Estimating the Impact of Medical Care Usage on Work Absenteeism by a Trivariate Probit Model with Two Binary Endogenous Explanatory Variables

Panagiota Filippou
Deloitte UK
panagiota.filippou@hotmail.com

Giampiero Marra
Department of Statistical Science
University College London
giampiero.marra@ucl.ac.uk

Rosalba Radice
Bayes Business School
City, University of London
rosalba.radice@city.ac.uk

David Zimmer
Department of Economics
Western Kentucky University
david.zimmer@wku.edu
corresponding author

July 12, 2022

Abstract

The aim of this paper is to estimate the effects of seeking medical care on missing work. Specifically, our case study explores the question: Does visiting a medical provider cause an employee to miss work? To address this, we employ a model that can consistently estimate the impacts of two endogenous binary regressors. The model is based on three equations connected via a multivariate Gaussian distribution, which makes it possible to model the correlations among the equations, hence accounting for unobserved heterogeneity. Parameter estimation is reliably carried out via a trust region algorithm with analytical derivative information. We find that, observationally, having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlates with a smaller 13 percent increase in the likelihood of missing work. However, after addressing potential endogeneity, neither type of visit appears to significantly relate to missing work. That finding also applies to visits that occur during the previous year. Therefore, we conclude that the observed links between medical usage and absenteeism derive from unobserved heterogeneity, rather than direct causal channels. The modeling framework is available through the `R` package `GJRM`.

Key Words: endogeneity, medical care, multivariate Gaussian, simultaneous equation model, work absenteeism.

1 Introduction

The aim of this work is to estimate the effects of seeking medical care on missing work. Labor economists have long recognized that injury or illness represents one of the most common reasons for worker absenteeism ("The Causes And Costs Of Absenteeism In The Workplace", Forbes, 7/10/2013). The U.S. Centers for Disease Control and Prevention (CDC) estimates that health-related worker absenteeism costs employers $225.8 billion annually, or about $1,685 per worker ("Worker Illness and Injury Costs U.S. Employers $225.8 Billion Annually", CDC Foundation Report, 1/28/2015). Those large dollar amounts call for rigorous studies exploring the specific channels through which medical events relate to missing work.

A voluminous body of research, scattered across a wide range of academic disciplines, explores the effects of health-related issues on productivity (Nicholson et al., 2006; Pauly, Nicholson and Polsky, 2008; Shultz, Chen and Edington, 2009; Zhang, Bansback and Anis, 2011; Zhang et al., 2015; Stromberg et al., 2017). Some of the extant literature looks specifically at absenteeism, while other studies investigate the related issue of "presenteeism", defined as ill employees showing up at work. At the risk of over-generalizing such a large body of literature, the consensus appears to be that health-related problems, including absenteeism, hinder worker productivity.

This paper explores a more narrowly-targeted question: Does visiting a medical provider cause an employee to miss work? Providing an answer to that question must confront two complications. First, employees likely possess unobserved (to the researcher) traits that increase their likelihood of visiting a medical provider while also increasing their chances of missing work. The most obvious such traits involve unobserved health problems, but other attributes, such as job satisfaction and attitudes toward health care providers, likely also muddle the observed

relationship between seeking medical care and missing work. In the jargon of econometrics, seeking medical care likely is endogenous with respect to missing work.

The second complication is that reasons for medical visits show substantial heterogeneity, and those reasons likely relate to the probability of missing work. Some visits to medical providers, which we label "curative", involve the diagnosis or treatment of some medical problem, while other visits, which we call "preventive", concern routine checkups and other wellness activities. Curative and preventive medical usage likely affect absenteeism differently for two reasons. First, preventive visits, presumably being more predictable and less urgent, might be easier to schedule around work hours. Second, the recently-passed Affordable Care Act includes provisions that nudge people toward preventive services and away from curative care, in the belief that shifting the mix of care will reduce aggregate medical expenses in the long run. Specifically, most co-pays for preventive services have been eliminated, while deductibles for curative services have increased (largely in response to the so-called "Cadillac Tax"). Some employers have gone further, by bringing medical professionals on-site, so that workers can obtain preventive services without missing work. Curative services, by contrast, might be less predictable, and therefore more difficult to schedule around work hours. Moreover, employers might require confirmation from a physician before granting sick leave. Therefore, curative visits and absenteeism might have a somewhat mechanical link, though no less economically damaging in light of the aforementioned CDC report on the aggregate harm from absenteeism. Thus, it seems likely that curative and preventive care have different effects on absenteeism, and therefore should be considered separately.

The effects of curative and preventive visits on absenteeism also might depend upon the *timing* of visits. The previous paragraph argues that preventive visits might not increase

absenteeism, while curative visits might. But if those visits happened during previous periods, and if preventive visits, all else equal, improve future health, then preventive visits might actually *reduce* absenteeism during subsequent periods. Meanwhile, previous-period curative visits might have no impact on future absenteeism, so long as treatment was effective.

To address this topic, we employ a statistical model that accommodates two endogenous explanatory variables in an outcome equation. Specifically, we consider a recursive trivariate probit model which can be regarded as an extension of the recursive bivariate model introduced by Marra and Radice (2011) (see also references therein). The adopted modeling framework is based on Filippou, Marra and Radice (2017) and Filippou et al. (2019), which makes it possible to specify and reliably estimate all the parameters of the trivariate model needed for this work. Model fitting is carried out via a trust region algorithm with analytical derivative information. The modeling framework is available through the `R` package `GJRM` (2022), is quick and easy to use, and provides easy-to-interpret treatment effects. Note that trivariate probit models can be traced back to the work by Ashford and Sowden (1970) on multivariate probit models. Chib and Greenberg (1998) later proposed a Bayesian approach for estimating such models. In terms of software, the available alternatives are function `mvprobit()` (Cappellari and Jenkins, 2003) in `STATA`, and the `R` `mvProbit` package (Henningsen, 2021). However, `mvProbit` may be unusably slow (as pointed out by the developer of the package) and it requires all equations to have the same set of covariates (a requirement that is not met in our case study). The `STATA` function `mvprobit()` has instead been found to produce significantly biased estimates of the correlation parameters (Filippou, Marra and Radice, 2017) which makes its application to our case study problematic. This is because endogenous effects can be consistently estimated if correlation coefficients can also be consistently estimated (e.g., Chib and Greenberg, 2007).

