Selection and Incentive Effects of Gatekeeping on Healthcare Utilisation in Germany

Masterarbeit zur Erlangung des akademischen Grades Master of Science – Volkswirtschaftslehre, Universität Leipzig

Christopher Schrey

Eingereicht im Juli 2019

Abstract

The aim of this piece of research was to assess whether or not gatekeeping participation is exogenous with respect to healthcare utilisation. To do so, an endogenous treatment regression model was be employed, which required maximum simulated likelihood estimation. Within this model, a latent factor structure is applied. This procedure allows to separate the effect of selection on unobservables from the causal incentive effect. Results imply that unobservable characteristics are conducive to both healthcare utilisation and gatekeeping participation. Even more so, selection effects outweigh the incentive effects. Overall, gatekeeping participation causally strengthens the role of the general practitioner, decreases emergency admissions, while decreasing both ambulatory and total medical costs. A likelihood-ratio test supports the rejection of exogeneity in three out of four utilisation measures. A literature review to make an informed decision about the required number of simulation draws was conducted. Also, extensive robustness analysis suggests that simulation error did not bias the results.

Keywords

gatekeeping participation · healthcare utilization · maximum simulated likelihood · incentive effects · selection effects

Contents

List of Tables

List of Figures

1 Introduction

Endogeneity arising through self-selection into treatment is a problem very often encountered in, but not limited to, health economics. One prominent situation in which health economists face problems with endogenous regressors is when the effect of health insurance status on healthcare utilisation, such as number of visits to the doctor, is to be estimated. Economic theory suggests that the overall effect of being health insured on healthcare utilisation can be decomposed into incentive and selection effects (Akerlof 1970; Arrow 1963). While the incentive effect refers to the restrictions and incentives imposed by the specific insurance plan with respect to healthcare utilisation, selection effects refer to the specific risk-structure of those attracted by the insurance plan. When participation in such health insurance plans is non-random, studies of incentive effects may be biased by self-selection on unobservables (Shane et al. 2012).

Self-selection on unobservables may occur when individuals with private information, such as risk preferences or awareness of future health states, select plans according to their private knowledge (Deb et al. 2006c). The same unobservable characteristics that affect insurance choice might also affect future healthcare utilisation. This is to say, individuals make insurance choice decisions with future healthcare needs in mind, whereas healthcare decisions are similarly based on healthcare needs, conditional on health insurance choice (Deb et al. 2006a).

Within the German statutory health insurance (SHI) system, health insurance schemes (*Krankenkassen*) offer different types of health insurance plans (*Wahltarife*), into which the insured may voluntarily enrol. Within these health insurance plans, the gatekeeping programme (*Hausarztzentrierte Versorgung*) is meant to incentivise economical utilization of scarce medical services, by making the general practitioner (*Hausarzt, GP*) a so called *gatekeeper*, by limiting otherwise free access to medical specialists. Recent studies dealing with the incentive effect of gatekeeping in the German SHI system are Hofmann et al. (2016), Klora et al. (2017), Schneider et al. (2016), Szecsenyi et al. (2018), and Wensing et al. (2017). While Szecsenyi et al. (2018) and Wensing et al. (2017) find that gatekeeping participation increases the number of GP contacts, Klora et al. (2017) find that GP consultations decrease, instead. While consensus seems to exist that gatekeeping participation decreases the number of emergency admissions, results concerning ambulatory and stationary costs are conflicting, as well. Hofmann et al. (2016) find that gatekeeping increases ambulatory cost, while Klora et al. (2017), Schneider et al. (2016) and Szecsenyi et al. (2018) find that ambulatory cost decrease as a result of gatekeeping participation. Similarly, with other healthcare utilisation and cost measures, the authors provide conflicting evidence with respect to effects of the gatekeeping programme in Germany. As participation in the gatekeeping programme is voluntary, thus non-random, selection on unobservables may possibly have biased their findings.

While selection on observables was only considered by Hofmann et al., none of the discussed pieces of research even considered selection on unobservables to bias their results. Accordingly, their conflicting estimates of the incentive effect of gatekeeping on healthcare utilisation might potentially be due to bias introduced by self-selection on unobservables. This raises the question, whether or not gatekeeping participation is exogenous with respect to healthcare utilisation. Accordingly, the null-hypothesis, that within the gatekeeping programme there is no effect of selection on unobservables on healthcare utilisation, emerges.

To deal with the underlying endogeneity issue, the endogenous treatment regression model, by Deb et al. (2006c), will be employed. It makes use of a latent factor structure that combines participation

and outcome equations to account for selection on unobservables. These latent factors enter both participation and healthcare utilisation equation to allow for simultaneous influences of insurance plan choice to affect healthcare utilisation, thus isolating the effect of selection on unobservables (Deb et al. 2006c). The latent factors serve as proxies for unobservable characteristics and can be interpreted as unobserved heterogeneity. Endogeneity is controlled for, as the same latent factors that affect gatekeeping participation also affect the healthcare utilisation decision. Problems in maximum likelihood estimation arise, as the latent factors cannot be observed, so that no closed-form solution to the respective integral exists (Deb et al. 2006c). Yet, when assuming that the underlying unobservable characteristics are standard normally distributed, maximum simulation likelihood (MSL) estimation remains an option. Thus, the effect of the unobservables can be integrated out, resulting in an unbiased estimate of the incentive effect. SHI claims data from the research database of the WIG2 institute will serve as basis for the analysis.[1]

[1] For further information on the research database of the WIG2 institute see: https://www.wig2.de/analysetools/wig2-forschungsdatenbank.html (received 04/15/2020).

2 Background

2.1 Incentive and Selection Effects

Economic theory suggests that the overall effect of being health insured on healthcare utilisation can be decomposed into two effects. First, being insured may have an impact on the insurant's incentives regarding healthcare utilisation. Arrow (1963, p. 962) suggests that health insurance "[...] *removes the incentive* [...] *to shop around for better prices for hospitalization and surgical care*". The case when the insurant's incentive to look for cost-saving healthcare alternatives is limited, due to being insured, is known as ex-post moral hazard. Within the German SHI system, freedom of provider choice may induce *ex-post moral hazard*, as being fully insured one may wish to consult a rather expensive specialist, even when consultation of a rather cheap GP might be sufficient. If the insurant had to partially cover the expenses, incentives to look for cost-saving alternatives might increase. Thus, health insurance schemes offer different types of health insurance plans to incentivise economical utilisation of scarce medical resources. The effect health insurance plans unfold on an individual's incentive towards healthcare utilisation will be referred to as *incentive effect*.[2]

Secondly, the specific risk-structures of individuals attracted by such health insurance plans, given that participation is voluntary and thus non-random, will have an effect on healthcare utilisation as well. Akerlof (1970) proposes that adverse (advantageous) selection may occur, whenever ill (healthy) individuals self-select into certain health insurance plans, resulting in higher (lower) costs for the insurance company. That is to say, the same factors that will lead to higher healthcare utilisation will also make enrolment into specific health insurance plans more likely. While some risk-related characteristics can be observed (e.g. age, gender, etc.), other factors that are conducive to both healthcare utilisation and enrolment remain unobserved. The same unobservable characteristics that affect insurance choice may also affect future healthcare utilisation, thus leading to potential unobserved correlation between insurance decision and decision to consume health services (Shane et al. 2012). These unobservable characteristics may potentially bias analysis of the incentive effect of health insurance plans on healthcare utilisation. Consequently, differences in healthcare utilisation among individuals with different insurance plans that is due to self-selecting into insurance plans according to risk-types (healthy vs. ill), will be referred to as *selection effect*.

2.2 Gatekeeping in Germany

hjg Within the German SHI system, one such health insurance plan is the gatekeeping programme. Within the gatekeeping programme, the insurant needs to make a commitment to choose one specific GP. With this commitment, the insurant may only consult the chosen GP, and may consult specialists only with referral of the GP, thus making the GP a gatekeeper. The gatekeeping GP will coordinate the patient's treatment with other physicians and will promote an exchange of information among the involved healthcare providers (Mehl et al. 2015). The benefit to the insurant lies in the better coordination of treatment among different specialists, managed by the gatekeeping GP. At the same time, from the SHI scheme's perspective, potentially unnecessary and expensive consultations of specialists (due to ex-post moral hazard) can be avoided, as one requires referral of the gatekeeping GP. Consequently, the gatekeeping programme is meant to substitute potentially unnecessary specialist consultations in favour of visits to the GP, thus ceteris paribus increasing the number of GP consultations. By

[2] The term *incentive effect* could be used synonymously with the expression treatment effect throughout this piece of research.

disincentivizing unnecessary specialist consultations, the overall ambulatory costs of the individual are supposed to decrease, while total cost-savings imply that costs should not be shifted from the ambulatory to the stationary sector (Mehl et al. 2015).

