Epidemiology Publish Ahead of Print

DOI: 10.1097/EDE.0000000000000425

Physician's prescribing preference as an instrumental variable: exploring assumptions using survey data.

Anna G.C. Boef[1], Saskia le Cessie[1,2], Olaf M. Dekkers[1,3,4], Peter Frey[5], Patricia M. Kearney[6], Ngaire Kerse[7], Christian D. Mallen[8], Vera J.C. McCarthy[6], Simon P. Mooijaart[9,10], Christiane Muth[11], Nicolas Rodondi[12], Thomas Rosemann[13], Audrey Russell[6], Henk Schers[14], Vanessa Virgini[12, 15], Margot W.M. de Waal[16], Alex Warner[17], Jacobijn Gussekloo[16], Wendy P.J. den Elzen[16]

[1]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands

[2]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands

[3]Department of Endocrinology and Metabolic Diseases, Leiden University Medical Centre, Leiden, The Netherlands

[4]Department of Clinical Epidemiology, Aarhus Medical Centre, Aarhus, Denmark

[5]Bern Institute of General Practice, University of Bern, Bern, Switzerland

[6]Department of Epidemiology and Public Health, University College Cork, Cork, Ireland

[7]Department of General Practice and Primary Health Care, University of Auckland, Auckland, New Zealand

[8]Arthritis Research UK Primary Care Centre, Keele University, Keele, Staffordshire, United Kingdom

[9]Institute for Evidence-based Medicine in Old age (IEMO), Leiden, The Netherlands

[10]Department of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands

[11]Institute of General Practice, Johann Wolfgang Goethe University, Frankfurt, Germany

[12] Department of General Internal Medicine, Inselspital, Bern University Hospital, Bern, Switzerland

[13]Institute of General Practice and Health Services Research, University of Zürich, Zürich, Switzerland

[14]Department of Primary and Community Care, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands

[15]Department of Internal Medicine, University Hospital of Zürich, Zürich, Switzerland

[16] Department of Public Health and Primary Care, Leiden University Medical Centre, Leiden, The Netherlands

[17] Research department of Primary Care and Population health, University College London, United Kingdom

Corresponding author

A.G.C. Boef, Department of Clinical Epidemiology, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden. T: +31 71 5264037, F: +31 715266994, E-mail: a.g.c.boef@lumc.nl.

Running head

Instrumental variable assumptions in survey data.

Conflicts of interest and sources of funding

The authors declare no conflicts of interest. This work was supported by the Netherlands Organisation for Health Research and Development (ZonMw, grant number 152002040). NR is supported by a grant from the Swiss National Science Foundation (SNSF 320030-150025). CDM is funded by the National Institute for Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care West Midlands, the NIHR School for Primary Care Research and a NIHR Research Professorship in General Practice (NIHR-RP-2014-04-026). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

ABSTRACT

Background: Physician's prescribing preference is increasingly used as an instrumental variable in studies of therapeutic effects. However, differences in prescribing patterns among physicians may reflect differences in preferences or in case-mix. Furthermore, there is debate regarding the possible assumptions for point estimation using physician's preference as an instrument.

Methods: A survey was sent to general practitioners (GPs) in The Netherlands, the United Kingdom, New Zealand, Ireland, Switzerland and Germany, asking whether they would prescribe levothyroxine to eight fictitious patients with subclinical hypothyroidism. We investigated (1) whether variation in physician's preference was observable and to what extent it was explained by characteristics of GPs and their patient populations and (2) whether the data were compatible with deterministic and stochastic monotonicity assumptions.

Results: Levothyroxine prescriptions varied substantially amongst the 526 responding GPs. Between-GP variance in levothyroxine prescriptions (logit scale) was 9.9 (95% CI 8.0;12) in the initial mixed-effects logistic model, 8.3 (6.7;10) after adding a fixed effect for country and 8.2 (6.6;10)after adding GP characteristics. The occurring prescription patterns falsified the deterministic monotonicity assumption. All cases in all countries were more likely to receive levothyroxine if a different case of the same GP received levothyroxine, which is compatible with the stochastic monotonicity assumption. The data were incompatible with this assumption for a different definition of the instrument.

Conclusions: Our study supports the existence of physician's preference as a determinant in treatment decisions. Deterministic monotonicity will generally not be plausible for physician's preference as an instrument. Depending on the definition of the instrument, stochastic monotonicity may be plausible.

INTRODUCTION

Instrumental variable (IV) analysis is increasingly used in observational studies of therapeutic effects, with the aim of circumventing confounding by indication. This method requires a variable (the instrument) that meets the following conditions: (1) is associated with treatment, (2) does not affect the outcome other than through treatment (exclusion restriction) and (3) does not share a common cause with the outcome (independence assumption).[1;2] One such instrument is physician's prescribing preference, which exploits the notion that prescribing by a medical doctor is influenced not only by prognostic characteristics of the patient, but also by a general preference of the doctor for some type of therapy when different treatment options are available.

