a scenario of intrinsic model misspecification in the sense that actual shocks η_i are drawn from a skewed distribution with excess kurtosis. Subsequently, η_i is falsely considered Gaussian for ML-IV estimation. For *benchmarking* purposes and the consideration of *density misspecification* the marginal distributions of η_i are

$$\text{(i) } \eta_i \sim \mathcal{N}\left(0, \sigma_\eta^2\right) \quad \text{and} \quad \text{(ii) } \eta_i \sim ((\chi^2(2) - 2)/2)\sigma_\eta, \tag{13}$$

respectively. To enhance model flexibility, the inefficiencies u_i^* enter the Poisson mean in scaled form $u_i = u_i^* h_i$,

$$h_i = \exp(\delta s_i), \quad \text{where } s_i \sim N(0, 1). \tag{14}$$

Regarding parameter settings, we set the values of φ to analyze the performance of alternative estimators under exogeneity ($\varphi = 0$) and degrees of moderate ($\varphi = .4$) and strong endogeneity ($\varphi = .8$). The remaining model parameters are set as $\kappa = 2$, $\alpha = \beta = \delta = .5$, $\sigma_\eta^2 = \sigma_{u^*}^2 = 1$ and $\gamma = 1$. We consider finite samples of alternative sizes $N = \{100, 250, 500\}$, and regard $N = 500$ as sufficient to unravel the asymptotic properties of the considered estimators. Each simulation experiment is replicated 1000 times. All computations were performed in R. We next encounter the alternative estimators and their adaptation to the simulated data-generating models (DGMs).

3.2 | Estimators

3.2.1 | The IV-free copula estimator

The proposed estimator $\hat{\theta}_{\text{Cop}}$ results from linking the observed endogenous variables to the unobserved inefficiency term u_i^* by means of a copula function. Since the DGM contains one endogenous regressor z_i, the copula in (9) and (10) is bivariate with density function

$$c\left(G\left(u_i^*\right), \tilde{F}(z_i)\right) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{ \frac{-\rho^2\left(\left(\Phi^{-1}\left(G\left(u_i^*\right)\right)\right)^2 + (\Phi^{-1}(\tilde{F}(z_i)))^2\right)}{2(1-\rho^2)} + \frac{2\rho\left(\Phi^{-1}\left(G\left(u_i^*\right)\right)\Phi^{-1}(\tilde{F}(z_i))\right)}{2(1-\rho^2)} \right\}. \tag{15}$$

In (15), ρ denotes the correlation between z_i and u_i^* based on the probability integral transform, and hence it governs the "degree" of endogeneity. Unlike alternative ML-IV approaches, the implementation of the copula estimator outlined in (9) neither requires instrumental variable information, for example, w_i, nor any distributional assumption regarding η_i.

3.2.2 | ML-IV estimator

Seeing the alternative parameter selections for the correlation parameter $\varphi = \{0, .4, .8\}$, $\hat{\theta}_{\text{ML}}$ is the estimator of choice and most likely to outperform $\hat{\theta}_{\text{Cop}}$ in case of $\varphi = 0$. Under alternative scenarios of $\varphi = \{.4, .8\}$, $\hat{\theta}_{\text{Cop}}$ is likely to benefit from its consistency, although it is unclear how it performs in comparison with suitably implemented ML-IV-based estimators. Owing to the stylized DGM in (12) and (13), it is straightforward to compare the copula estimator with ML-IV estimation. Assuming that the proper instrument w_i is available and the distribution of η_i is known, the likelihood of an ML-IV-type estimator obtains as

$$L_{\text{IV1}}(\theta|Y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}) = \prod_{i=1}^{N} \int_0^{\infty} \left[\frac{\lambda_i^{Y_i}\exp(-\lambda_i)}{Y_i!} f^{**}\left(\eta_i, u_i^*\right) \right] du_i^* \tag{16}$$

according to Drivas et al. (2014). The joint distribution $f^{**}\left(\eta_i, u_i^*\right)$ in (16) can be computed by means of Sklar's theorem (Sklar, 1959). The marginal distribution $g\left(u_i^*\right)$ follows from the assumption of half normality. Moreover, shocks η_i obtain as $\eta_i = x_i - \gamma w_i$, and exhibit a marginal normal distribution $\eta_i \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$ by assumption. The log-likelihood implied by (16) allows for straightforward maximization with respect to $\theta = \left(\kappa, \alpha, \beta, \delta, \gamma, \sigma_{u^*}^2, \sigma_\eta^2, \varphi\right)'$ to obtain $\hat{\theta}_{\text{IV1}}$. The actual performance of this estimator likely differs for the two scenarios encountered in (13), for which only the first underlying distribution corresponds to a correctly specified model density in (16). As further estimates and valuable diagnostic information, $\hat{\theta}_{\text{IV1}}$ includes a variance estimate for η_i and the estimated correlation between η_i and u_i^* ($\hat{\varphi}$). To unravel potential risks associated with the choice of instrumental information, we next consider a counterpart of $\hat{\theta}_{\text{IV1}}$ where the selected instrument is weak.

We further modify the ML-IV estimator in (16) such that the instrument w_i is replaced by $w_i^* = \exp(w_i)$. Subsequently, w_i^* is scaled to exhibit the same mean and variance as w_i, and $\eta_i = z_i - \gamma w_i^*$ enters the model density $L_{\text{IV2}}(\theta|Y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w})$, which is otherwise in full analogy to L_{IV1} in (16). Since w_i^* provides less precise information for