2.3 Variable Summary

The research database of the WIG2 Institute contains anonymized SHI claims data (*Routinedaten*) from 2010–2018.[3] Over this period, the database includes information about the year of birth, gender and insurance periods for about 4.5 million insured persons, as well as information about the billed services from the following areas:

- doctors
- pharmacies
- hospitals
- inability to work
- remedies and aids
- home nursing and
- travel expenses

as well as the costs of the services billed. Consequently, the explanatory and outcome variables need to be chosen deliberately from a multitude of healthcare utilisation and health insurancerelated records. The set of socio-demographic variables to choose from SHI claims data is rather limited, in contrast to, e.g. healthcare related surveys. Typical proxies for healthcare needs are age, gender, income and morbidity variables, such as presence of chronic conditions (O'Donnell et al. 2007). As the effect of age on healthcare utilization might not be linear, the age term will additionally enter the healthcare utilisation equation in a quadratic form (as in, e.g. Cameron et al. (2008, 1988)). Individuals younger than 20 years of age will be excluded, as their healthcare utilisation behaviour might be strongly and unobservably influenced by their parents. Income can be considered to be a determinant of healthcare utilisation as well, as individuals with lower income typically have greater healthcare needs (O'Donnell et al. 2007). An individual's morbidity will be approximated by the sum out of four chronic conditions (as in Cameron et al. (1988)): obesity, coronary artery disease, chronic obstructive pulmonary disease and chronic pain. Regional characteristics are known to influence healthcare utilisation as well, as access to healthcare is not evenly distributed among regions (Weinhold et al. 2014, 2018). To account for differences among rural and urban populations, population density per zip-code area will be considered as explanatory variable, to control for possibly supply induced demand.[4] To avoid further endogeneity, both income and the number of chronic diseases refer to the year (2016) prior to the year in which healthcare utilisation is measured (2017). Also, to avoid heteroskedasticity, both income and population density are logarithmised (natural logarithm, denoted as ln). The explanatory variables extracted and employed summarised in **Tab. 1**.

[3] Claims data are not publicly available, as they contain sensitive patient records. The employed sample is available from the author at request.

[4] Zip-code areas and population density are derived online from OpenStreetMap (2019).

***Tab. 1**: Summary Statistics of Gatekeeping Participants vs. Non-Participants*

	Participants		Non-participants	
	Mean	Std. Dev.	Mean	Std. Dev.
Explanatory Variables				
Age	46.76	12.07	43.56	12.12
Age^2	2332.31	1057.87	2043.93	1040.38
Sex (male)	0.55	0.50	0.56	0.50
ln (income)	10.28	1.18	10.27	1.19
ln (pop. density)	6.76	0.83	6.74	0.88
Number chronic. cond.	0.17	0.43	0.10	0.32
Healthcare outcomes				
GP	2.23	2.9	1.38	2.07
Emergency	0.06	0.29	0.05	0.26
ln (ambulatory)	5.98	1.00	5.19	1.89
ln (total)	4.60	3.58	3.75	2.83
Weights				
Days insured in 2017	364.61	3.78	364.10	10.41
Observations	2281		42719	

Number of visits to the GP, number of emergency admissions, ambulatory and total costs of healthcare will serve as healthcare utilisation measures. Gatekeeping is meant to substitute possibly avoidable and rather expensive specialist consultations with rather cheap GP consultations. That is to say, the incentives provided by the gatekeeping programme are meant to increase, all else being equal, visits to the GP. Therefore, the main healthcare utilisation outcome of interest will be an individual's number of visits to a GP within one year. As the gatekeeping programme limits otherwise free access to specialists, it might be possible, that individuals substitute specialist consultations with emergency admissions at the hospital, given that one does not need referral for an emergency admission (Deb et al. 2006c). As the overall aim of the gatekeeping programme is to reduce ex-post moral hazard-induced costs, ambulatory costs (i.e. costs of visits to all types of physicians) of an individual is another outcome of interest. Also, the gatekeeping programme is not only meant to reduce ambulatory, but overall healthcare costs. Thus, the costs of drugs, sick-leave and costs of hospital treatment will be analysed as well and will be referred to as *total costs*. Tab. 1 also provides an overview of the healthcare outcomes. Within all subsequent econometric analysis, individuals will be weighted by the amount of days in which they were insured (equivalent to days in which they could be observed) in 2017, which can be seen in Tab. 1. Out of the 45,000 randomly drawn individuals 2,281 (5.0 %) were participating in the gatekeeping programme.

3 Econometric Methods

Before the endogenous treatment regression approach of Deb et al. (2006c), that jointly estimates healthcare utilisation and gatekeeping participation decision will be introduced, the latter two equations need to be discussed. Also, the Negative Binomial regression model, which will be employed to properly address the two count data outcomes *visits to the GP* and *emergency admissions* deserves some attention. Afterwards, the maximum simulated likelihood approach, that is necessary for estimation, will be introduced.

3.1 Healthcare Utilisation and Gatekeeping Decision

3.1.1 Healthcare Utilisation

To properly estimate the incentive effect of gatekeeping on healthcare utilisation, other determinants of healthcare seeking behaviour need to properly covered. Thus, healthcare utilisation will be determined by different sets of variables, as already discussed in section 2.3. Apart from the gatekeeping incentive effect γ where gatekeeping participation will be captured with the dummy variable d_i^{gk}, unobservable characteristics I_i with respective factor loadings λ are believed to influence healthcare utilisation as well. Additionally, an individual's age, gender, labour income, population density within the zip-code area of residence as well as the number of chronic conditions will be assumed to determine healthcare seeking behaviour. These observable characteristics will be denoted as x_i variables. Thus, healthcare utilisation behaviour, such as the number of doctor visits y_i, is modelled through density function *f*, such that (Deb et al. 2006b, p. 311):

$$f(y_i|\ \mathbf{x}_i, \mathrm{d}_i^{gk}, \mathrm{I}_i) = f(\mathbf{x}_i\beta + \mathrm{d}_i^{gk}\gamma + \mathrm{I}_i\lambda) \tag{3.1}$$

3.1.2 Gatekeeping decision

The gatekeeping participation outcome will be modelled with the dummy variable d_i^{gk}, with values 0 for non-participants and 1 for participants, respectively. The equation will be modelled via logistic regression.

The gatekeeping decision is assumed to be made with respect to future healthcare needs. Thus, gatekeeping decision and healthcare utilisation behaviour are both determined by the above mentioned (observable) x_i variables. Similarly, unobservable characteristics I_i with respective factor loadings λ are assumed to jointly determine both healthcare utilisation and gatekeeping decision. Gatekeeping participation decision is modelled through a density function *g* that characterizes binary choice (Deb et al. 2006b, p. 310):

$$\Pr[\mathrm{d}_i^{gk}|\ \mathbf{x}_i, \mathrm{I}_i] = g(\mathbf{x}_i\phi + \mathrm{I}_i\lambda). \tag{3.2}$$

Here, endogeneity occurs, as the unobservable characteristics I_i enter both healthcare utilisation and participation equation, while the participation equation also enters the utilisation equation separately.

3.2 Negative Binomial Model

The number of visits to the doctor are characterised by being non-negative, integervalued and right-skewed, making doctor visits a textbook example for count data (e.g. Cameron et al. 2005, 2008, 2010).

Appropriate count data models are, among others, the Poisson and Negative Binomial (NB) models. Within the microeconometric literature, the Poisson model is usually used as a starting point (e.g. in Cameron et al. 2005, 2008; Greene 2008), even though not the model of choice, before the Negative Binomial model is introduced.