Observationally, we find that having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlates with a smaller 13 percent increase in the likelihood of missing work. Those numbers support the supposition that preventive visits, being more predictable and less urgent, are easier to schedule around work hours. Similarly, previous-year visits also correlate with increases in absenteeism, with the effect of curative visits again being larger. However, after addressing potential endogeneity using the proposed model, neither type of visit appears to significantly relate to missing work. That conclusion also applies to visits that occur in the previous year. Therefore, we conclude that the observed links between medical usage and absenteeism derive from unobserved heterogeneity, rather than direct causal channels.

2 Data

The adopted framework draws inspiration from, and is informed by, our empirical case study. Therefore, we discuss the data before turning to some modeling details.

The data used in this study come from the Medical Expenditure Panel Survey (MEPS), collected and published by the Agency for Healthcare Research and Quality, a unit of the U.S. Department of Health and Human Services. The MEPS enjoys a reputation as the most detailed and complete source of information on individual-level medical spending and usage in the U.S. Of particular importance for this study, the MEPS also includes rich information on individual-level employment-related details. We focus on data from the 2012, 2013, 2014 and 2015 waves, which, at the time of this writing, are the most recent public releases that include details on office-related health care usage. We extract individual-level socioeconomic information from the main "Full Year Consolidated Data" files, focusing on all males, 20 years of age or older, who report working full time (at least 35 hours per week) for the full calendar

year. The main variable of interest is a binary indicator for whether the person missed any work for health-related reasons. (We avoided using MEPS's counted measured of missed days, in part due to changes in the way that variable has been recorded across survey years, and in part due to the fact that those counted measures show very large probably mass at zero and very small mass at positive values. Thus, most of the interesting statistical variation falls along a binary dimension.) We then link such individual-level information to medical usage event-level details from the "Office-Based Medical Provider Visits" files. Crucially, those event-level files record the reason for visiting a medical provider, allowing us to ascertain whether the visit was for diagnosis or treatment, which we label as "curative", or for a general checkup, which we label as "preventive".

Table 1 presents sample means. The most important numbers, appearing near the top of the table, show that, among subjects who had neither curative nor preventive office-based visits, only 23 percent reported missing any work, compared to 55 percent who had both types of visits. The middle two panels, which focus on subjects who had one type of visit but not the other, suggest that curative visits associate with far larger probabilities of missing work. The table also reveals socioeconomic differences across the four medical usage categories. Having any type of office visit appears to positively correlate with age, education, and marital status. Subjects who work for employers that offer paid sick leave, or who work for government organizations, perhaps not surprisingly, appear more likely to report having either type of office-base visit. Also not surprisingly, subjects who report having office-based visits appear to be in worse health.

Our estimation sample also includes 10 occupation dummies, although we do not report their frequencies in the table in order to prevent it from becoming unwieldy. Those occupation

	Curative = NO Preventive = NO n = 10,591	Curative = YES Preventive = NO n = 3,656	Curative = NO Preventive = YES n = 3,199	Curative = YES Preventive = YES n = 3,323
Miss any work	0.23	0.53	0.29	0.55
Age	39.4	42.5	46.3	49.5
Black	0.16	0.10	0.19	0.12
Hispanic	0.40	0.26	0.23	0.17
College degree	0.22	0.33	0.34	0.42
Married	0.54	0.65	0.68	0.71
Family size	3.49	3.15	3.04	2.81
Fair or poor health	0.06	0.10	0.09	0.12
Paid sick leave	0.47	0.62	0.64	0.67
Government job	0.09	0.16	0.17	0.20
Occupation	10 dummies	10 dummies	10 dummies	10 dummies
Easy to phone USC	0.21	0.39	0.41	0.45
Easy to contact USC after hours	0.09	0.15	0.19	0.17

Table 1: Sample means.

dummies are collapsed versions of categories defined by the U.S. Census Bureau. The ten categories are (1) managerial, business, and finance; (2) professional; (3) service; (4) sales; (5) office and administrative support; (6) farming, fishery, and forestry; (7) construction, extraction, and maintenance; (8) production, transportation, and materials moving; (9) military; and (10) unclassified.

Finally, the bottom of the table reports the proportion of subjects who claim that it is "not at all difficult" to contact their usual source of care (USC) by phone, and also the ease with which the person can contact his or her USC after work hours. We argue below, both economically and statistically, that those variables represent appropriate identifying instruments, for two reasons. First, in the U.S., office-based visits with medical providers typically require an appointment, with those appointments usually arranged by phone. Therefore, the ease of

	Curative Yr 1 = NO Preventive Yr 1 = NO n = 3,231	Curative Yr 1 = YES Preventive Yr 1 = NO n = 1,082	Curative Yr 1 = NO Preventive Yr 1 = YES n = 1,012	Curative Yr 1 = YES Preventive Yr 1 = YES n = 1,009
Miss any work Yr 2	0.26	0.42	0.33	0.49

Table 2: Sample means, two-year subsample.

contacting one's USC should predict the likelihood of having an office-based visit, which the numbers in Table 1 seem to confirm. Second, the ease of contacting one's USC should not relate to missing work, aside from indirectly through its link to office visits. Though not formally testable, we offer suggestive evidence of that lack of correlation below.