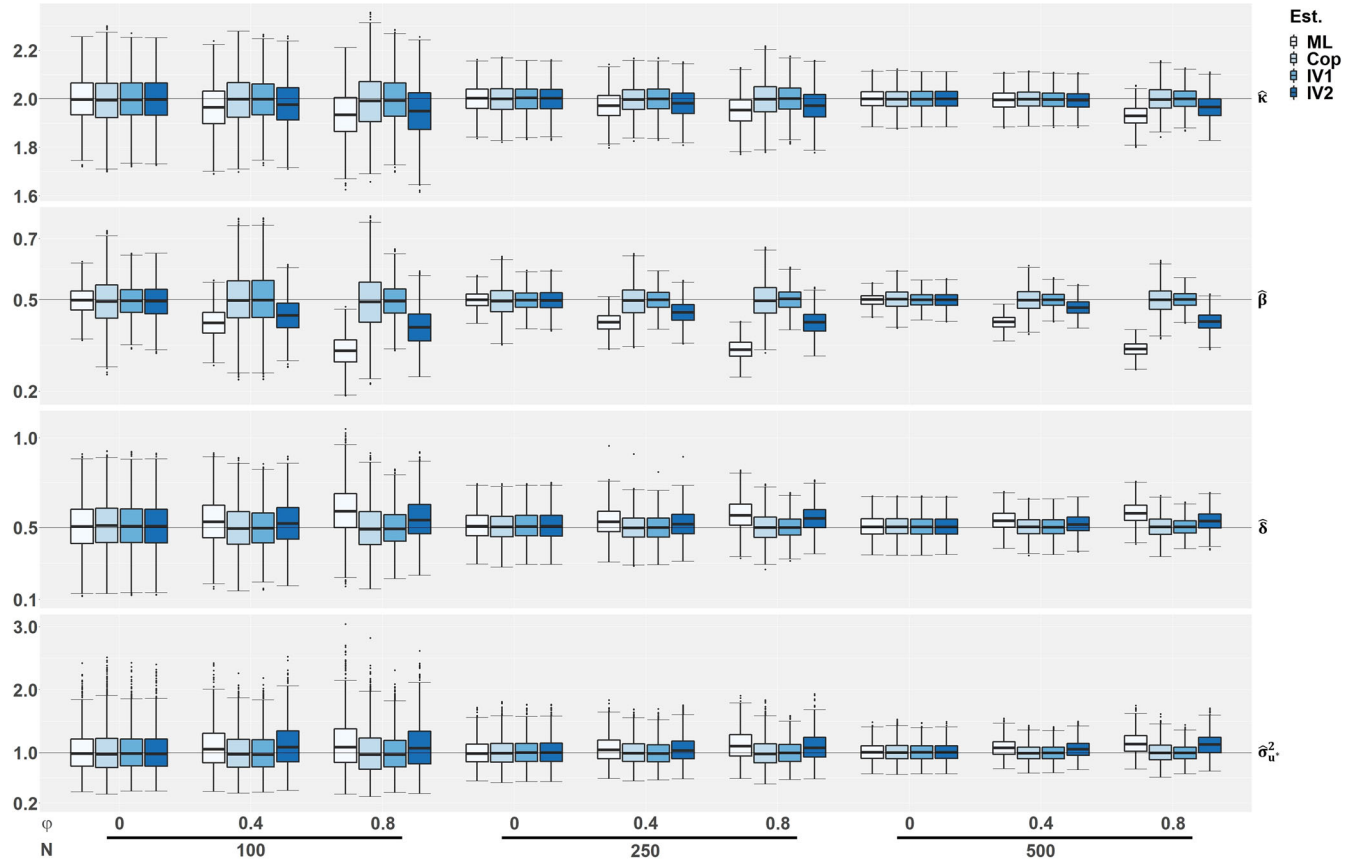

FIGURE 2 Boxplots for parameter estimates $\hat{\kappa}$ (top), $\hat{\beta}$ (second top), $\hat{\delta}$ (second lowest), and $\hat{\sigma}^2_{u^*}$ (lowest) obtained from the estimators *ML*, *Cop*, *IV*1, and *IV*2 based on the DGM of scenario (*i*) in (13). Cop, copula; DGM, data-generating model; IV, instrumental variable; ML, maximum-likelihood.

the endogenous variable z_i, the joint density $f^{**}\left(\eta_i, u_i^*\right)$ is misspecified, and the dependence between η_i and u_i^* is only insufficiently captured. As a result, the estimator $\hat{\theta}_{\mathrm{IV2}}$ might suffer from remaining endogeneity if $\varphi > 0$.[9]

3.3 | Distributional results for parameter estimates

The stylized boxplots displayed in Figure 2 provide an overall assessment of the simulation outcomes.[10] Parameter-specific boxplots are provided with respect to the same scale. Hence, the graphical representations suitably highlight the asymptotic properties of the alternative estimators. All estimators seem to converge to some asymptotic limit. Specifically, unbiased estimators, for example, *ML* under exogeneity, *Cop*, *IV*1 converge to the true underlying model parameters, while biased estimators (*ML* under endogeneity, *IV*2) approach a false limit with shrinking variance as the sample size increases.

The boxplots of parameter estimates $\hat{\kappa}$, $\hat{\delta}$, and $\hat{\sigma}^2_{u^*}$ show quite similar outcomes for all alternative estimators under exogeneity of input variables ($\varphi = 0$). In these DGMs, sizeable performance differentials of alternative estimators emerge for the estimation of the elasticity parameter β only. Unsurprisingly, the ML estimator is the method of choice in such models showing the smallest variations around the true parameter of $\beta = .5$. The boxplots for both variants of ML-IV estimation are close to outcomes for *ML*, while the proposed estimator suffers in relative terms from the largest variation around the true parameter.

For DGMs subjected to endogeneity, simulation outcomes change markedly even in the cases of the smallest sample sizes ($N = 100$). For all parameters and sample sizes and both levels of induced endogeneity ($\varphi = .4, .8$), ML estimation and the IV estimator with weak instrumentation (*IV*2) come with distributions of parameter estimates that lack a centering around the underlying true parameters. In these scenarios, properly implemented ML-IV estimation (*IV*1) is

the best performing throughout. However, the practical implementation of this estimator is always at risk of misspecification. Therefore, it is interesting to see that relative to $IV1$ estimation, the efficiency losses of copula estimation (Cop) are at most moderate and rather small in larger samples ($N \geq 250$). Although this result holds for all parameters, moderate inefficiencies of the proposed estimator are most prevalent for the elasticity parameter β.

The relative performance of alternative estimators drastically changes in case of a false density specification and endogenous input variables—scenario (ii) in (13) and $\varphi > 0$ shown in Figure S1 in the Supporting Information. Under such realistic conditions, the proposed estimator is unique among all rival estimators (ML, $IV1$, and $IV2$) in providing unbiased assessments of all model parameters. Apart from unbiased parameter estimation, the proposed estimator is apparently the method of choice in terms of having the smallest estimation errors on average and for the vast majority of single Monte Carlo replications.