The Negative Binomial model can be seen as a generalisation of the Poisson case. The Poisson regression model is derived from the Poisson distribution (Cameron et al. 2008, p. 668). The number of visits to the doctor, y_i, given x_i, d_i^{gk} and I_i is Poisson-distributed with density (Cameron et al. 2005, p. 668)

$$f(y_i \mid \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \text{ with } y_i = 0, 1, 2, \ldots \tag{3.3}$$

with intensity parameter μ_i,

$$\mu_i = \exp(\mathbf{x}_i\beta + \mathrm{d}_i^{gk}\gamma + \mathrm{I}_i\lambda). \tag{3.4}$$

A distinctive feature of the Poisson regression model is a property known as *equidispersion*. Equidispersion refers to the assumption that y_i has mean μ_i equal to its variance, such that mean and variance are described by (Cameron et al. 2005, p. 668)

$$\mathrm{E}[y_i] = \mathrm{V}[y_i] = \mu_i. \tag{3.5}$$

Yet, overdispersion (i.e. variance > mean) may occur when the process leading to a first doctor visit might be different from the process resulting in subsequent doctor visits. This might be the case if the first visit to the doctor is solely the patient's choice, whereas subsequent consultations are also determined by the physician (Cameron et al. 2005, p. 674). Similarly, overdispersion may occur, if spells of illness make the occurrence of additional doctor visits more likely, thus violating the underlying independence assumption of the Poisson model (Cameron et al. 2008, p. 71).

These cases can be taken into account by introducing a scalar parameter $\alpha \leq 0$ which reflects such unobserved heterogeneity (Cameron et al. 2005, p. 675). Therefore, the conditional variance of y_i can be enhanced with the scalar parameter y_i, such that

$$\mathrm{V}[y_i|\ \mu_i,\ \alpha] = \omega_i = \mu_i + \alpha{\mu_i}^2. \tag{3.6}$$

Hence, variance is quadratic in the mean.[5] The Poisson case emerges when $\alpha = 0$. Given that $\alpha > 0$ and $\mu > 0$, conditional variance exceeds the conditional mean in the Negative Binomial case, in contrast to the rather restrictive equidispersion assumption ofhe Poisson regression model (Cameron et al. 2005, p. 676). The mean function remains the same as in the Poisson case

$$\mathrm{E}[y_i|\mu_i,\ \alpha] = \mu_i. \tag{3.7}$$

Within the Negative Binomial models the overdispersion parameter α is gamma-distributed, where $\Gamma(\alpha)$ implies (Cameron et al. 2008, p. 374)

[5] Cameron et al. (2005, 2008) refer to this case as Negative Binomial II, whereas they refer to the case in which variance is linear in the mean as Negative Binomial I. Hereinafter, for the sake of simplicity, the term Negative Binomial (NB) refers to what Cameron et al. (2005, 2008) refer to as Negative Binoial II.

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt, \ \alpha > 0. \tag{3.8}$$

Then the NB model has density (Cameron et al. 2008, p. 71)

$$f(y|\ \mu,\ \alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{\mu}. \tag{3.9}$$

The Negative Binomial model is an appropriate model when analysing count data outcomes, as it addresses both the restricted support of the outcome *y*, as well as its discrete nature. Thus, it avoids potential insufficiencies that would come along with the use of linear models (Cameron et al. 2005, p. 667). While the Poisson regression model also addresses heteroskedasticity, it requires equidispersion, whereas the Negative Binomial also handles overdispersion while containing the Poisson model (i.e. equidispersion) as a limiting case (if α = 0). Thus, the Negative Binomial model will be employed for further analysis. Cameron et al. (2008, p. 77) suggest data to be overdispersed, if its sample variance is twice the sample mean. Accordingly, looking at Tab. 1, the sample mean of the outcome visits to the GP is definitely overdispersed, with sample mean 2.23 and variance 8.58 (standard deviation: 2.93). Also, the outcome *emergency admissions* seems slightly overdispersed, with sample mean 0.06 and variance 0.08 (standard deviation: 0.29). Consequently, it seems appropriate to assume overdispersion and to employ the Negative Binomial model.

After maximum likelihood estimation interest lies in interpretation of the coefficients. Differentiating the exponential conditional mean

$$\mathrm{E}[y\ |\mathbf{x}_i, \mathrm{d}_i^{gk}, \mathrm{I}_i] = \exp(\mathbf{x}_i\beta + \mathrm{d}_i^{gk}\gamma + \mathrm{I}_i\lambda), \tag{3.10}$$

with respect to a one-unit change in the j^{th} regressor yields (Cameron et al. 2008, p. 80)

$$\frac{\delta \mathrm{E}[y\ |\mathbf{x}_i, \mathrm{d}_i^{gk}, \mathrm{I}_i]}{\delta x_j} = \beta_j \exp(\mathbf{x}_i\beta + \mathrm{d}_i^{gk}\gamma + \mathrm{I}_i\lambda). \tag{3.11}$$

Thus, β_j equals the proportionate change in the conditional mean. As individuals differ with respect to x, marginal effects differ across individuals as well. When an intercept term is used in the regression, calculation of average marginal effect is straightforward and simplifies to (Cameron et al. 2008, p. 80)

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\delta \mathrm{E}[y\ |\mathbf{x}_i, \mathrm{d}_i^{gk}, \mathrm{I}_i]}{\delta x_j} = \beta_j \bar{y}, \tag{3.12}$$

that is, the average marginal effect is equal to the coefficient size times sample mean.

Interpretation of a dummy regressor coefficient, such as the coefficient γ of the gatekeeping dummy d_i^{gk}, is similarly straightforward. A change in the dummy from zero to one yields

$$\frac{\mathrm{E}[y\ |\mathbf{x}_i, d_i^{gk} = 1,\ \mathrm{I}_i]}{\mathrm{E}[y\ |\mathbf{x}_i, d_i^{gk} = 0,\ \mathrm{I}_i]} = \frac{\exp(\mathbf{x}_i\beta + \gamma + \mathrm{I}_i\lambda)}{\exp(\mathbf{x}_i\beta + \mathrm{I}_i\lambda)} = \exp(\gamma). \tag{3.13}$$

That is to say, the conditional mean becomes exp(γ) times larger due to a change in the dummy from zero to one.

3.3 Maximum Simulated Likelihood

Healthcare utilisation and gatekeeping participation decisions are jointly modelled. Endogeneity arises, as the unobservable characteristics I_i enter both healthcare utilization and gatekeeping participation equations, while the gatekeeping equation also enters the utilisation equation separately.

After conditioning healthcare utilisation and gatekeeping participation equation on explanatory variables, including the common latent factors, the joint probability of gatekeeping participation d_i^{gk} and number of visits to the doctor (or emergency admissions) y_i becomes the product of the now conditionally independent probabilities (Deb et al. 2006c, p. 312):

$$\Pr(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i, \mathrm{I}_i) = \overbrace{g(\mathbf{x}_i\phi + \mathrm{I}_i\lambda)}^{gatekeeping} \times \overbrace{f(\mathbf{x}_i\beta + d_i^{gk}\gamma + \mathrm{I}_i\lambda)}^{healthcare\ utilisation}. \tag{3.14}$$

Difficulties in estimation arise, as the I_i are unknown (Deb et al. 2006c, p. 312). Consequently, the integral to Equation 3.14 does not have a closed form solution, which is a prerequisite for maximum likelihood estimation. Yet, if there is no such closed-form solution, maximum likelihood estimation may still be feasible, if an appropriate approximation can be found to evaluate the likelihood function (Cameron et al. 2005, p. 384). If the approximation of such an integral without analytical solution is achieved by simulation techniques, the estimator is called a *simulation-based estimator* (Cameron et al. 2005, p. 384). The way that integrals can be solved numerically, the use of this approximation within maximum likelihood estimation and different approaches to obtain such an approximation will be introduced subsequently.

3.3.1 Estimation

An integral can be solved numerically with simulation techniques. Let function *f*, as in Equation 3.1, depend for notational simplicity only on the unobservable characteristics I. Let the unobservable characteristics I be i.i.d. draws from the standard normal distribution, with probability density function denoted as $h(\mathrm{I})$. The integral to be evaluated is given by (Cameron et al. 2005, p. 390)

$$\mathrm{E}[f(\mathrm{I})] = \int f(\mathrm{I})h(\mathrm{I})d\mathrm{I}. \tag{3.15}$$

Here, simulation refers to the fact that integrating over a density is a form of averaging (Train 2009, p. 5). The simulation estimate of Equation 3.15, referred to as *simulator*, is given by (Cameron et al. 2005, p. 391)

$$\hat{\mathrm{E}}[f(\mathrm{I})] = \frac{1}{S}\sum_{s=1}^{S} f(\mathrm{I}^s), \tag{3.16}$$

where $\mathrm{I}^1, \ldots, \mathrm{I}^s$ denote S (S = 1, …, S) random draws of I from its density $h(\mathrm{I})$. As $\mathrm{S} \to \infty$ the estimate $\hat{\mathrm{E}}[f(\mathrm{I})]$ converges in probability to $\mathrm{E}[f(\mathrm{I})]$ (Cameron et al. 2005, p. 391).