We also seek to explore whether visits during the previous year affected absenteeism in the subsequent year. To that end, we exploit the (limited) panel structure of the MEPS. With subjects present in the survey for two years, we stack the 2012/2013, 2013/2014, and 2014/2015 panels to create a subsample of the subjects from our parent sample. Samples sizes for this subsample are necessarily smaller than those in Table 1, because we eliminate subjects present for only one year, and we focus only on subjects who remained employed full time throughout the two years. Our goal is to investigate whether visits during year 1 associate with absenteeism in year 2. Table 2 repeats the top portion of Table 1, but focuses on this two-period subsample. Overall, the numbers appear similar to those reported in Table 1, both in terms of magnitude and direction.

3 Trivariate probit model with endogenous binary regressors

The endogeneity issue can be understood in terms of a regression model from which important covariates have been omitted (since not readily available) and hence become part of the model's error term. To control for this form of unmeasured heterogeneity in the empirical context of

this paper, the employed approach builds on a first equation modeling the first endogenous dummy variable, a second equation for the second endogenous dummy variable, and an outcome equation which determines the response variable and that depends on the endogenous binary regressors. The three equations are then connected via a multivariate Gaussian distribution which makes it possible to model the correlations among the equations, hence accounting for unobserved heterogeneity. Below, we provide some details on the model specification employed for our case study, and briefly discuss estimation, inference and the calculation of average treatment effects.

3.1 Model definition

The model can be expressed in terms of latent continuous variables as

$$y_{1i}^* = \mathbf{v}_{1i}^\top \boldsymbol{\gamma}_1 + \varepsilon_{1i}, \tag{1}$$

$$y_{2i}^* = \mathbf{v}_{2i}^\top \boldsymbol{\gamma}_2 + \varepsilon_{2i}, \tag{2}$$

$$y_{3i}^* = \psi_1 y_{1i} + \psi_2 y_{2i} + \mathbf{v}_{3i}^\top \boldsymbol{\gamma}_3 + \varepsilon_{3i}, \tag{3}$$

for $i = 1, \ldots, n$, where n is the sample size, y_{1i}^*, y_{2i}^* and y_{3i}^* are related to the endogenous and outcome variables such that $y_{mi} = 1$ if $y_{mi}^* > 0$ and 0 otherwise, for $m = 1, \ldots, 3$, $\mathbf{v}_{mi}$ contains (e.g., binary, categorical and continuous) covariates, vector $\boldsymbol{\gamma}_m$ represents the effects of the exogenous variables in $\mathbf{v}_{mi}$ with each effect taking any value in $\mathbb{R}$, $\psi_1 \in \mathbb{R}$ and $\psi_2 \in \mathbb{R}$ are the impacts of the endogenous binary variables y_1 and y_2, and the error terms follow the Gaussian distribution $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^\top \overset{iid}{\sim} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \vartheta_{12} & \vartheta_{13} \\ \vartheta_{21} & 1 & \vartheta_{23} \\ \vartheta_{31} & \vartheta_{32} & 1 \end{pmatrix}.$$

The error variances in $\boldsymbol{\Sigma}$ are, as usual, normalized to unity while the off-diagonal elements represent the correlations among the error terms and $\vartheta_{kz} = \vartheta_{zk}$ for $z \neq k$. The correlations can

take values in the range $[-1, 1]$, although some restrictions have to be imposed to guarantee the positive-definiteness of $\boldsymbol{\Sigma}$ (see the next section).

Since the model includes only unidirectional effects (the endogenous variables affect the outcome but the outcome does not affect them), this system can be regarded as a special case of the multivariate recursive model discussed by Wilde (2000). Although the model is theoretically identified (Wilde, 2000), consistent estimation of ψ_1 and ψ_2 is typically more reliably achieved in the presence of instrumental variables (i.e., extra covariates or regressors in the model that are associated with the endogenous variables, are not directly related to the outcome, and are independent of the unobserved confounders) (e.g., Little, 1985). We therefore elect to include such variables in $\mathbf{v}_{1i}$ and $\mathbf{v}_{2i}$, but not in $\mathbf{v}_{3i}$.

Note that the above model could have been defined in a more classical way, where the responses would be y_1, y_2 and y_3 and the covariate vectors $\mathbf{v}_{1i}$, $\mathbf{v}_{2i}$ and $(\mathbf{v}_{3i}^\top, y_1, y_2)$, respectively, with $\mathbf{v}_{1i}$ and $\mathbf{v}_{2i}$ including the instruments. In fact, the employed model can be regarded as a classical trivariate probit model with a particular specification stemming from the specific aims of the case study.

The implementation available through the `GJRM R` package allowed us to explore more general model specifications than those implied by equations (1), (2) and (3). For instance, we considered different link functions for the three equations (see Filippou et al., 2019) as well as smooth functions of the continuous regressors (to estimate the covariate effects flexibly and in a data-driven manner). However, the substantive conclusions of our case study did not change, hence we refrained from discussing a considerably more complex modelling strategy.

3.2 Parameter estimation and inference

Given an observed random sample of responses, y_{1i}, y_{2i} and y_{3i}, and covariates, $\mathbf{v}_{1i}$, $\mathbf{v}_{2i}$ and $\mathbf{v}_{3i}$, for $i = 1, \ldots, n$, the log-likelihood of the model can be written as