Moreover, with the exception of estimating the elasticity parameter β, the distributional characteristics of copula-based estimates are almost entirely unaffected by introducing endogeneity, that is, changing the correlation parameter from $\varphi = 0$ to .4. This is noteworthy, since the model structure is much less complex under exogeneity in comparison with the case of DGMs subjected to endogeneity. Copula-based estimates of β are slightly more dispersed for DGMs subjected to endogeneity in comparison with DGMs featuring exogenous factor inputs. The DGMs considered feature the "realistic" property that—under endogeneity ($\varphi = .4, .8$)—production inputs and inefficiency are positively related (Alexopoulos, 2011). Hence, it is interesting to highlight the directional patterns of biased parameter estimates. Conditional on the endogenous DGMs considered, estimation biases materialize in the form of underestimating both the frontier parameter β and the intercept κ, while the parameters governing the deviation from the frontier tend to be overestimated ($\delta, \sigma_{u^*}^2$). Accordingly, misinterpretations and incorrect conclusions are likely caused by neglected endogeneity in practical applications of Poisson frontier models.

3.4 | Summary of simulation results

In summary, the graphical illustration of the parameter estimates obtained from alternative estimators under a variety of DGMs allows for a few short-hand conclusions: First and unsurprisingly, ML estimation is the method of choice to estimate Poisson frontier models under exogeneity of input variables. Second, if applied properly, that is, with access to suitable external information and correct distributional assumptions, ML-IV estimation is prime to cope with endogenous input variables. Third, if an analyst feels insecure about available instrument information or might doubt the distributional assumptions, ML-IV estimation is at risk of delivering systematically biased assessments of all model parameters. Fourth, with access to scarce or insecure external information the proposed estimator is the method of choice in Poisson frontier models. The same conclusion applies whenever an analyst feels uncertain about distributional assumptions. Fifth, seeing that copula-based estimation is unique in providing unbiased estimates in many realistic cases, it is worth highlighting that efficiency losses relative to ML and ML-IV estimation are moderate even if the latter estimators do not suffer from misspecification in any direction.

The simulation results discussed in particular for the suggested IV-free approach cover somewhat ideal settings where the model implementations align with core aspects of data generation. In the Supporting Information, we examine estimator performance in scenarios of model misspecification regarding (i) the choice of the copula function, (ii) more complex patterns of endogeneity, (iii) discrete or binomial endogenous regressors, and (iv) small samples coupled with weak dependence. From these additional simulations we can draw three general remarks. First, estimation under model misspecification comes at the cost of larger estimation uncertainty. Especially the estimation of variance parameters is at risk of suffering from model misspecification. However, losses in estimation accuracy rapidly shrink when extending the sample information. Second, when dealing with binomial endogenous regressors, the proposed method remains biased in small samples as a result of a lack of identification. However, the Monte Carlo results hint at an improved estimator performance in extended samples in such cases of limited variation in endogenous regressors. Third, the proposed estimator performs well in small samples in the case of low degrees of dependence between inefficiency and the endogenous variable.

4 | KNOWLEDGE GENERATION IN THE EUROPEAN PHARMACEUTICAL INDUSTRY

In this section, we illustrate the proposed method to handle endogeneity in Poisson frontier models empirically, and estimate a production function for the number of patents awarded to a firm in a given year. The empirical model is set up conditional on our initial considerations summarized in Figure 1. Noticing that a major fraction of costs is incurred in the R&D stage in the pharmaceutical industry, an unbiased assessment of the payoff to R&D expenditures offers valuable insights into productivity levels regarding intraindustry competition.

We provide alternative empirical assessments of patent generation as a manifestation of the knowledge production of firms (Hausman et al., 1984). As these estimates differ regarding their performance under endogeneity, our analysis is informative on the adverse effects of unmodeled (ML under exogeneity) or only partly controlled endogeneity (ML-IV with lagged explanatory variables). Apart from modeling knowledge generation of European pharmaceutical firms in an unbiased manner, further purposes of our comparative analysis include (i) highlighting the adverse effects of endogeneity on common ML-based evaluations, (ii) unraveling the case of potential endogenous interrelations between input selection and productive inefficiency, and (iii) investigating the role of knowledge spillover effects in supporting innovation processes.

4.1 | Data

The empirical analysis conditions on a cross-sectional sample of 137 independent European pharmaceutical firms in 2016. The pharmaceutical industry is one of the most productive and profitable industrial sectors and it is often considered as highly research-intense (Meliciani, 2000; Pammolli et al., 2011). An intense commitment to R&D is decidedly among the main reasons for a firm's long-run success, since high innovative efficiency is supposed to establish sustainable competitive advantages (Chen & Chang, 2010). By restricting our attention to firms within a single industry, we can best focus on the unraveling of endogenous interrelations and their implications for the empirical understanding of knowledge generation (N. Wang & Hagedoorn, 2014). In particular, we avoid difficulties associated with interindustrial or panel analysis as additional endogeneities channeled through unobserved heterogeneity are ruled out.

As formalized by the Poisson stochastic frontier model in (3)–(6), the number of patent grants is used to approximate the outcome of knowledge production.[11] Data on patents are retrieved from the PATSTAT database.[12] Following the related literature (E. C. Wang, 2005; E.C. Wang & Huang, 2005), R&D expenditures (*R&D*) and the number of research employees (*MP*) are considered as input factors in the knowledge production process. The variable *R&D* is obtained from profit and loss reports, measured as total business expenditures on research and development in 2016 prices and exchange rates in units of 1000 Euro. The variable *MP* is based on the number of full-time equivalent research staff (technicians, researchers, engineers, etc.) registered on the company's payroll.[13] Data have been drawn from the Amadeus database of the Bureau van Dijk. Formalizing a Cobb–Douglas production process, both input variables ($\boldsymbol{z}_i$ in Equation 11) enter the model in natural logarithms. Accordingly, the coefficient estimates attached to these endogenous variables provide direct quantifications of the elasticities of knowledge output with respect to R&D spendings and research staff. Pointing to returns to scale in performing innovation investments, total elasticities less than 1 (greater than 1) indicate decreasing (increasing) returns to scale (P. Wang, 1998).