This simulation approach can be applied to maximum likelihood estimation, resulting in *maximum simulated likelihood* estimation (Cameron et al. 2005, p. 393). Thus, as one cannot observe the latent factors I_i, their effect can be integrated out (Deb et al. 2006c, p. 312):

$$\Pr(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i) = \int \Pr(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i, \mathrm{I}_i)h(\mathrm{I}_i)d\mathrm{I}_i. \tag{3.17}$$

As no closed-form solution to Equation 3.17 exists, the integral can be approximated numerically with simulation-techniques. Thus, an individual's simulated likelihood contribution is given by

$$\Pr(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i) = \frac{1}{S}\sum_{s=1}^{S}\Pr(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i, \mathrm{I}_i^s), \tag{3.18}$$

where the effect of I_i has been integrated out of Equation 3.18, which is to say, the left hand side of Equation 3.18 does no longer contain values of I_i (Deb et al. 2006a, p. 749). Thus, the simulated log-likelihood function is given by

$$\ln \mathrm{L}(y_i, \mathrm{d}_i^{gk}|\mathbf{x}_i) = \sum_{i=1}^{N}\ln\left(\frac{1}{S}\sum_{s=1}^{S}\left[\underbrace{g(\mathbf{x}_i\phi + \mathrm{I}_i\lambda)}_{gatekeeping} \times \underbrace{f(\mathbf{x}_i\beta + d_i^{gk}\gamma + \mathrm{I}_i\lambda)}_{healthcare\ utilisation}\right]\right). \tag{3.19}$$

3.3.2 Properties

Simulation error may be introduced, as the estimator is simulated rather than calculated precisely (Train 2009, p. 253). Simulation error can be decomposed into simulation chatter, simulation noise and simulation bias (Train 2009). Simulation chatter is the result of using different random draws at each likelihood iteration. While simulation chatter might render MSL estimation impossible, it can be easily tackled, simply by using the same simulation draws for each observation (Cameron et al. 2010). Consequently, simulation chatter does neither depend on S nor on N. Simulation noise may be introduced as a deviation of each simulated value from its expectation. Simulation noise cancels out, thus it decreases with either increasing S, or with increasing N, even if S is fixed (Train 2009). Simulation bias may be introduced as a consequence of averaging over the natural logarithm, as the MSL simulator $\ln \hat{f}$, which is the average over S subsimulators, is biased for $\ln f$, even if the simulator $\hat{f}$ is unbiased for f (Cameron et al. 2005).[6]

After all, for the simulation bias to disappear, S and $N \to \infty$, while S must increase faster than $\sqrt{N}$, such that $\sqrt{N}/S \to 0$ (Gouriéroux et al. 1997). If this requirement is met, MSL is asymptotically normal, efficient and equivalent to maximum likelihood estimation (Gouriéroux et al. 1997; Train 2009).

However, this ratio does not state what S should be for given N, it only describes the properties of the MSL estimator as N increases (Greene 2008, p. 592). When choosing an appropriate S one has to keep in mind that S refers to the number of draws per observation. Accordingly, the total number of random variables, that is to say, memory consumption, equals number of observations N times draws per observation S. Consequently, increasing S comes with (potentially prohibitively) high computational cost. To find an appropriate number of S, Drukker (2006, p. 154) suggests to begin with $S = N^{0.55}$ and repeat the estimation with higher values of S, until point estimates and log-likelihood settle. Similarly, Greene (2008, p. 592) concludes that researches are left with no choice but experimenting with higher values of S (for fixed N) until numerical stability of the estimates is achieved. Deb et al. (2006b, p. 254), on the other hand, recommend using as large of S as computational reasonable.

[6] Gouriéroux et al. (1997) suggests an asymptotic bias-adjusted MSL-estimator. As this bias adjustment requires bias to be small, Cameron et al. (2005) add, that the usefulness of this procedure may vary from case to case, as the small bias-assumption may not always hold.

In search for guidance, applied pieces of health econometric research that employed MSL-estimation were analysed. The results, with respect to choices of S, are summarised in **Tab. 2**. Not only the quantity of the simulation draws is of importance, their quality matters as well. Quasi-random draws, such as the Halton-sequence, greatly reduce the required amount of simulation draws for a given level of precision (Greene 2008; Train 2009). Halton-draws are more evenly distributed than pseudo-random draws since they are negatively correlated by design (Train 2009, p. 224). While Halton-draws are rather deterministic than random, the randomness of draws is not as important as their uniform coverage over the domain of integration (Drukker et al. 2006). Consequently, Haltondraws were employed in the identified pieces of literature, as displayed in Tab. 2. Backed by the review, within the subsequent MSL-estimation 5,000 Halton-sequence draws will be used for each of the 45,000 observations, resulting in a ratio $\sqrt{N}/S = \frac{\sqrt{45{,}000}}{5{,}000} = 0.042$, which is slightly below the ratio achieved by Deb et al. (2006c).

***Tab. 2**: Overview of Choices Regarding S in the Literature*

Author	N	S	$\frac{\sqrt{N}}{S}$	Random variates
Bratti et al. (2011)	2,467	1,600	0.031	Halton
Deb et al. (2006c)	8,129	2,000	0.045	Halton
Deb et al (2006a)	26,514	1,000	0.162	Halton
Deb et al. (2006b)	5,033	400	0.177	Halton
Gardiol et al. (2005)	31,540	100	1.776	No information
Shane et al. (2012)	Did not report S			
Garrido et al. (2012)	Did not report S			

4 Results

Computations were conducted using Stata 14.2 by employing Deb (2009) user-written command `mtreatreg`, which is the Stata implementation of Deb et al. (2006c) *endogenous treatment regression model*. The `mtreatreg` command's ado-file was manipulated in a way such that it handles a binary choice equation, while otherwise only multinomial outcomes were possible for the participation equation. The way the respective adofile was manipulated is described in the appendix. Parallel to Deb et al. (2006c), the analysis begins with estimating the effect of gatekeeping participation on the respective healthcare outcomes without taking endogeneity into account. This procedure will be referred to as *exogeneity assumption*. Then, the exercise will be repeated by employing Deb et al. (2006c) endogenous treatment regression model, thus taking endogeneity into account (referred to as *endogeneity assumption*).

4.1 Assuming Exogeneity

To begin with, results that assume gatekeeping participation to be exogenous with respect to healthcare utilisation are discussed. First, gatekeeping participation is estimated with a logit model in column (1) of **Tab. 3**. While the coefficients of both age terms are significantly different from zero, the age-term is positive, while the aqe-squared-term is negative. Also, the number of chronic conditions is significant and positive. Thus, individuals who suffer from more chronic conditions are more likely to subscribe to the gatekeeping programme.

Assuming exogeneity of gatekeeping participation with respect to healthcare utilisation, the four outcomes of interest are estimated in columns (2) to (5) of Tab. 3, where the interest lies in the coefficient of the gatekeeping dummy. Visits to the GP and emergency admissions are estimated with a NB model and displayed in columns (2) and (3), respectively.

***Tab. 3**: Results: exogeneity Assumption*

	(1) Gatekeeping	(2) GP	(3) Emergency	(4) ln(ambulatory)	(5) ln(total)
Gatekeeping (γ)		0.379**	0.199**	0.656***	0.575***
Age	–0.026*	0.016***	–0.050***	–0.039***	0.008
Age^2	0.001***	0.000***	0.001***	0.001***	0.000***
Sex (male)	–0.025	–0.302***	–0.157***	–0.932***	–0.563***
ln(Income)	0.010	–0.019***	–0.039**	0.028***	–0.109***
ln(pop. Density)	0.035	–0.004	–0.020	–0.004	–0.002
Number chronic cond.	0.404***	0.461***	0.494***	0.756***	1.482***
Constant	–3.290***	–0.427***	–1.531***	5.699***	3.782***
Ln(α)		0.209***	1.578***		
Observations	45 000	45 000	45 000	45 000	45 000
Pseudo R2	0.01	0.03	0.01		
R2				0.111	0.100

*Legend: * p < 0.10, ** p < 0.05, *** p < 0.01*

The coefficients of the gatekeeping enrolment dummies are both positive and significantly different from zero. Thus, being enrolled in gatekeeping is associated with a 46.1 % (exp(0.379) ≈ 1.461) increase in visits to the GP, while emergency admissions increase by 22.0 % (exp(0.199) ≈ 1.220). In both equations the $\ln(\alpha)$ coefficients are larger than 0, meaning $\alpha > 1$, indicating that the choice of the NB model over the Poisson is justified, as in the Poisson case $\alpha = 0$.