$$\begin{aligned}\ell(\boldsymbol{\delta}) \;=\; & \sum_{i=1}^{n} \{y_{1i}y_{2i}y_{3i}\log(p_{111i}) + y_{1i}y_{2i}(1-y_{3i})\log(p_{110i}) + y_{1i}(1-y_{2i})y_{3i}\log(p_{101i}) + \\ & (1-y_{1i})y_{2i}y_{3i}\log(p_{011i}) + (1-y_{1i})(1-y_{2i})(1-y_{3i})\log(p_{000i}) + \\ & (1-y_{1i})(1-y_{2i})y_{3i}\log(p_{001i}) + (1-y_{1i})y_{2i}(1-y_{3i})\log(p_{010i}) + \\ & y_{1i}(1-y_{2i})(1-y_{3i})\log(p_{100i})\}\,, \end{aligned} \tag{4}$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \psi_1, \psi_2, \boldsymbol{\gamma}_3^\top, \vartheta_{12}, \vartheta_{13}, \vartheta_{23})^\top$, and the joint Gaussian probabilities of the three responses conditional on the covariates are denoted as $p_{\bar{e}_1\bar{e}_2\bar{e}_3 i} = \mathbb{P}(y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3)$ with $\bar{e}_m \in \{0, 1\}, \forall m$; see Filippou, Marra, Radice (2017) for technical details. Since the ϑ_{zk} parameters take values in $[-1, 1]$, during fitting, the correlation coefficients are transformed as $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk}) \in \mathbb{R}$. Positive-definiteness of $\boldsymbol{\Sigma}$ is achieved by including range restrictions: for fixed values of ϑ_{13} and ϑ_{23}, $\vartheta_{12} \in \left(\vartheta_{13}\vartheta_{23} - \sqrt{(1-\vartheta_{13}^2)(1-\vartheta_{23}^2)}, \vartheta_{13}\vartheta_{23} + \sqrt{(1-\vartheta_{13}^2)(1-\vartheta_{23}^2)}\right)$. In practice, such a restriction is imposed using the eigenvalue method. Estimation of $\boldsymbol{\delta}$ is achieved using the carefully constructed trust region algorithm in `GJRM`. This is based on first and second order analytical derivatives of trivariate integrals which have been developed in Filippou, Marra and Radice (2017) and Filippou et al. (2019).

Confidence intervals and p-values are constructed using the Bayesian result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}^{-1}\right)$, where the arguments of the multivariate Gaussian denote the estimated parameter vector and the inverse of minus the Hessian matrix.

3.3 Average treatment effects

The effect of the endogenous variables y_{1i} and y_{2i} on the probability that y_{3i} is equal to 1, given covariate information, is of interest. For the case of the effect of y_{1i} on y_{3i}, this can be calculated as

$$P(y_{3i}|y_{1i}=1, y_{2i}, \mathbf{v}_{3i}^{\top}) - P(y_{3i}|y_{1i}=0, y_{2i}, \mathbf{v}_{3i}^{\top}),$$

where $P(y_{3i}=1|y_{1i}=1, y_{2i}, \mathbf{v}_{3i}^{\top}) = \Phi(\eta_{3i}^{(y_{1i}=1)})$, $P(y_{3i}=1|y_{1i}=0, y_{2i}, \mathbf{v}_{3i}^{\top}) = \Phi(\eta_{3i}^{(y_{1i}=0)})$, $\eta_{3i}^{(y_{1i}=\bar{e}_1)}$ denotes the predictor in the outcome equation evaluated at $y_{1i}=\bar{e}_1$ with $\bar{e}_1 \in \{0,1\}$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. Similarly, for the impact of y_{2i} on y_{3i}.

This is known as the causal treatment effect (TE, Angrist, 1996), and it measures the causal difference in outcomes between individuals that receive the treatment ($y_{1i}=1$ or $y_{2i}=1$) and individuals who do not receive it ($y_{1i}=0$ or $y_{2i}=0$). For each individual, only one of the two potential outcomes can be observed. The other outcome is the counter-factual. The average TE (ATE) in a specific sample is given by $1/n \sum_{i=1}^{n} \text{TE}_i$ where TE_i denotes the TE of individual i (e.g., Abadie, 2004).

Intervals for the ATEs of interest are obtained by using the posterior simulation approach detailed in Filippou, Marra and Radice (2017) and Filippou et al. (2019). These have been programmed and made available through `GJRM`.

4 Simulation study

The aim of this section is to assess the empirical effectiveness of the recursive trivariate probit model employed in this paper. In the following, two binary endogenous variables, a binary response, an instrument, two observed confounders, and correlated error terms are denoted as y_1,

Distribution			Mean		% Bias		RMSE	
			Trivariate	Univariate	Trivariate	Univariate	Trivariate	Univariate
$\mathcal{N}$	$n = 1000$	ψ_1	0.48	1.69	3.40	237.69	0.26	1.19
		ψ_2	-0.52	-1.10	3.25	120.07	0.28	0.61
	$n = 4000$	ψ_1	0.49	1.68	2.64	237.28	0.14	1.18
		ψ_2	-0.5	-1.10	0.56	120.32	0.13	0.60
t	$n = 1000$	ψ_1	0.51	1.63	2.70	225.83	0.32	1.13
		ψ_2	-0.56	-1.02	11.21	105.13	0.34	0.54
	$n = 4000$	ψ_1	0.52	1.62	4.61	223.34	0.16	1.12
		ψ_2	-0.55	-1.02	10.20	104.45	0.17	0.52
χ^2	$n = 1000$	ψ_1	0.62	-0.38	24.53	175.34	0.27	0.88
		ψ_2	-0.63	-0.14	25.02	72.33	0.24	0.37
	$n = 4000$	ψ_1	0.64	-0.37	28.62	174.36	0.19	0.87
		ψ_2	-0.65	-0.14	29.42	72.16	0.18	0.36

Table 3: Mean, % bias and root mean squared error (RMSE) of the estimates obtained when fitting the traditional probit (univariate) and the recursive trivariate probit model (trivariate) in the *presence* of an instrument to 1000 data-sets generated using the trivariate normal ($\mathcal{N}$), Student's t and χ^2 distributions for the error terms. The sample sizes considered were 1000 and 4000. True values for ψ_1 and ψ_2 are 0.5 and -0.5, respectively.

y_2, y_3, v_3, v_1, v_2, ε_1, ε_2 and ε_3, respectively. We simulated y_1, y_2 and y_3 using several trivariate distributions for the error terms (i.e., normal, Student's t and χ^2). The error correlations were set to 0.3 between ε_1 and ε_2, 0.6 between ε_1 and ε_3 and -0.2 between ε_2 and ε_3. Variables v_1, v_2 and v_3 were generated from uniform distributions over $[0, 1]$. We considered a more general set up, compared to that used in our case study, in which non-linear covariate effects between y_1 and v_2 and between y_2 and v_2 were present. The coefficients that relate y_1 and y_2 to y_3 were set to 0.5 and -0.5, respectively. The sample sizes were 1000 and 4000, respectively, and each scenario was replicated 1000 times. We also considered the case where no instrument (v_3) is available. Full details on the data generating process, using `R` syntax, are reported in the Appendix. Several other settings were tried out (different values for: ψ_1, ψ_2, correlation coefficients, covariate effects); the substantial findings remained the same.