4.2 | Evidence on knowledge production

We treat both input variables *R&D* and *MP* as potentially endogenous and employ the proposed copula estimator in comparison with the standard ML approach and ML-IV estimation. Recall that the copula and the ML-IV estimators differ in terms of informational needs. Whereas the former is IV-free, truly external information as required by the latter is hardly available. Hence, we naïvely instrument the input variables with their own 1-year lagged realizations as in Cincera (1997).[14] We differentiate two empirical models. For the first detection of endogeneity in input selection, we set up Model I without scaling of inefficiencies, that is, we impose the restriction $\boldsymbol{\delta} = 0$ in (5). This model coincides with its counterparts in Fé and Hofler (2013) and P. Wang (1998). Hence, it allows for a straightforward comparison of empirical results with literature findings. Moreover, in some sense it isolates endogeneity effects in a stylized manner.

TABLE 1 Estimation results using the ML, the proposed copula, and the ML-IV estimator

	Model I			Model II		
Parameter	**ML**	**Proposed**	**ML-IV**	**ML**	**Proposed**	**ML-IV**
const	2.216	1.601	2.115	2.216	1.567	2.054
	(.0401)	(.0709)	(.4137)	(.0414)	(.0817)	(.4098)
log(*R&D*)	.2178	.4482	.2556	.2521	.4376	.2902
	(.0351)	(.0998)	(.0899)	(.0358)	(.0928)	(.0861)
log(*MP*)	.1562	.1956	.1504	.1770	.2079	.1669
	(.0131)	(.0684)	(.0503)	(.0179)	(.0722)	(.0488)
uni	–	–	–	−.3948	−.4107	−.4114
				(.0401)	(.0957)	(.1071)
$\sigma^2_{u^*}$	2.664	1.120	2.293	2.556	1.183	2.109
	(.1084)	(.1902)	(.2071)	(.1112)	(.2009)	(.2075)
$\rho_{(\log(R\&D),u^*)}$	–	.5251	–	–	.486	–
		(.0537)			(.0499)	
$\varphi_{(\eta_{\log(R\&D)},u^*)}$	–	–	.1209	–	–	.1153
			(.1084)			(.1157)
$\rho_{(\log(MP),u^*)}$	–	.2916	–	–	.251	–
		(.0592)			(.0536)	
$\varphi_{(\eta_{(\log(MP)},u^*)}$	–	–	.0009	–	–	−.0052
			(.0909)			(.0933)

Note: Standard errors in parentheses are obtained by means of bootstrap procedures with 1000 replications. For ML-IV estimation, the covariates *R&D* and *MP* are instrumented with their respective one-period lagged counterparts. The estimated coefficients of the instruments are omitted to economize on space.
Abbreviations: IV, instrumental variable; ML, maximum-likelihood.

To obtain Model II, we augment Model I with inefficiency determinants as outlined in Section 2.3. Specifically, we consider university spillover effects as external determinants of scaled inefficiencies. Hence, the more general model allows a joint investigation of endogenous interrelations between inefficiencies of knowledge generation and research input variables. In addition, it provides a structural view on inefficiency in response to firm location within the geographic proximity of scientific institutions.

4.3 | The stylized Poisson frontier model

4.3.1 | Endogeneity biases and the independence of patent outcomes

The left-hand side panel of Table 1 reports estimation results for Model I obtained through ML, copula, and ML-IV estimation.[15] As we have observed from the simulation studies, the standard errors of the proposed copula estimator are relatively large in small samples, but sufficiently small to draw meaningful inferential conclusions on model parameters. By comparing ML estimates with those from the copula approach, we can implicitly examine whether the production inputs are endogenous. Supporting the case for endogeneity, the estimated elasticities are substantially different. In detail, the model (significantly) changes the estimated output elasticities from .2178 (.1562) to .4482 (.1956) for *R&D* expenditures (research staff, *MP*). The copula estimator also provides tests of regressor endogeneity for $\log(R\,\&\,D)$ and $\log(MP)$. Wald statistics for testing the null hypotheses of $\rho_{(\log(R\&D),u^*)} = 0$ and $\rho_{(\log(MP),u^*)} = 0$ are $.5251/.\,0537 = 9.778$ ($p < .001$) and $.2916/.\,0592 = 4.925$ ($p < .001$), respectively. Hence, both diagnostics indicate the presence of endogeneity for both regressors with high significance. As already insinuated by Cincera (1997),

instrumenting with lagged explanatory variables is not sufficient to cope with estimation biases. Specifically, the coefficient estimates obtained from ML-IV estimation only slightly differ from their ML counterparts.

The Poisson model outlined in (1) and (2) builds implicitly upon the assumption of *independent* patent outcomes between the firms. Although our sample information does not allow an explicit testing of this assumption it is worth recalling that we employ bootstrap techniques for inferential purposes (see also footnote 8). In particular, we employ a resampling scheme of drawing tuples of firm-specific observations independently with replacement. Although a rigorous proof of bootstrap validity in the present Poisson SFA model is beyond the scope of this paper, we observe that the bootstrap means of parameter estimates are very close to the data-based estimates throughout (not shown). For instance, the R&D elasticity estimate of .4482 is very close to the mean elasticity estimate from 1000 bootstrap samples of .4319. Similarly, the bootstrap means of the employment elasticity (.2102) and the correlation estimates $\rho_{(\log(R\&D),u^*)}$ (.5202) and $\rho_{(\log(MP),u^*)}$ (.3077) are very close to the data-based estimates documented in Table 1 for the copula estimator. From these descriptive results for the core behavioral model parameters we conclude that the assumption of independent patent outcomes between the firms is not overly restrictive for the considered sample. Given considerable evidence for endogenous regressors, we proceed with the more general model with scaling property (Model II) for further investigations.[16]

4.3.2 | R&D effectiveness and patent output

The right-hand side panel of Table 1 documents estimation results for Model II comprising scaled inefficiencies. Model augmentation by means of scaling the inefficiency terms only marginally changes the frontier coefficient estimates for all estimators. As demonstrated in Section 3, neglecting potential endogeneity is associated with an underestimation of output elasticities, which is in line with the results of Cincera (1997), who finds a slight increase in coefficient estimates after weakening endogeneity. Accordingly, an unbiased assessment of production elasticities reveals that factor productivity is actually higher when relaxing the assumption of (strictly) exogenous inputs. Although the production elasticities are underestimated under assumed exogeneity, two important effects that have been uncovered earlier in the Monte Carlo study are visible likewise.