Results of the OLS-regression of logarithmised ambulatory costs and logarithmised total medical costs are displayed in columns (4) and (5) of Tab. 3, respectively. In these two outcomes, the gatekeeping dummy coefficients are positive and significantly different from zero. Thus, being enrolled in gatekeeping, is associated with a 92.7 % (exp(0.656) ≈ 1:927) increase in ambulatory costs as well as a 77.7 % (exp(0.575) ≈ 1.777) increase in total medical costs. So far it could be seen that gatekeeping enrolment is associated with a statistically significant increase in all four healthcare utilisation measures. Yet, it remains unclear whether or not these estimates are biased by unobservable self-selection into gatekeeping.

4.2 Taking Endogeneity into Account

The possibility of gatekeeping being endogenous with respect to healthcare utilization is considered subsequently. Within the endogenous treatment regression model by Deb et al. (2006c) each outcome equation is estimated jointly with a participation equation.[7] The estimates of the gatekeeping participation equations of each of the four models are similar, as the same outcome (i.e. gatekeeping participation) was estimated with the same set of observable characteristics for the same individuals (only slightly differing due to different values for λ).

The results of the gatekeeping participation equation, modelled as a logistic outcome, are displayed in column (1) of **Tab. 4**. The gatekeeping participation equation displayed in column (1) of Tab. 4) was jointly estimated with the visits to the GP equation (column 2 of Tab. 4).8 Here, bot age-coefficients and the number of chronic conditions coefficient are significantly different from zero.

Within the four healthcare utilisation measures of Tab. 4, the factor loadings λ of the latent factors as well as the estimated coefficients of the gatekeeping dummy variable γ are of main interest. The factor loading coefficients λ have a natural interpretation as proxies for selection on unobservables and can be interpreted in the same way as coefficients of observable characteristics (Deb et al. 2006c, p. 308). A significant positive (negative) factor loading coefficient can be interpreted as unobserved factors, that both increase probability of participating in gatekeeping while also leading to higher (lower) utilisation, relative to someone who was randomly assigned to gatekeeping (Deb et al. 2006c, p. 321). The λ factor loading coefficients are highly significant for all four healthcare utilisation measures. That is to say, for each healthcare outcome, the same unobserved characteristics that positively contribute to utilisation at the same time make enrolment into gatekeeping more likely. Thus, the null-hypothesis that selection on unobservables does not occur can be rejected. That is to say, there are non-trivial effects of selection on unobservables into the gatekeeping programme.

After correcting for selection on unobservables (i.e. λ) the coefficients of the gatekeeping participation dummies γ in Tab. 4 columns (2) to (4) are unbiased with respect to self-selection. Thus, the remaining effect of gatekeeping participation can be interpreted as a causal incentive effect (Deb et al. 2006c, p. 325). The γ coefficients are significantly different from zero in all four healthcare utilisation measures.

7 Here, the normal outcomes ambulatory costs and total costs will be estimated with MSL instead of OLS.

8 A full table including all omitted gatekeeping participation equations can be seen in the appendix.

The gatekeeping programme increases GP consultations (versus non-participants) by 10.0 % (exp(0.096) ≈ 1.100). Also, gatekeeping (significantly) decreases emergency admissions by 55.7 % (exp(–0.814) ≈ 0.443). Both ambulatory and total costs are decreased by 63.4 % (exp(–1.004) ≈ 0.366) and 32.8 % (exp(–0.398) ≈ 0.672), respectively. Consequently, after taking into account selection effects (i.e. selection bias), the gatekeeping programme unfolds non-trivial incentive effects with respect to healthcare utilisation. Also, the ln(α) coefficients, associated with the NB model, are smaller than under the exogeneity assumption. This may be the case, as the α coefficients reflect unobserved heterogeneity, similar to the λ coefficients. This finding indicates that introducing the latent factors properly captures some of the underlying unobserved heterogeneity.

***Tab. 4**: Results: MSL*

	(1) Gatekeeping	(2) GP	(3) Emergency	(4) ln(ambulatory)	(5) ln(total)
Gatekeeping (γ)		0.096**	0.914**	1.004***	0.389***
Age	–0.030**	0.015***	–0.053***	–0.043***	0.006
Age^2	0.001***	0.000***	0.001***	0.001***	0.000***
Sex (male)	–0.037	–0.309***	–0.150***	–0.931***	–0.564***
ln(Income)	0.010	–0.018***	–0.039**	0.030***	–0.109***
ln(pop. Density)	0.038	–0.004	–0.011	–0.000	–0.001
Number chronic cond.	0.445***	0.480***	0.545***	0.791***	1.508***
Constant	–3.685***	–0.452***	–2.092***	5.765***	1.035***
Gatekeeping (λ)	0.306***	0.306***	1.113***	1.754***	
Ln(α)	0.099***	0.099***	0.201		
Observations	45 000	45 000	45 000	45 000	45 000
Simulation draws	5 000	5 000	5 000	5 000	5 000
log-L (endog.)	–79 878.82	–79 878.82	–17 842.83	–97 483.22	–116 753.41
log-L (exog.)	–79 896.74	–79 896-74	–17 852.89	–97 799.65	–116 754.84

Legend: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3 Test for Exogeneity of Gatekeeping

The null-hypothesis of exogeneity of gatekeeping may be additionally challenged with a likelihood-ratio test, as described in Deb et al. (2006b, p. 253). Here, a test for exogeneity is a test for the hypothesis that λ is equal to zero, where the likelihood-ratio statistic follows a χ^2 distribution with one degree of freedom, as one parameter, λ, is tested (Deb et al. 2006b, p. 253). The constrained log-likelihood ("log-L (exog.)", as displayed in Tab. 4 is given by the sum of the log-likelihoods of the participation and outcome equation that are calculated under the exogeneity assumption (Deb et al. 2006b, p. 253).[9] The results of the likelihood-ratio test are displayed in **Tab. 5**. The null hypothesis of exogeneity of

[9] With the `mtreatreg` command both participation and outcome equation are estimated under the exogeneity assumption with a regular maximum likelihood approach to provide starting values for the MSL estimation. Thus, ereturn provides both the likelihood assuming exogeneity "log-L (exog.)" and the likelihood accounting for endogeneity "log-L (endo.)".

gatekeeping can decisively be rejected for the healthcare utilisation measures *visits to the GP, emergency admissions and ambulatory costs*, while it can only be weakly rejected for the outcome *total costs*.

***Tab. 5**: Likelihood-Ratio Test for Exogeneity of Gatekeeping*

	GP	Emergency	ln(ambulantory)	ln(total)
log-L (exog.)	–79878.82	–17842.83	–97483.22	–116753.41
log-L (endo.)	–79896.74	–17852.89	–97799.65	–116754.84
LR	35.84	20.13	632.85	2.85
p-value	0.00	0.00	0.00	0.09

4.4 Incentive and Selection Effects

The marginal effect of gatekeeping participation, calculated under the exogeneity assumption, incorporates both the causal incentive effect as well as the selection effect (Deb et al. 2006b, p. 325).

Similarly, the marginal effect of gatekeeping participation estimated under the endogeneity assumption dentifies the causal incentive effect, as if under experimental conditions (Deb et al. 2006b, p. 325). Thus, the difference of the marginal (incentive) effect of gatekeeping, obtained with and without taking self-selection into account, serves as a approximation to the selection effect (Deb et al. 2006c, p. 325). The corresponding marginal effects are summarised in **Tab. 6**. Here, the first row of Tab. 6 corresponds to the marginal effect of gatekeeping under the exogeneity assumption (as in Tab. 3), while row two corresponds to the marginal effect of gatekeeping obtained while taking endogeneity into account (as in Tab. 4). Thus, the results in the second row of Tab. 6 represent the causal incentive effect of gatekeeping on the healthcare utilisation measures. The results of row three of Tab. 6 are calculated as the difference of rows one and two and thus provide an approximation of the magnitude of the selection effect of gatekeeping on healthcare utilisation, as explained above.