The findings are summarized in Tables 3 and 4, which compare the results obtained when using the traditional univariate probit (ignoring endogeneity) and the recursive trivariate probit

model, in the presence and absence of an instrument, respectively. Focusing on Table 3, the results confirm that the recursive trivariate probit model is appropriate for correcting for endogeneity and that, as expected, the traditional univariate probit model performs poorly in all scenarios considered. Under misspecification of the model's distribution (Student's t and χ^2), the performance of the trivariate model worsens, although it still yields lower bias and RMSE than those of the univariate probit model, and, importantly, the mean estimates from the trivariate model are not far from the true values. In the absence of an instrument (Table 4), under correct specification of the model's distribution, the parameter estimates of the trivariate model exhibit higher bias and variability as compared to the results of Table 3, although the situation significantly improves as the sample size increases. (For comparison, our case study relies on a sample size of more than 20,000 observations.) Under misspecification, the performance of the trivariate model worsens considerably, although it still outperforms that of the univariate probit model. These results provide evidence of the overall good performance of the trivariate model and that, in practice, using an instrument should make empirical findings more robust especially under misspecification.

Following a reviewer's suggestion, we considered sample sizes smaller than 1000, specifically $n = 200, 500$. The main conclusions in terms of comparisons across models (trivariate and univariate), availability of an instrument (presence or absence) and mis-specification (Student's t and χ^2 distributions) were essentially unchanged. However, as expected, the overall performance deteriorated and, for $n = 200$, the rate of convergence failure was around 35% on average, hence hinting at the fact that the model employed in this paper is not likely to produce sensible estimates at small sample sizes. Note that such small sample sizes are not common in empirical studies such as the one considered in the this paper.

Distribution			Mean		% Bias		RMSE	
			Trivariate	Univariate	Trivariate	Univariate	Trivariate	Univariate
$\mathcal{N}$	$n = 1000$	ψ_1	0.61	1.77	21.62	254.73	0.53	1.28
		ψ_2	-0.66	-1.18	31.70	136.38	0.67	0.69
	$n = 4000$	ψ_1	0.51	1.77	2.15	254.29	0.24	1.27
		ψ_2	-0.51	-1.18	2.05	136.69	0.32	0.68
t	$n = 1000$	ψ_1	0.77	1.67	54.27	234.69	0.69	1.18
		ψ_2	-0.90	-1.07	79.97	114.65	0.86	0.58
	$n = 4000$	ψ_1	0.75	1.66	51.83	232.24	0.46	1.16
		ψ_2	-0.87	-1.06	75.09	113.63	0.58	0.57
χ^2	$n = 1000$	ψ_1	0.57	-0.59	14.89	219.02	0.66	1.10
		ψ_2	-0.44	-0.05	10.87	90.20	0.34	0.46
	$n = 4000$	ψ_1	0.78	-0.59	57.34	218.55	0.35	1.09
		ψ_2	-0.58	-0.05	16.46	89.88	0.15	0.45

Table 4: Mean, % bias and root mean squared error (RMSE) of the estimates obtained when fitting the traditional probit (univariate) and the recursive trivariate probit model (trivariate) in the *absence* of an instrument to 1000 data-sets generated using the trivariate normal ($\mathcal{N}$), Student's t and χ^2 distributions for the error terms. The sample sizes considered were 1000 and 4000. True values for ψ_1 and ψ_2 are 0.5 and -0.5, respectively.

Finally, we also checked the coverage probabilities of the ATEs. We considered the Gaussian error case, with instrument, for $n = 1000, 4000$. In all cases explored, coverage probabilities for the 95% intervals of both ATEs were in the range $(0.93, 0.95)$.

5 Case study

Does visiting a medical provider cause an employee to miss work? The topic has important implications in light of the aforementioned CDC report that economic costs associated with worker absenteeism total in the hundreds of billions annually. However, unmeasured attributes might simultaneously associate with medical care usage *and* worker absenteeism. Furthermore, curative and preventive visits likely exert disparate influences on absenteeism, with those disparate effects depending, in part, on the timing of visits. We explore this topic using the proposed trivariate setup.

This section first explores the validity of the instrument. It then skips straight to the

paper's main punchline. Finally, it presents and discusses some of the more nuanced findings from our trivariate model.

5.1 Instrument validity

Although our model is technically identified via nonlinear functional forms, especially, as suggested by our simulations, in settings such as this one that use large sample sizes, more robust identification of our model hinges on our two instruments – (1) the ease of contacting one's USC by phone and (2) the ease of contacting one's USC after work hours – significantly affecting the likelihood of office visits, while also *not* affecting the likelihood of missing work, other than indirectly through its effect on having office visits.

To investigate the first condition, Table 5 reports probit estimates for each type of visit. Estimates for the instruments, appearing near the top of the table, reveal that being able to easily phone one's USC correlates significantly with increased probabilities of having a curative visit and a preventive visit. Despite the after hours measure showing weaker influence, the instruments appear to significantly and non-trivially affect the likelihood of each type of office visit.