On the one hand, inefficiency variances are substantially higher for both the ML and the ML-IV estimation in comparison with copula-based parameter estimation. Assessments of productive efficiency are usually based on $\mathbb{E}\left[\hat{u}_i^*\right]$, which depends on the estimate of $\sigma_{u^*}^2$.[17] Accordingly, the inefficiency variance can be related to a firm's average distance from full efficiency. The estimated average distances from full efficiency are for the ML and the copula estimator, respectively. Although these values are not directly interpretable, they are useful for model comparison. An overestimation of $\sigma_{u^*}^2$ implies that the firms are falsely considered as more inefficient than is actually the case. For instance, as implied by our model outcomes, pharmaceutical firms have been 35% more efficient on average as implied by standard ML-based model evaluations.[18] Furthermore, a sizable shrink in the estimated intercept is evident when moving from ML or ML-IV estimates to copula-based parameter assessments. Recall that the intercept estimate can be attributed to total factor productivity (Kumbhakar & Lovell, 2003). Accordingly, it can be interpreted to reflect accumulated knowledge, that is, a firm's "knowledge base" (Hausman et al., 1984). Although both knowledge accumulation and innovative activity are significantly responsible for patent generation, an overestimate of the intercept joint with underestimates of output elasticities implies that knowledge accumulated is falsely considered to be more important for patent generation in comparison with current production inputs. As a result, inconclusive implications and misleading suggestions for improving knowledge generation processes are likely induced by endogeneity biases.

The estimates for the correlation coefficients $\rho_{(R\&D,u^*)}$ and $\rho_{(MP,u^*)}$ are both significant and positive, providing additional evidence of correlation between unobserved technical inefficiency and production inputs. Specifically, the correlation coefficients are estimated as $\hat{\rho}_{(R\&D,u^*)} = .486$ and $\hat{\rho}_{(MP,u^*)} = .251$. These statistics directly capture mutual dependencies and provide valuable economic information, as they enable assessing how the firms respond to changes in inefficiency on average. Sizeably positive correlation estimates indicate pronounced input adjustments for implicit changes in the production technology. A positive sign of both correlation estimates is intuitively reasonable. For instance, (adverse) external technological shocks are likely to reduce efficiency and result indirectly in less output. At the same time, the production inputs have to be simultaneously increased to retain the output level. Exemplifying such shocks, one might notice regulations of the European Union that apply to the pharmaceutical industry and reflect

that pharmaceutical research outcomes are related to the population's health (Baines et al., 2018). Consequently, stronger production restrictions or sharper regulations extended to drug approvals might be considered as potential manifestations of technological shocks (Baines et al., 2018).

To complement our assessment of knowledge generation of European pharmaceutical firms in 2016, we compare our results to those found in Fé and Hofler (2013) and P. Wang (1998) for the case of US producers in 1976. Both studies use the same US data and a Poisson model that is similar to our Model I in terms of the covariate information employed. Since Fé and Hofler (2013) neglect (potential) endogeneity, a comparison with our ML results is natural. Our ML estimate of the elasticity of innovative output with respect to R&D spendings (.2521) is substantially smaller. In P. Wang (1998), the associated coefficient ranges from .6716 up to 1.588, whereas Fé and Hofler (2013) estimate 1.207, which is more than four times larger than our estimate. Despite cross-country heterogeneity, a sizeable reduction of R&D effectiveness in the over course of more than 40 years is consistent with literature findings diagnosing a steady decline in pharmaceutical payoffs to R&D (Hashimoto & Haneda, 2008, report a decline of 50% from 1983 to 1992). Such diminishing returns in knowledge generation might be ascribed to intensified competition (Pammolli et al., 2011), or the fact that pharmaceutical innovations have become incrementally spread over a large variety of products (Aghion & Howitt, 1998; Ha & Howitt, 2007).

4.3.3 | Knowledge spillovers

Seeing a steady decline in pharmaceutical R&D effectiveness (Hashimoto & Haneda, 2008; Pammolli et al., 2011), a promising complement to produce one's own innovations by investing in R&D is to absorb external knowledge—that is, "diffusion spillovers" (Nelson, 2009). A growing body of literature has highlighted that the capacity to absorb local external knowledge holds particular relevance for firms in the pharmaceutical industry to remain innovative (Furman et al., 2005; Harhoff, 2005; Runiewicz-Wardyn, 2013).

To approximate such effects, the dummy variable *uni* indicates if a university or public research institution with more than 1000 employees (excluding auxiliaries) is located within a 10-km road distance from a firm (for a similar approach, see Hall et al., 2005). As can be seen in Table 1, all three estimated Poisson frontier models lead to the same conclusion in qualitative terms. A negative and significant effect of *uni* indicates that geographic proximity to a university or public research institution is associated with higher levels of efficiency in knowledge generation. Thus, we can confirm the findings of Siegel et al. (2003), namely that university presence facilitates knowledge spillovers which stimulate innovative activity.

To assess the model implications in an economically meaningful way, it is necessary to jointly evaluate the quantitative results with inefficiency variance.[19] The copula estimator reveals a variance estimate of 1.183 and an estimate of −.4107 for the coefficient of *uni*. Since these estimates jointly determine innovative efficiency, we are able to attribute efficiency differences to local knowledge spillovers originating from universities. In detail, pharmaceutical firms located in geographic proximity to a university are on average 56% more efficient than firms without such a neighbor.[20] Since universities are powerful sources of external knowledge (Szücs, 2018), our results indicate that the capacity to absorb such local spillovers largely enhances innovativeness. From the firm perspective, similar research taking place in private research establishments and universities hosting a pharmaceutical faculty could be the major source of spillovers due to technological transfers and the acquisition of key business skills (Powell, 1998).