An individual randomly assigned to gatekeeping would have 0.153 more visits to the GP per year compared to a person assigned to non-gatekeeping (incentive effect, row two). A person who deliberately chose to enrol into gatekeeping would have 0.430 more visits to the GP per year compared to a person who chose not to enrol into gatekeeping (selection effect, row three). As can be seen, the (marginal) selection effect clearly outweighs the (marginal) incentive effect in all four healthcare outcomes. What can also be seen is that the properly identified incentive effect (endogeneity assumption) clearly diverges from the biased incentive effect (exogeneity assumption) in each case. Except for *visits to the GP*, the incentive effect identified under the endogeneity assumption even has a different sign than the biased ones (exogeneity assumption).

***Tab. 6**: Incentive and Selection Effects of Gatekeeping on Healthcare Utilisation*

	GP	Emergency	ln(ambulantory)	ln(total)
Incentive Effect (exog.)	0.583	0.011	0.656	0.575
Incentive Effect (endo.)	0.153	–0.017	–1.004	–0.398
LR	0.430	0.028	1.660	0.973

Note: The displayed effects are the respective marginal effects.

4.5 Robustness with Respect to Simultation Draws

The MSL-results do sensitively depend upon on the number of simulation draws employed, as a too small amount of S (relative to N) may introduce simulation error. Whether or not enough simulation draws were used remains a difficult question to answer (Cameron et al. 2005, p. 396). Literature suggests to experiment with different sizes of S until numerical stability of the estimators is achieved (Cameron et al. 2005; Drukker 2006; Greene 2009). Consequently, each outcome was estimated eleven times, with different sizes of S, beginning from S = 100 and then going in increments of 500 from S = 500 to S = 5; 000. The robustness of the λ coefficient with respect to different values of S for the four healthcare outcomes (with the same underlying sample, thus N = 45,000) are summarised graphically and discussed below.

Visits to the GP

The λ coefficient for the healthcare measure *visits to the GP* seems to be well-behaved with respect to different sizes of S (cf. **Fig. 1**). Starting from S = 100; λ is remarkably stable with respect to S. These results imply that choosing S = 5; 000 is more than sufficient to tackle simulation error.

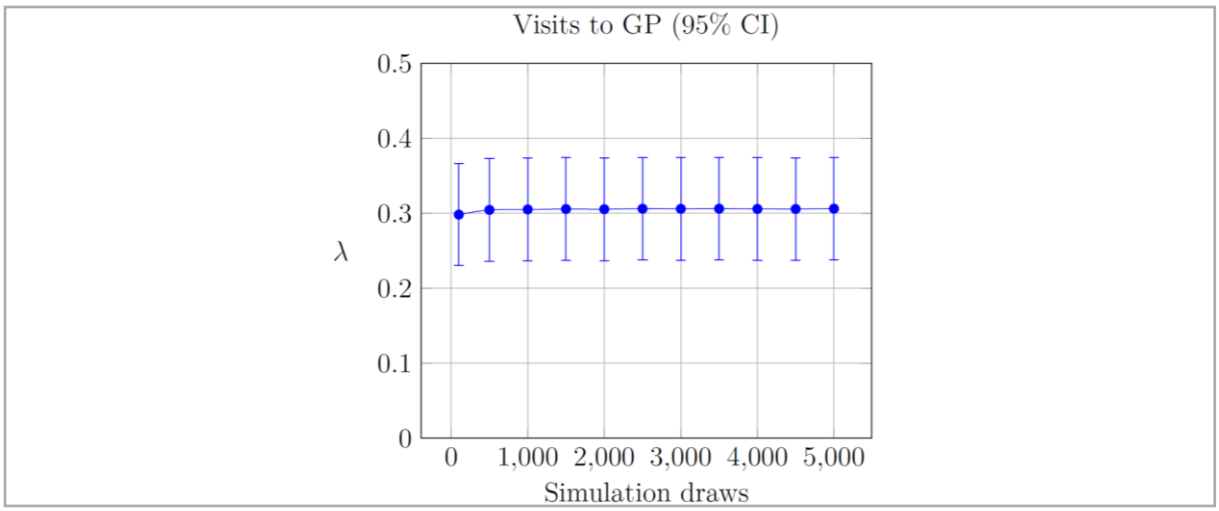

***Fig. 1**: Visits to GP – Robustness of λ w.r.t. Simulation Draws*

Emergency Admissions

The λ coefficient for the healthcare measure *emergency admissions* fluctuates somewhat for different values of S (cf. **Fig. 2**). Nevertheless, beginning with values from S = 1; 000, the λ coefficient seems to be relatively close to its final value, derived from S = 5, 000. Interestingly, for all values but S = 3,000, the λ coefficient is significantly different from zero. This notable deviation at S = 3; 000 emphasises the importance of conducting such robustness analysis. After all, the results imply that S = 5, 000 is a sufficient amount of simulation draws.

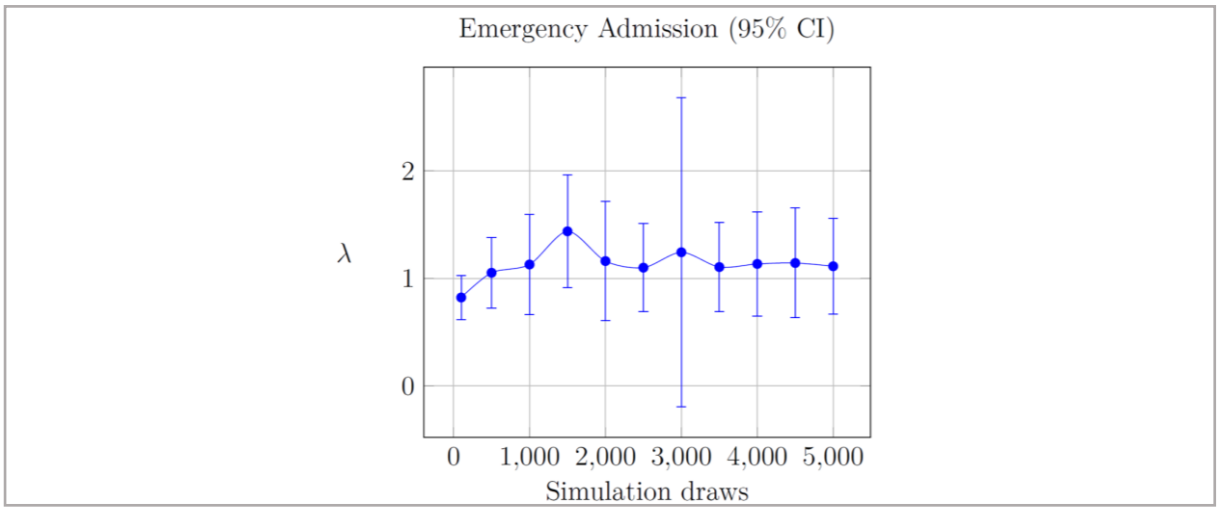

***Fig. 2**: Emergency Admissions – Robustness of λ w.r.t. Simulation Draws*

Ambulatory Costs

The λ coefficient for the healthcare measure *ambulatory costs* exhibits a converging behaviour for increasing values of S (cf. **Fig. 3**). For each increase in S the λ coefficient seems to converge more closely the value derived from S = 5, 000. For values larger than S = 3, 500 the λ coefficient settles. Thus, the results create confidence that the number of simulation draws (i.e. S = 5, 000) was sufficiently large.