As for the second condition, we sketch a simple economic model, building from a canonical model of labor supply (Varian, 1992, pp. 145-146). Suppose a person maximizes utility $U(c, L)$ where c denotes consumption, and L represents leisure (i.e., time not spent working). The person has $\overline{L}(e)$ total hours in the day in which to work, so that $\overline{L}(e) - L$ represents labor supply. Drawing from Gould (2004), the term e denotes "effort" expended obtaining health care, such that $\frac{\partial \overline{L}(e)}{\partial e} < 0$. For example, difficulty contacting one's usual source of care by phone would require expending more effort, which, in turn, reduces available time in the day. The consumer chooses c and L to maximize his utility function subject to his budget constraint

	Curative visit		Preventive visit	
	Coeff.	St. err.	Coeff.	St. err.
Easy to phone USC	0.305	0.022	0.362	0.022
Easy to contact USC after hours	−0.032	0.030	0.033	0.030
Age	0.014	0.001	0.027	0.001
Black	−0.422	0.029	−0.043	0.028
Hispanic	−0.282	0.023	−0.225	0.024
College degree	0.072	0.025	0.130	0.026
Married	0.170	0.023	0.184	0.024
Family size	−0.076	0.007	−0.088	0.007
Fair or poor health	0.405	0.033	0.289	0.034
Paid sick leave	0.192	0.021	0.215	0.022
Government job	0.171	0.029	0.120	0.030
Intercept	−0.969	0.052	−1.711	0.055
Occupation dummies	yes		yes	

Table 5: Coefficients from univariate probit models for the curative and preventive visit variables.

$pc + wL = w\overline{L}(e)$ where p is the per-unit price of consumption, and w is the wage rate. The budget constraint implies that the person "sells" his endowment of available time at price w and then buys some back in the form of leisure. The utility maximization problem requires that the marginal rate of substitution between c and L equates to the ratio in prices,

$$\frac{\frac{\partial U}{\partial c}}{\frac{\partial U}{\partial L}} = \frac{p}{w},$$

which depends on e only indirectly via its influence on $\overline{L}$. Consequently, from a theoretical perspective, the term e appears plausibly excludable from labor/leisure decisions.

To afford protection against potential violations of this second condition, we include a set of detailed control variables, both person- and job-specific. For person-specific, we include age, race/ethnicity, education, marital status and family size. Moreover, to account for the possibility that people with health problems might select providers based on ease of phone access, we include a indicator of whether the person self-reports being in "fair or poor" health. For job-specific, we include dummy indicators for whether the job offers paid sick leave, and

	Ignoring endogeneity		Trivariate model	
	Estimate	95% interval	Estimate	95% interval
Curative visit	0.263	(0.251, 0.278)	0.106	(−0.061, 0.289)
Preventive visit	0.045	(0.030, 0.059)	0.044	(−0.066, 0.169)

Table 6: Treatment effects of office visits on missing work.

whether the employer is some sort of state or federal government agency. We also include a set of occupation dummies. After including those predictors, it seems plausible that any remaining unobserved confounders exert little, if any, impact on a person's propensity to miss work for medical-related reasons.

5.2 Main finding

Skipping ahead to our main punchline, Table 6 shows average treatment effects of office-based visits on the probability of missing work. The left-hand panel, under the header "ignoring endogeneity", shows estimates derived from simple probit models that do not correct for potential endogeneity of office visits. Those estimates suggest that having a curative visit increases the probability of missing work by about 26 percentage points. Compared to the sample mean of missing work (0.34), that 26 point increase corresponds to an approximate 76 percent increase in the probability of missing work. Meanwhile, having a preventive visit leads to a more modest 4.5 percentage point increase (13 percent relative to the mean) of missing work. However, those estimates should be interpreted with caution, since unobserved attributes might simultaneously correlate with medical care usage as well as one's propensity to miss work.

Shown in the right-hand panel of Table 6, our trivariate model, which helps to deal with such endogeneity bias, finds that neither type of office visit appears to significantly affect missed work. Instead, the observed link between medical care usage and missed work seems to derive almost entirely from unobserved heterogeneity that simultaneously drives both.

	Curative visit		Preventive visit		Miss any work	
	Coeff.	St. err.	Coeff.	St. err.	Coeff.	St. err.
Easy to phone USC	0.302	0.022	0.365	0.022	—	—
Easy to contact USC after hours	−0.025	0.029	0.036	0.030	—	—
Curative visit	—	—	—	—	0.298	0.243
Preventive visit	—	—	—	—	0.127	0.191
Age	0.014	0.001	0.027	0.001	− 0.005	0.001
Black	−0.425	0.029	−0.043	0.028	−0.193	0.041
Hispanic	−0.285	0.023	− 0.228	0.024	−0.230	0.027
College	0.073	0.025	0.131	0.026	−0.096	0.026
Married	0.172	0.023	0.187	0.024	−0.062	0.025
Family size	−0.077	0.007	−0.088	0.007	−0.041	0.007
Fair or poor health	0.408	0.033	0.290	0.034	0.450	0.037
Paid sick leave	0.193	0.021	0.218	0.022	0.276	0.022
Government job	0.170	0.029	0.119	0.030	0.271	0.030
Occupation dummies	yes		yes		yes	
Intercept	−0.964	0.052	−1.712	0.055	−0.346	0.073

Table 7: Main estimation results from trivariate model.

5.3 Full presentation of estimates

Table 7 presents the full estimation results from the trivariate model. The top row shows that being able to easily phone one's USC increases the propensity of having both types of visits. As for the other control variables, blacks and Hispanics report fewer curative visits than their nonblack/nonHispanic counterparts, while Hispanics also report fewer preventive visits. Marriage positively associates with visits, while family size has a negative impact. Being in "fair or poor" health, not surprisingly, associates with larger probabilities of having both types of visits. Working for an employer than offers paid sick leave increases the likelihood of both types of visits, as does working for a government agency. The right hand panel of Table 7 reports estimates from the outcome equation. Blacks and Hispanics are less likely to miss work than their counterparts, marriage and family size negatively associate with missing work, while health problems positively correlate with missing work. Paid sick leave and government employment have positive effects.

	Estimate	95% interval
$\theta_{\text{curative visit, preventive visit}}$	0.264	(0.241, 0.285)
$\theta_{\text{curative visit, missed work}}$	0.257	(0.005, 0.485)
$\theta_{\text{preventive visit, missed work}}$	0.049	(−0.136, 0.222)

Table 8: Estimated correlations and related 95% intervals.