5 | CONCLUSION

Previous studies have highlighted five important aspects that should be taken into account when modeling firm-level innovation processes, namely (i) innovation-relevant knowledge generated by the firm is unobserved, (ii) the discrete nature of patents should be taken into account when using them as an approximation of innovative output, (iii) the process turning R&D into knowledge is likely affected by inefficiencies that (iv) are subject environmental factors, and (v) R&D inputs are typically endogenous. We establish an empirical model to simultaneously consider these aspects regarding the pharmaceutical industry. The suggested Poisson frontier model allows for inefficiency scaling, and accounting for regressor endogeneity does not require instrumental information. Rather, building upon the Gaussian SFA model of Tran and Tsionas (2015) feedback linking R&D expenditures and inefficiency is quantified by means of a (Gaussian) copula. We examine the finite sample behavior of the proposed estimator by means of Monte Carlo simulations. The results show that the coefficient estimates of the proposed method are not affected by endogeneity. From a set of complementary simulation experiments, we conclude that

the performance of the proposed estimator remains unaffected in case of departures from some underlying model assumptions, such as non-Gaussian dependence structures.

However, the proposed approach is not without limitations. First, although the Gaussian copula is highly flexible, especially when dealing with multiple endogenous inputs as in an empirical application, it still builds upon the assumption of linear dependency of its margins. Second, Monte Carlo results show that unbiased estimates come at the cost of high variances in small samples. Although instrumental variable estimation is the method of choice under endogeneity, weak instrumentation or density misspecifications are pertinent threats to such fully parametric estimation. Nevertheless, we expect that the proposed method provides useful alternatives to many empirical problems addressed in the form of Poisson frontier models.

We employ the proposed estimator to quantify a knowledge production function for European pharmaceutical firms in 2016. In particular, innovation processes are subject to technological interactions between innovative efficiency and R&D. On the one hand, we can confirm the arguments of Alexopoulos (2011) that stylized technological shocks are positively linked to the inputs of knowledge production. On the other hand, endogeneity biases are pertinent, since restrictive assumptions of regressor exogeneity result in a marked underestimation of an R&D elasticity and efficiency. In addition to the assessment of endogenous input selection, we re-examine the effect of university knowledge spillovers on innovative efficiency. Our results support the "absorptive capacity" hypothesis (Cohen & Levinthal, 1989, 1990). Since pharmaceutical research is costly, innovation policy should explicitly develop the capacity to absorb local technology spillovers. In particular, universities hosting a pharmaceutical faculty stimulate beneficial externalities that foster innovativeness, which can be exploited for efficiency enhancements. Accordingly, a deeper understanding of the channels of external pharmaceutical knowledge acquisition would benefit the strategic positioning of firms to become more innovative.

ACKNOWLEDGMENTS

We gratefully acknowledge helpful comments and suggestions received from three anonymous reviewers and the editor Daniel Spulber. Financial support by the German Research Association (DFG) Research Training Group 1644 "Scaling problems in Statistics" (grant no. 152112243) is also gratefully acknowledged. Much of the research documented in this paper was conducted while the corresponding author was employed at the Chair of Statistics and Econometrics at the Georg-August University of Göttingen, Germany. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Rouven E. Haschka http://orcid.org/0000-0002-2916-9745

ENDNOTES

[1] We discuss the important origins of endogeneity in knowledge generation models below in Section 2.2. The linkage of endogenous covariate information and stochastic inefficiency can lead to biased estimates for causal effects if applied methods build upon assumptions of regressor exogeneity. As a result, the outcome of managerial decisions can hardly be assessed under the presumption of independent actions combined with the invariance of all other factors.

[2] To be precise, total factor productivity is related to $\exp(\kappa)$.

[3] Although scaling functions are widely employed in Gaussian SFA models (Kumbhakar & Lovell, 2003), they have not yet been used for the class of count data models.

[4] In contrast to Fé and Hofler (2013, 2020), who approximate the integral using Halton sequences, we follow Tran and Tsionas (2015) and use Gauss–Konrod quadrature with 50 points.

[5] We assess adverse effects on the performance of the estimator when dealing with binary endogenous regressors in the Supporting Information.

[6] We characterize the performance of the proposed estimator under misspecification of the copula function in the Supporting Information.

[7] The rescaling factor $1/(N+1)$ instead of $1/N$ ensures that the empirical cumulative distribution is well bounded in $(0, 1)$.

[8] We use the derivative-free simplex method for numerical optimization of Nelder and Mead (1965) to obtain the ML estimates of model parameters θ. The proposed copula estimator can be classified as a semiparametric method, as it combines the nonparametric distribution of endogenous regressors with the parametric distribution of the inefficiency term. In semiparametric models, the derivation of Hessian-based asymptotic standard errors can be extremely difficult such that bootstrap procedures have become a natural alternative to assess estimation uncertainty (Park & Gupta, 2012). Accordingly, we use simple bootstrap techniques to estimate standard errors and

confidence intervals. For fully nonparametric approaches in SFA models for continuous outcomes, we refer the reader to Simar et al. (2016) and Mastromarco and Simar (2021).

[9] The correlation between the endogenous variable z_i and the instrument w_i is on average about .705. Accordingly, w_i can be considered as a strong instrument for z_i. The "weak" instrument w_i^* exhibits an average correlation, which is smaller by a factor of .6715.

[10] The coefficient attached to the exogenous variable x (α) did not exhibit a bias (see Tables S1 and S2 in the Supporting Information). For this reason, boxplots for $\hat{\alpha}$ are not included, although they can be provided by the authors upon request.

[11] Mansfield (1982, 1986) showed that in the pharmaceutical industry the patenting rate of inventions is about 84%. More recently, Danguy et al. (2019) document patenting rates between 79% and 74%. Noticing that not all knowledge output is patented, patents are only an imperfect indicator to approximate knowledge production. Nonetheless, they hold particular relevance in the pharmaceutical industry, since the patenting rate in this sector is among the highest in any high-technology sectors (Danguy et al., 2019; Griliches, 1998).

[12] We restrict our analysis to European patents, since the PATSTAT database only provides information about patents granted by the European Patent Office.

[13] External agency workers were disregarded due to data limitations.

[14] Input variables lagged by 1 year are available for only 94 firms out of the full sample of 137 firms. Accordingly, ML-IV estimates are based on the reduced sample.

[15] The copula estimator is set up with a three-dimensional Gaussian copula to model the interdependence between *R&D*, *MP*, and the (unobserved) inefficiency term, which is assumed to be half normally distributed. Regarding the ML-IV estimator, we additionally assume that the endogenous inputs *R&D* and *MP* are marginally normal.