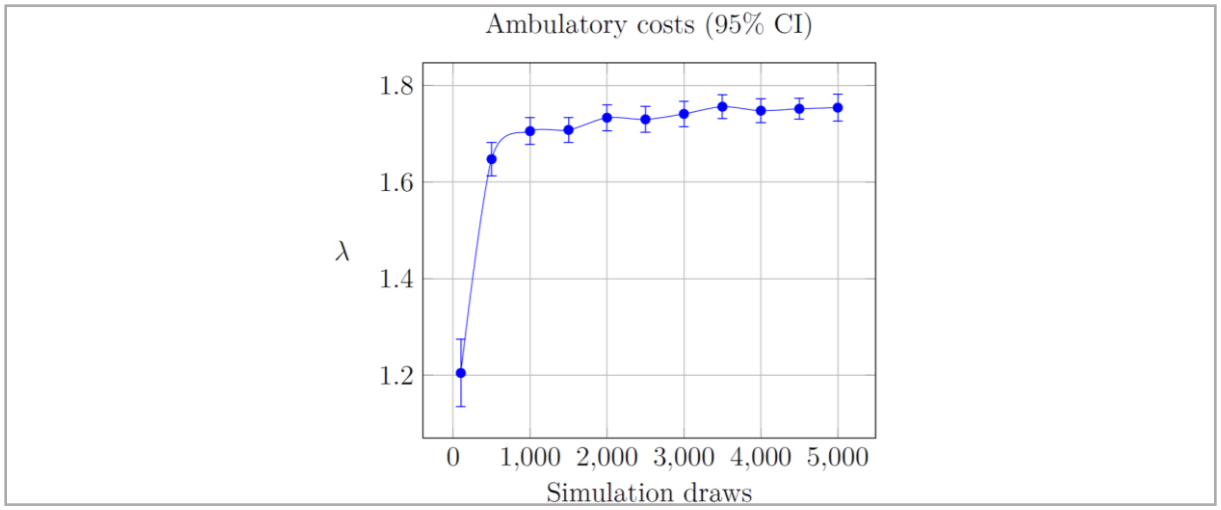

***Fig. 3**: Ambulatory Costs – Robustness of λ w.r.t. Simulation Draws*

Total Costs

The λ coefficient for the healthcare measure total costs exhibits a relatively steady behavior for different values of S (cf. **Fig. 4**). While the λ coefficient for S = 100 is not significantly different from zero, all following values, beginning from S = 500, are almost identical to the final value derived from S = 5,000. Once again it seems that S = 5,000 is sufficiently large to tackle simulation error. Regarding all four healthcare measures, the robustness analysis strengthens confidence that results are not biased by simulation error.

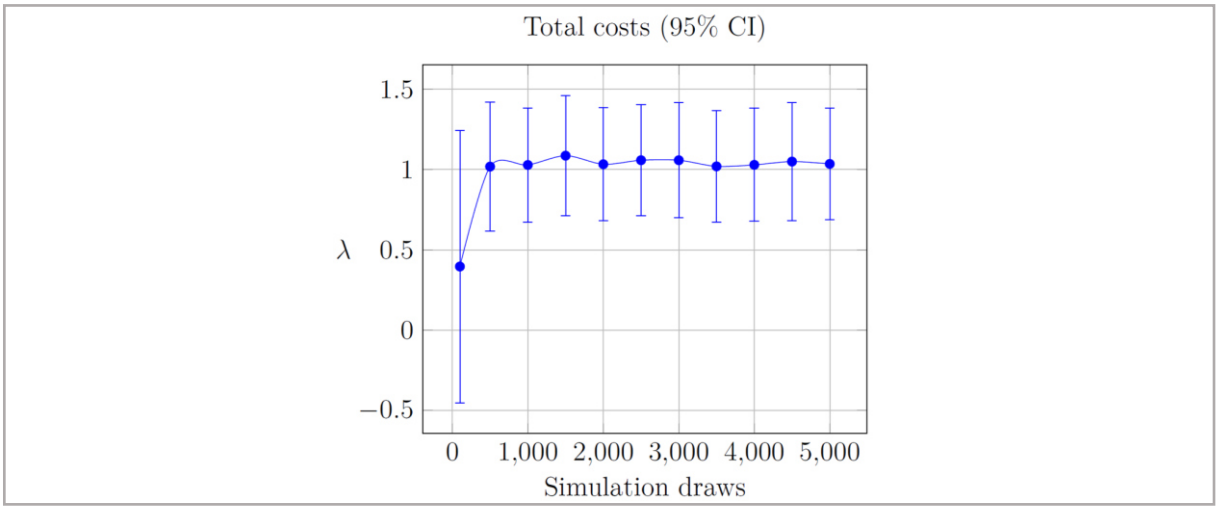

***Fig. 4**: Total Costs – Robustness of λ w.r.t. Simulation Draws*

5 Discussion

While the aim of employing the MSL-approach within this piece of research was to tackle bias introduced by selection on unobservables, simulation bias might instead be introduced if an inappropriate amount of simulation draws S was used. Whether or not one has used a sufficient amount of simulation draws remains a challenging question to answer (Cameron et al. 2005, p. 396). As no empirical guidance exists, theory suggests to experiment with different sizes of S until numerical stability of the estimator is achieved (Cameron et al. 2010; Drukker et al. 2006; Greene 2008). While some (e.g. Atella et al. 2008; Bratti et al. 2011; Deb et al. 2006a,c) report to have relied on such experimentation, Deb et al. (2006b, p. 254) recommend using as large of an S as computational reasonable. Due to having relative large N, Deb et al. (2006a) report to have used smaller S than desired, to ensure convergence of their model. Even more so, Atella et al. (2008) report that one of their models did not converge after four days of CPU time. Deb et al. (2006c) state that their choice of S was based on other empirical studies that use MSL. Similarly, within this piece of research, the choice of S was based on other empirical studies that employ MSL. Also, robustness analysis suggests that the amount of simulation draws S employed in this piece of research was sufficiently large, such that simulation error can be ruled out. Thus, confidence in the validity of the results is strengthened.

The MSL-approach is computationally burdensome, as it makes extensive use of simulation techniques (Cameron et al. 2005, p. 384), as also explicitly mentioned in several pieces of applied research (e.g. Atella et al. 2008; Deb et al. 2006a,b). Consequently, as computational resources are fixed, favouring relatively large S (thus, high precision) comes at the cost of limiting sample size N. It could be shown that S = 5; 000 was sufficiently large, yet, this insight could only by gained ex-post. Even though a literature review was conducted to motivate the choice of S, the outcome was unclear beforehand. Thus, the trade-off between simulation draws and sample-size was decided in favour of precision. If MSL simulation error properties were generally better understood, a better trade-off between sample size and simulation draws might have been possible. Nevertheless, the employed sample size is still considerably larger than those used in other studies that were discussed earlier (cf. Tab. 2). Thus, the underlying sample size does not seem to be a limiting factor.

Within this piece of research, all variables in the gatekeeping participation equation were included in the healthcare utilisation equations as well. Deb et al. (2006c) suggest to employ exclusion restrictions, by specifying instrumental variables in the insurance choice equation that are excluded from the utilisation equation. To do so, Deb et al. (2006c) employ variables that they consider to be insurance choice-related, while being exogenous to healthcare consumption, such as employment information (union membership, number of employees in firm, whether or not firm has multiple locations, etc.), which are included in their healthcare survey. Yet, as such information is not available within claims data, no such exclusion restriction was employed within this piece of research. Whether or not the lack of such exclusion restriction may have affected the results remains unclear. Yet, Deb et al. (2006b) note that identification may be sufficiently achieved even if all variables in the insurance equation are included in the utilisation equation.

6 Conclusion

The aim of this piece of research was to assess whether or not gatekeeping participation is exogenous with respect to healthcare utilisation (*visits to the GP, emergency admissions, ambulatory costs, total costs*). To do so, the endogenous treatment regression model by Deb et al. (2006c) was employed. A literature review to make an informed decision about the required number of simulation draws was conducted. Results imply that the latent factors, representing unobservable characteristics that jointly determine healthcare utilisation behaviour and gatekeeping participation, are significantly different from zero in all four cases. Thus, the null-hypothesis of exogeneity of gatekeeping participation could clearly be rejected. Also, a likelihood-ratio test strongly rejected the exogeneity assumption of gatekeeping for the outcomes *visits to the GP, emergency admissions, ambulatory costs*, while weakly rejecting it for *total costs*. Thus, when not taking endogeneity of gatekeeping into account, effects of gatekeeping on healthcare utilisation are biased by unobservable self-selection. But, if endogeneity is properly controlled for, causal incentive effects of gatekeeping on healthcare utilisation can be identified. Also, the effect of selection on unobservables can be quantified. Extensive robustness analysis show that the employed amount of simulation draws was more than sufficient to rule out simulation bias.