Table 8 shows estimates of the correlations. First, the link between curative and preventive visits is positive, and precisely estimated, at 0.264. The interpretation is that unobserved attributes that increase a person's likelihood of having a curative visit also, perhaps not surprisingly, increase his chances of having a preventive visit. The next two rows reveal disparate patterns for endogeneity bias for the two types of visits. The positive dependence term in the second row of the table suggests that unobserved attributes that increase the likelihood of obtaining curative care also increase a person's chances of missing work. Such a pattern would be evident if unobserved (to the researcher) health problems simultaneously increase the probabilities of obtaining curative services and missing work. In the univariate treatment effect reported in Table 6 that ignores endogeneity, that positive dependence becomes absorbed into the treatment effect, creating an upward bias. The third row, however, indicates much smaller endogeneity bias with respect to preventive care. Overall, the finding of such stark differences between those latter two dependence terms offers further evidence that curative and preventive services are different types of care, with distinctly different links to work absenteeism.

5.4 Visits in the previous year

Would a curative visit during the previous year, assuming treatment was effective, have any effect on absenteeism during the subsequent year? And would a preventive visit during the previous year actually *reduce* absenteeism during the subsequent year? To investigate those questions, we focus on the aforementioned subset of individuals who remain in the survey

	Ignoring endogeneity		Trivariate model	
	Estimate	95% interval	Estimate	95% interval
Curative visit	0.127	(0.104, 0.149)	0.069	(−0.264, 0.470)
Preventive visit	0.047	(0.019, 0.070)	−0.006	(−0.281, 0.319)

Table 9: Treatment effects of office visits on missing work, two-year subsample.

for two years, and who remain employed full time during those two years. We use the same trivariate probit setup to estimate the effects of visits during year 1 on absenteeism during year 2.

The left-hand panel of Table 9 shows estimates from simple probits that ignore endogeneity. Similar to estimates reported in Table 6, those effects are positive and precisely estimated. Yet, also similar to Table 6, accounting for endogeneity renders those estimates indistiguishable from zero. Table 10 presents full estimates of results, with estimates appearing similar to those reported in Table 7. Furthermore, dependence numbers reported in Table 11 also appear similar to those reported in Table 8.

Thus, the implication is that, whether the visit occurred in the previous year or the current one, both types of visits *correlate* with increased absenteeism, with curative visits showing larger effects. However, those effects appear to derive from unobserved heterogeneity. Once that heterogeneity is addressed, neither type of visit appears to directly *cause* absenteeism.

6 Conclusion

This paper explored whether visiting a medical provider causes an employee to miss work. Ignoring endogeneity, we find large, and statistically significant, effects. Specifically, having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlates with a smaller 13 percent increase in the probability of missing work. Those observed links between health events and absenteeism for the most

	Curative visit		Preventive visit		Miss any work	
	Coeff.	St. err.	Coeff.	St. err.	Coeff.	St. err.
Easy to phone USC	0.328	0.039	0.359	0.040	—	—
Easy to contact USC after hours	−0.117	0.054	−0.013	0.054	—	—
Curative visit	—	—	—	—	0.195	0.640
Preventive visit	—	—	—	—	−0.016	0.527
Age	0.012	0.002	0.028	0.002	0.0001	0.003
Black	−0.337	0.052	−0.046	0.051	−0.306	0.084
Hispanic	−0.256	0.043	− 0.273	0.045	−0.233	0.048
College	0.093	0.046	0.145	0.047	−0.140	0.047
Married	0.201	0.042	0.209	0.044	−0.085	0.046
Family size	−0.089	0.012	−0.086	0.013	−0.043	0.014
Fair or poor health	0.335	0.059	0.234	0.061	0.249	0.073
Paid sick leave	0.150	0.038	0.192	0.040	0.355	0.040
Government job	0.181	0.052	0.067	0.053	0.275	0.059
Occupation dummies	yes		yes		yes	
Intercept	−0.901	0.097	−1.798	0.103	−0.400	0.193

Table 10: Main estimation results from trivariate model, two-year subsample.

	Estimate	95% interval
$\theta_{\text{curative visit, preventive visit}}$	0.272	(0.230, 0.314)
$\theta_{\text{curative visit, missed work}}$	0.113	(−0.484, 0.743)
$\theta_{\text{preventive visit, missed work}}$	0.110	(−0.382, 0.536)

Table 11: Estimated correlations and related 95% intervals, two-year subsample.

part corroborate what appears in the existing literature. However, after addressing potential endogeneity, neither type of visit appears to significantly relate to missing work. Thus, we do not detect *causal* effects of health events on absenteeism, although we cannot infer what that finding might imply about effects on productivity in general. Those findings come from a trivariate model, where one equation models an outcome of interest, which depends, among other things, on two endogeneous variables. The other two equations model the two endogenous explanatory variables. The three equations are connected via a multivariate Gaussian distribution, which makes it possible to model the correlations among the equations, hence accounting for unobserved heterogeneity.

The employed model should prove useful for empirical problems with two endogeneous explanatory variables. For example, husband and wife employment decisions might endogenously affect childcare decisions. Or the votes of two senators from the same state might endogenously affect economic conditions in that state. The model proposed in this paper offers an intuitive, and easy-to-estimate, route to explore such topics.

Acknowledgments

The first two authors were supported by the Engineering and Physical Sciences Research Council [EP/T033061/1].

References

Abadie, A., Drukker, D., Herr, J. L., Imbens, G. W. (2004). "Implementing matching estimators for average treatment effects in stata", Stata Journal, 4, 290-311.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). "Identification of causal effects using instrumental variables", Journal of the American Statistical Association, 91(434), 444-

455.

Ashford, J., Sowden, R. (1970). "Multi-variate probit analysis", Biometrics, 26, 535-546.

Cappellari, L., Jenkins, S. P. (2003). "Multivariate probit regression using simulated maximum likelihood", Stata Journal, 3, 278-294.