[16] As an alternative to considering Poisson distributed patent outcomes, discrete counts could also be assumed to follow a more flexible negative binomial distribution. Explicit estimation results for the negative binomial model are available upon request from the authors. Except for intercept estimates, ± 2 standard errors rule-of-thumb confidence intervals constructed around copula-based estimates retrieved under the negative binomial model include the estimates as documented for the Poisson models I and II in Table 1.

[17] Specifically, it holds that $\mathbb{E}\left[u_i^*\right] = \sigma_{u^*}\frac{\phi(0)}{\Phi(0)}$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function of the standard normal distribution, respectively.

[18] By taking the ratio, that is, $1 - .8444/1.302$.

[19] Since $u_i^* \sim N^+\left(0, \sigma_{u^*}^2\right)$, then $u_i = \exp(\delta s_i)u_i^*$ is $\sim N^+\left(0, \exp(2\delta s_i)\sigma_{u^*}^2\right)$, see Haschka et al. (2020).

[20] By taking the ratio, that is, $1 - \exp(2 \times (-.4107) \times 1)/\exp(2 \times (-.4107) \times 0)$. The remaining two estimators provide nearly identical results. In detail, they report efficiency gains of 54.6% (ML) and 56.1% (ML-IV).

REFERENCES

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, *37*, 3–13.

Aghion, P., & Howitt, P. (1998). *Endogenous growth theory*. MIT Press.

Alexopoulos, M. (2011). Read all about it!! What happens following a technology shock? *American Economic Review*, *101*(4), 1144–1179.

Amsler, C., Prokhorov, A., & Schmidt, P. (2016). Endogeneity in stochastic frontier models. *Journal of Econometrics*, *190*(2), 280–288.

Arora, A., Ceccagnoli, M., & Cohen, W. M. (2003). *R&D and the patent premium* [NBER Working Paper No. 9431].

Baines, D., Bates, I., Bader, L., Hale, C., & Schneider, P. (2018). Conceptualising production, productivity and technology in pharmacy practice: A novel framework for policy, education and research. *Human Resources for Health*, *16*(1), 51.

Becker, J.-M., Proksch, D., & Ringle, C. M. (2022). Revisiting Gaussian copulas to handle endogenous regressors. *Journal of the Academy of Marketing Science*, *50*, 46–66. (forthcoming).

Blume, L., & Fromm, O. (2000). *Regionalökonomische Bedeutung von Hochschulen: Eine empirische Untersuchung am Beispiel der Universität Gesamthochschule Kassel*. Deutscher Universitätsverlag.

Chen, Y.-S., & Chang, K.-C. (2010). The relationship between a firm's patent quality and its market value—The case of US pharmaceutical industry. *Technological Forecasting and Social Change*, *77*(1), 20–33.

Cincera, M. (1997). Patents, R&D, and technological spillovers at the firm level: Some evidence from econometric count models for panel data. *Journal of Applied Econometrics*, *12*(3), 265–280.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*. Springer.

Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: The two faces of R&D. *The Economic Journal*, *99*(397), 569–596.

Cohen, W. M., & Levinthal, D. A. (1990). The implications of spillovers for R&D investment and welfare: A new perspective. *Administrative Science Quarterly*, *35*(1990), 128–152.

Correa, C. M. (2004). Ownership of knowledge: The role of patents in pharmaceutical R&D. *Bulletin of the World Health Organization*, *82*, 784–787.

Czarnitzki, D., & Toole, A. A. (2011). Patent protection, market uncertainty, and R&D investment. *The Review of Economics and Statistics*, *93*(1), 147–159.

Danaher, P. J., & Smith, M. S. (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science*, *30*(1), 4–21.

Danguy, J., De Rassenfosse, G., & van Pottelsberghe de la Potterie, B. (2019). The R&D-patent relationship: An industry perspective. *European Investment Bank (EIB), Luxembourg*, *14*(1), 170–195.

Datta, H., Ailawadi, K. L., & Van Heerde, H. J. (2017). How well does consumer-based brand equity align with sales-based brand equity and marketing-mix response? *Journal of Marketing*, *81*(3), 1–20.

Daub, C.-H. (2008). Nachhaltige Unternehmen unter Innovationsdruck. *Marketing Review St. Gallen*, *25*(4), 18–22.

Dittmer, S., & Strätz, E. (2012). Gewusst wie - General Management in wissenschaftsinstitutionen: Führungskräfte zwischen Autonomie und Innovationsdruck. *Wissenschaftsmanagement*, *4*, 22–27.

Drivas, K., Economidou, C., & Tsionas, E. G. (2014). *A Poisson stochastic frontier model with finite mixture structure* [MPRA Working Paper, No. 57485, 1–18]. https://mpra.ub.uni-muenchen.de/57485/

Fahrmeir, L., & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society*, *50*(2), 201–220.

Fé, E., & Hofler, R. (2013). Count data stochastic frontier models, with an application to the patents–R&D relationship. *Journal of Productivity Analysis*, *39*, 271–284.

Fé, E., & Hofler, R. (2020). sfcount: Command for count-data stochastic frontiers and underreported and overreported counts. *The Stata Journal*, *20*(3), 532–547.

Fu, X., & Yang, Q. G. (2009). Exploring the cross-country gap in patenting: A stochastic frontier approach. *Research Policy*, *38*(7), 1203–1213.

Furman, J. L., Kyle, M. K., Cockburn, A. M., & Henderson, R. (2005). Public & private spillovers: Location and the productivity of pharmaceutical research. *Annals of Economics and Statistics*, *79*, 165–188.

Griliches, Z. (1998). Patent statistics as economic indicators: A survey. In Z. Griliches (Ed.), *R&D and productivity: The econometric evidence* (pp. 287–343). University of Chicago Press.

Ha, J., & Howitt, P. (2007). Accounting for trends in productivity and R&D: A Schumpeterian critique of semi-endogenous growth theory. *Journal of Money, Credit and Banking*, *39*(4), 733–774.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, *36*(1), 16–38.

Harhoff, D. (2005). R&D spillovers, technological proximity, and productivity growth—Evidence from German panel data. *Schmalenbach Business Review*, *52*(3), 238–260.

Haschka, R. E. (2022). Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research*, 1–22.

Haschka, R. E., & Herwartz, H. (2020). Innovation efficiency in European high-tech industries: Evidence from a Bayesian stochastic frontier approach. *Research Policy*, *49*, 104054.