Results imply that the selection effects outweigh the incentive effects. The effect of selection on unobservables is positive in all four outcomes, indicating that individuals self-select into the gatekeeping programme according to their unobservable healthcare needs. It could also be seen that the gatekeeping programme does successfully unfold the implied incentive effects, as the role of the GP is strengthened, emergency admissions are avoided, while reducing ambulatory and total (medical) costs. Consequently, it is possible to identify both selection and incentive effects of gatekeeping on healthcare-utilisation.

References

Akerlof, G.-A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics 84*(3), 488–500.

Arrow, K.-J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review 53*(5), 941–973.

Atella, V., & Deb, P. (2008). Are primary care physicians, public and private sector specialists substitutes or complements? Evidence from a simultaneous equations model for count data. *Journal of health economics (27)*3, 770–785.

Bratti, M., & Miranda, A. (2011). Endogenous Treatment Effects for Count Data Models with Sample Selection or Endogenous Participation. *Health economics* 9, 1090–1109.

Cameron, A.-C., & Trivedi, P.-K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Cameron, A.-C., & Trivedi, P.-K. (2008). *Regression analysis of count data*. (Econometric Society monographs, 7. print. Vol. 30). Cambridge University Press.

Cameron, A.-C., & Trivedi, P.-K. (2010). *Microeconometrics using Stata*. Rev. ed. College Station: Stata Press.

Cameron, A.-C, Trivedi, P.-K., Milne, F., & Piggott, J. (1988). A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. *The Review of Economic Studies (55)*1, 85.

Deb, P. (2009). *MTREATREG: Stata module to fits models with multinomial treatments and continuous, count and binary outcomes using maximum simulated likelihood*. Online: https://EconPapers.repec.org/RePEc:boc:bocode:s457064.

Deb, P., Li, C., Trivedi, P. K., & Zimmer, D. M. (2006a). The effect of managed care on use of health care services: results from two contemporaneous household surveys. *Health economics (15)*7, 743–760.

Deb, P., & Trivedi, P. K. (2006b). Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. *Stata Journal (6)*2, 246–255.

Deb, P., & Trivedi, P. K. (2006c). Specification and simulated likelihood estimation of a non–normal treatment–outcome model with selection: Application to health care utilization. *The Econometrics Journal (9)*2, 307–331.

Drukker, D. M. (2006). Maximum simulated likelihood: Introduction to a special issue. *Stata Journal (6)*2, 153–155(3).

Drukker, D. M., & Gates, R. (2006). Generating Halton sequences using Mata. *Stata Journal (6)*2, 214–228(15).

Gardiol, L., Geoffard, P.-Y., & Grandchamp, C. (2005). Separating selection and incentive effects in health insurance. *PSE Working Papers 38*.

Garrido, M. M., Deb, P., Burgess, J. F., & Penrod, J. D. (2012). Choosing models for health care cost analyses: issues of nonlinearity and endogeneity. *Health services research (47)*6, 2377–2397.

Gouriéroux, C., & Monfort, A. (1997). *Simulation-based Econometric Methods*. Oxford University Press.

Greene, W. (2008). *Econometric analysis*. (6. Ed.). Upper Saddle River: Pearson Prentice Hall.

Greene, W. (2009). Models for count data with endogenous participation. *Empirical Economics (36)*1, 133–173.

Hofmann, S. M., & Mühlenweg, A. M. (2016). Gatekeeping in German Primary Health Care: Impacts on Coordination of Care, Quality Indicators and Ambulatory Costs. *CINCH – Health Economics Research Center 5*.

Klora, M., Zeidler, J., May, M., Raabe, N., & Graf von der Schulenburg, J.-M. (2017). Evaluation der hausarztzentrierten Versorgung in Deutschland anhand von GKV-Routinedaten der AOK Rheinland/Hamburg. *Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen 120*, 21–30.

Mehl, E., & Weiß, I. (2015). Selektivverträge am Beispiel der Hausarztmodelle. C. Thielscher (Ed.). *Medizinökonomie*. FOM-Edition, FOM Hochschule für Oekonomie & Management. Wiesbaden: Springer Gabler, 633–662.

O'Donnell, O., van Doorslaer, E., Wagstaff, A., & Lindelow, M. (2007). *Analyzing Health Equity Using Household Survey Data. The World Bank.* OpenStreetMap (2019). Online: https://www.suche-postleitzahl.org.

Schneider, A., Donnarchie, E., Tauscher, M., Gerlach, R., Maier, W., Mielck, A., Linde, K., & Mehring, M. (2016). Costs of coordinated versus uncoordinated care in Germany: results of a routine data analysis in Bavaria. *BMJ open (6)*6, e011621.

Shane, D., & Trivedi, P. K. (2012). What Drives Differences in Health Care Demand? The Role of Health Insurance and Selection Bias. *Health, Econometrics and Data Group Working Papers 12*(09).

Szecsenyi, J., & Gerlach, F. (2018). *Evaluation der Hausarztzentrierten Versorgung (HZV) in Baden-Württemberg: Zusammenfassung der Ergebnisse – Ausgabe 2018*. Universitätsklinikum Heidelberg: Abteilung Allgemeinmedizin und Versorgungsforschung, & Goethe-Universität Frankfurt am Main: Institut für Allgemeinmedizin (Hrsg.). Online: https://aok-bw-presse.de/fileadmin/mediathek/dokumente/hzv-evaluation_2018.pdf.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Weinhold, I., & Gurtner, S. (2014). Understanding shortages of sufficient health care in rural areas. *Health Policy (118)*2, 201–214.

Weinhold, I., & Gurtner, S. (2018). Rural - urban differences in determinants of patient satisfaction with primary care. *Social science & medicine 212*, 76–85.

Wensing, M., Szecsenyi, J., Stock, C., Kaufmann Kolle, P., & Laux, G. (2017). Evaluation of a program to strengthen general practice care for patients with chronic disease in Germany. *BMC health services research (17)*1, 62.

Appendix

mtreatreg.ado File

As described above, the user written `mtreatreg.ado` file was manipulated, such that it does not only handle multinomial, but also binary outcomes for the treatment equation. For the sake of transparency and reproducibility, the manipulations are summarised here. The manipulated ado-file was saved as `mtreatreg2.ado`. The `mtreatreg2.ado`-file is also made available by the author.

Line Nr.	Changed into	Comment
15	`program mtreatreg2, sortpreserve`	new command name
92	If ‘nalt’ < 2 {	min. number of treatments 2 instead of 3

Results: Full MSL-Results

The full table including all MSL-Results is displayed subsequently (cf. **Tab. 7**). Here, each healthcare utilisation outcome is displayed along with its own gatekeeping participation equation. As can be seen, the gatekeeping participation results are very similar to each other, as they use the same set of observable characteristics on the same outcome, only slightly differing due to different values of λ.

***Tab. 7**: Full MSL-Results; * p < 0.10, ** p < 0.05, *** p < 0.01*

	GP	Emergency	ln(ambulatory)	ln(total)
Gatekeeping Participation				
Age	–0.030**	–0.029***	–0.050***	0.031**
Age^2	0.001***	0.001***	0.001***	0.001***
Sex (male)	–0.037	–0.032***	–0.226***	–0.047
ln(Income)	0.010	0.010	0.030*	0.010
ln(pop. Density)	0.038	0.038	0.037	0.037
Number chronic cond.	0.445***	0.442***	0.543***	0.453***
Constant	–3.685***	–3.708***	–3.275***	–3.655***
Healthcare Utilisation				
Gatekeeping (λ)	0.096**	–0.814***	–1.004***	–0.398**
Age	0.015***	–0.053***	–0.043***	0.006
Age^2	0.000***	0.001***	0.001***	0.000***
Sex (male)	–0.309***	–0.150***	–0.931***	–0.564***
ln(Income)	–0.018***	–0.039**	0.030***	–0.109***
ln(pop. Density)	–0.004	–0.011	0.000	–0.001
Number chronic cond.	0.480***	0.545***	0.791***	1.508***
Constant	–0.452***	–2.092***	5.765***	3.833***
Gatekeeping (λ)	0.306***	1.113***	1.754***	1.035***
Ln(α)	0.099***	0.201		
Observations	45 000	45 000	45 000	45 000
Simulation draws	5 000	5 000	5 000	5 000
log-L (endog.)	–79 878.8	–17 842.8	–97 483.2	–116 753.4
log-L (exog.)	–79 896.7	–17 852.9	–97 799.6	–116 754.8