Chib, S., Greenberg, E. (1998). "Analysis of multivariate probit models", Biometrika, 85, 347-361.

Chib, S., Greenberg, E. (2007). "Semiparametric Modeling and Estimation of Instrumental Variable Models", Journal of Computational and Graphical Statistics, 16(1), 86-114.

Filippou, P., Kneib, T., Marra, G., Radice, R. (2019). "A Trivariate Additive Regression Model with Arbitrary Link Functions and Varying Correlation Matrix", Journal of Statistical Planning and Inference, 199, 236-248.

Filippou, P., Marra, G., Radice, R. (2017). "Penalized Likelihood Estimation of a Trivariate Additive Probit Model", Biostatistics, 18(3), 569-585.

Gould, E. (2004). "Decomposing the Effects of Children's Health on Mother's Labor Supply: Is it Time or Money?" Health Economics, 13, 525-541.

Henningsen, A. (2021). "mvProbit: Multivariate Probit Models." R package version 0.1-10.

Hansen, L. (1982). "Large Sample Properties of Generalized Method of Moments Estimators." Econometrica, 50, 1029-1054.

Little, R.J.A. (1985). "A note about models for selectivity bias", Econometrica, 53(6), 1469-1474.

Marra, G., Radice, R. (2011). "Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity", Canadian Journal of Statistics, 39, 259-279.

Marra, G., Radice, R. (2022). "GJRM: Generalized Joint Regression Modeling", R package version 0.2-6, URL https://cran.r-project.org/package=GJRM.

Nicholson, S., Pauly, M., Polsky, D., et al. (2006). "Measuring the effects of workloss on productivity with team production", Health Economics, 15, 111-123.

Pauly, M., Nicholson, S., Polsky, D., et al. (2008). "Valuing reductions in on-the-job illness: 'presenteeism' from managerial and economic perspectives", Health Economics, 17, 469-485.

Schultz, A., Chen, C., Edington, D. (2009). "The cost and impact of health conditions on presenteeism to employers a review of the literature", Pharmacoeconomics, 27, 365-378.

Strömberg, C., Aboagye, E., Hagberg, J., Bergström, G., Lohela-Karlsson, M. (2017). "Estimating the Effect and Economic Impact of Absenteeism, Presenteeism, and Work Environment–Related Problems on Reductions in Productivity from a Managerial Perspective", Value in Health, 20, 1058-1064.

Varian, H. (1992). Microeconomic Analysis: Third Edition. W.W. Norton and Company: New York.

Wilde, J. (2000). "Identification of multiple equation probit models with endogenous dummy regressors", Economics Letters, 69(3), 309-312.

Zhang, W., Bansback, N., Anis, A. (2011). "Measuring and valuing productivity loss due to poor health: a critical review", Social Science and Medicine, 72, 185-192.

Zhang, W., Sun, H., Woodcock, S., Anis, A. (2015). "Illness related wage and productivity losses: valuing 'presenteeism", Social Science and Medicine, 147, 62-71.

Appendix

Code to simulate data and estimate trivariate binary models.

```
# Load library, set seed and sample size
library(GJRM)
set.seed(1)
n <- 1000

# Set correlation matrix of error terms
Sigma      <- matrix(0.6, 3, 3); diag(Sigma) <- 1
Sigma[1,2] <- Sigma[2,1] <-  0.3
Sigma[2,3] <- Sigma[3,2] <- -0.2

# Set correlation matrix of covariates
SigmaC <- matrix(0.5, 3, 3); diag(SigmaC) <- 1

# Function to generate non-linear covariate effects
f1 <- function(x) cos(pi*2*x) + sin(pi*x)
f2 <- function(x) x + exp(-30*(x - 0.5)^2)

# Generate errors from trivariate normal distribution
e <- rMVN(n, rep(0,3), Sigma)

# Generate covariates from multivariate uniform distribution
cov <- rMVN(n, rep(0,3), SigmaC)
cov <- pnorm(cov)
v1  <- cov[, 1]
v2  <- cov[, 2]
v3  <- cov[, 3]

# Generate the endogenous and response variables
y1 <- ifelse(-1    +                    2*v1 - f1(v2) +      v3 + e[,1] > 0, 1, 0)
y2 <- ifelse( 0.25 -                 1.25*v1 + f2(v2) - 1.25*v3 + e[,2] > 0, 1, 0)
y3 <- ifelse(-0.75 + 0.5*y1 - 0.5*y2+ 0.25*v1 +    v2            + e[,3] > 0, 1, 0)

# Construct a dataframe
dataSim <- data.frame(y1, y2, y3, v1, v2, v3)

# Fit the recursive trivariate probit model
f.l <- list(y1 ~           v1 + s(v2) + v3,
            y2 ~           v1 + s(v2) + v3,
            y3 ~ y1 + y2 + v1 +   v2       )
out  <- gjrm(f.l, data = dataSim, Model = "T",
             margins = c("probit", "probit", "probit"))
conv.check(out)    # convergence check
```

```
summary(out)       # summary output
plot(out, eq = 1)  # plotting function
plot(out, eq = 2)
```

For the case without instrument, `v3` is dropped from the equations above.

To allow the error terms to be Student's t (with two degrees of freedom) or χ^2 (with two degrees of freedom) distributed, respectively, the above `R` code was be easily modified by replacing

```
e <- rMVN(n, rep(0,3), Sigma)
```

with

```
library(mvtnorm)
e <- rmvt(n, rep(0,3), sigma = Sigma, df = 2)
```

or with

```
library(copula)
norm.cop <- ellipCopula("normal", param = c(0.3, -0.6, 0.2),
                        dim = 3, dispstr = "un")
myMvd <- mvdc(copula = norm.cop, margins = c("chisq", "chisq", "chisq"),
              paramMargins = list(list(df = 2), list(df = 2),  list(df = 2)) )
e <- rMvdc(mvdc = myMvd, n = n)
```