Haschka, R. E., Herwartz, H., Struthmann, P., Tran, V. T., & Walle, Y. M. (2021). The joint effects of financial development and the business environment on firm growth: Evidence from Vietnam. *Journal of Comparative Economics*, *50*(2), 486–506. (forthcoming).

Haschka, R. E., Schley, K., & Herwartz, H. (2020). Provision of health care services and regional diversity in Germany: Insights from a Bayesian health frontier analysis with spatial dependencies. *The European Journal of Health Economics*, *21*(1), 55–71.

Hashimoto, A., & Haneda, S. (2008). Measuring the change in R&D efficiency of the Japanese pharmaceutical industry. *Research Policy*, *37*(10), 1829–1836.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents–R&D relationship. *Econometrica*, *52*, 909–937.

Holgersson, M. (2013). Patent management in entrepreneurial SMEs: A literature review and an empirical study of innovation appropriation, patent propensity, and motives. *R&D Management*, *43*(1), 21–36.

Jaffe, A. B. (1986). *Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value* [NBER Working Paper No. 1815].

Jaffe, A. B. (1989). Characterizing the 'technological position' of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, *18*(2), 87–97.

Kumbhakar, S., & Lovell, C. A. K. (2003). *Stochastic frontier analysis*. Cambridge University Press.

Kutlu, L. (2010). Battese–Coelli estimator with endogenous regressors. *Economics Letters*, *109*(2), 79–81.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.

Manchanda, P., Rossi, P. E., & Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, *41*(4), 467–478.

Mansfield, E. (1982). How economists see R&D. *Research Management*, *25*(4), 23–29.

Mansfield, E. (1986). Patents and innovation: An empirical study. *Management Science*, *32*(2), 173–181.

Mastromarco, C., & Simar, L. (2021). Latent heterogeneity to evaluate the effect of human capital on world technology frontier. *Journal of Productivity Analysis*, *55*(2), 71–89.

Meliciani, V. (2000). The relationship between R&D, investment and patents: A panel data analysis. *Applied Economics*, *32*(11), 1429–1437.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Nelsen, R. B. (2006). An introduction to copulas. In *Springer series in statistics*. Springer.

Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005.

Oshri, I., Kotlarsky, J., & Willcocks, L. P. (2015). *The handbook of global outsourcing and offshoring*. Palgrave MacMillan.

Pakes, A., & Griliches, Z. (1984). Patents and R&D at the firm level: A first look. In Z. Griliches (Ed.), *R&D, patents, and productivity* (pp. 55–72). University of Chicago Press.

Pammolli, F., Magazzini, L., & Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*, *10*(6), 428.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, *31*(4), 567–586.

Pieri, F., Vecchi, M., & Venturini, F. (2018). Modelling the joint impact of R&D and ICT on productivity: A frontier analysis approach. *Research Policy*, *47*(9), 1842–1852.

Powell, W. W. (1998). Learning from collaboration: Knowledge and networks in the biotechnology and pharmaceutical industries. *California Management Review*, *40*(3), 228–240.

Prokhorov, A., Tran, K. C., & Tsionas, M. G. (2020). Estimation of semi-and nonparametric stochastic frontier models with endogenous regressors. *Empirical Economics*, 1–26.

Runiewicz-Wardyn, M. (2013). *Localized knowledge spillovers, agglomeration externalities, and technological dynamics in high-tech industries. Evidence based on the EU regions*. Springer.

Scherer, F. M. (2001). The link between gross profitability and pharmaceutical R&D spending. *Health Affairs*, *20*(5), 216–220.

Schilling, M. A. (2015). Technology shocks, technological collaboration, and innovation outcomes. *Organization Science*, *26*(3), 668–686.

Siebert, R. B. (2017). A structural model on the impact of prediscovery licensing and research joint ventures on innovation and product market efficiency. *International Journal of Industrial Organization*, *54*, 89–124.

Siegel, D. S., Westhead, P., & Wright, M. (2003). Science parks and the performance of new technology-based firms: A review of recent UK evidence and an agenda for future research. *Small Business Economics*, *20*(2), 177–184.

Simar, L., Vanhems, A., & Van Keilegom, I. (2016). Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics*, *190*(2), 360–373.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'institut de Statistique de L'Universit de Paris*, *8*, 229–231.

Szücs, F. (2018). Research subsidies, industry-university cooperation and innovation. *Research Policy*, *47*(7), 1256–1266.

Tran, K. C., & Tsionas, E. G. (2013). GMM estimation of stochastic frontier models with endogenous regressors. *Economics Letters*, *118*, 233–236.

Tran, K. C., & Tsionas, E. G. (2015). Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters*, *133*, 85–88.

Tran, K. C., & Tsionas, M. G. (2021). Efficient semiparametric copula estimation of regression models with endogeneity. *Econometric Reviews*, 1–20. (forthcoming).

Wang, E. C. (2005). R&D efficiency and economic performance: A cross-country analysis using the stochastic frontier approach. *Journal of Policy Modeling*, *29*(2), 345–360.

Wang, E. C., & Huang, W. (2005). Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach. *Research Policy*, *36*(2), 260–273.

Wang, H.-J., & Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis*, *18*(2), 129–144.

Wang, N., & Hagedoorn, J. (2014). The lag structure of the relationship between patenting and internal r&d revisited. *Research Policy*, *43*(8), 1275–1285.

Wang, P., Cockburn, l. M., & Puterman, M. L. (1998). Analysis of patent data—A mixed-Poisson-regression-model approach. *Journal of Business & Economic Statistics*, *16*(1), 27–41.

Windmeijer, F. A., & SantosSilva, J. M. (1997). Endogeneity in count data models: An application to demand for health care. *Journal of Applied Econometrics*, *12*(3), 281–294.

Yu-ming, W. (2009). An empirical analysis of R&D cooperation and regional knowledge spillovers based on knowledge production function. *Studies in Science of Science*, *27*(11), 1486–1494.

Zhang, X., Kumar, V., & Cosguner, K. (2017). Dynamically managing a profitable email marketing program. *Journal of Marketing Research*, *54*(6), 851–866.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Haschka, R. E., & Herwartz, H. (2022). Endogeneity in pharmaceutical knowledge generation: An instrument-free copula approach for Poisson frontier models. *Journal of Economics & Management Strategy*, 31, 942–960. https://doi.org/10.1111/jems.12491