Because underlying preference cannot be observed, physician's preference-based IV studies use an estimate of physician's preference based on prescribing behaviour. The question remains, however, whether differences in prescribing behaviour between physicians truly reflect differences in preference rather than just differences in their patient populations. Furthermore, the three main IV conditions described above are only sufficient for the estimation of bounds of a treatment effect.[3] To obtain a point estimate an additional (fourth) assumption is required. The assumption of no heterogeneity of treatment effects, under which the average treatment effect in the population can be estimated, is often implausible.[3] A frequently used alternative is the monotonicity assumption, first

described by Imbens and Angrist.[4] According to the original (deterministic) monotonicity assumption, the instrument may only be related to treatment monotonically in one direction for all subjects.[2;4-7] A less strict, stochastic version of the monotonicity assumption has been proposed, as we will explain later.[5-7]

The notion that physician's underlying prescribing preference affects prescribing behaviour cannot be proven in IV study data (at the most, the assumption that physician's estimated prescribing preference is unrelated to characteristics of the physician's patient population can be explored to some extent). Furthermore, the deterministic monotonicity assumption is generally not verifiable within IV study data and the validity of the stochastic monotonicity assumption can only be explored to some extent. Swanson et al. recently proposed using a study design in the form of a survey, asking physicians what their treatment decision would be for the same set of cases, to assess the monotonicity assumption empirically.[2] Here we perform such a study, using data from a survey originally performed with the aim of establishing differences in treatment strategies of general practitioners (GPs) for subclinical hypothyroidism by country and by patient characteristics.[8] These data were therefore not primarily intended for our current study, but can nevertheless provide a valuable insight into the plausibility of the different monotonicity assumptions. Our aims are twofold, (1) to establish whether variation in

physician's preference regarding treatment of subclinical hypothyroidism is observable when GPs are presented with the same set of patients and to what extent this variation is explained by characteristics of the GPs and (2) to establish to what extent the data are compatible with the deterministic and stochastic monotonicity assumptions.

METHODS

Study data

The survey procedures have been described in detail elsewhere.[8] An online survey was e-mailed to 2710 GPs in The Netherlands, Germany, England, Ireland, Switzerland and New Zealand. It contained eight fictitious cases of women with subclinical hypothyroidism. All cases had a normal BMI, non-specific complaints resulting in fatigue and a normal free thyroxine level. Cases varied in age (70 years/ 85 years), vitality status (vita /vulnerable) and thyroid stimulating hormone (TSH) (6 mU/L/15 mU/L), (Table 1). For each case, GPs were asked if they would start treatment, and, if so, what levothyroxine starting dose they would choose. For the purposes of this study, we only consider the responses on whether treatment would be started. Furthermore, GPs were asked questions about their gender, years of experience as a GP, the percentage of elderly patients registered in their practice, the time since the last diagnosis of subclinical hypothyroidism and the time since last starting levothyroxine treatment in a patient with subclinical hypothyroidism. For the full survey, we

refer to Appendix 2 of Den Elzen et al. , which reports the study for which the survey was originally performed.[8] The survey study was exempt from ethical review in in the Netherlands, Germany, England, Switzerland, and New Zealand, as it discussed only fictional patients. In Ireland, the Clinical Research Ethics Committee of the Cork Teaching Hospital approved the survey.[8]

Possible assumptions for point estimation

Deterministic monotonicity

For a dichotomous instrument the deterministic monotonicity assumption is usually defined as the absence of 'defiers'.[1;2;6;9] The IV analysis then estimates a local average treatment effect among the 'compliers'.[1;4] These 'compliers' or 'marginal patients' are those patients who would receive treatment at the 'encouraging' value of the instrument (e.g. preference for treatment), but not at the 'non-encouraging' value of the instrument (e.g. preference for no treatment).[1;5;9;10] As discussed by Swanson et al. and Small et al, for physician's preference as an IV, the compliance class (whether the patient is a complier, defier, always taker or never taker) is generally not well defined.[2;6]

Hernán and Robins have formulated the deterministic monotonicity assumption for physician's preference as a continuous instrument.[3] This would translate to the example of subclinical hypothyroidism as follows: if physician A would treat a certain patient with subclinical hypothyroidism with levothyroxine, then

all physicians with a preference greater than or equal to the preference of physician A should treat that patient with levothyroxine. It is this assumption which we will assess for our survey data. It would correspond to global monotonicity as described by Swanson et al. [2] (Local monotonicity was also described by Swanson et al: for this somewhat more relaxed version of the assumption monotonicity must hold for specific pairs of physicians.[2]) For continuous instruments, the local average treatment effect is a weighted average of treatment effects in multiple subgroups of patients (e.g. subgroups of patients who would receive levothyroxine from physicians with a certain preference but not from physicians with a lower preference).[1;3]

Stochastic monotonicity

The alternative to deterministic monotonicity proposed is the stochastic monotonicity assumption, which states that the instrument should be related to treatment monotonically across subjects within strata of a sufficient set of measured and unmeasured common causes of treatment and the outcome.[6]

If we view the cases in our survey not as individual cases but as strata of patients with the same relevant characteristics, the stochastic monotonicity

assumption requires GPs' preference to be related to treatment monotonically in one direction across patients in each of these strata. This means that the probability of levothyroxine treatment for patients treated by GPs with preference A should be at least as high as for patients treated by GPs with a lower preference, within all strata of patients.

Under the stochastic monotonicity assumption, the effect estimated is a weighted average of treatment effects in the different strata of patients, with more weight given to those strata in which the instrument is strongest.[5;7] Small et al. have named this the strength-of-IV weighted average treatment effect (SIVWATE).[6] We point out that, in their identification framework for the SIVWATE and local average treatment effect, Small et al. formulate the three main IV assumptions differently to how we formulated these assumptions in our introduction.[6]

Analysis

Variation in preference for levothyroxine and its determinants

For each GP who completed all survey questions, we calculated the total number of cases treated with levothyroxine, as a measure of the GP's relative preference for treatment with levothyroxine in subclinical hypothyroidism.

To investigate the effect of GP characteristics on their tendency to prescribe levothyroxine, we used mixed-effects logistic regression. All cases completed

by the GPs were included, with treatment with levothyroxine (no/yes) as the outcome. We ran the following (pre-specified) models:

Model 1: A random effect for GP and fixed effects for characteristics of the case (age 70 or 85, TSH 6 or 15 mU/L, vital or vulnerable disposition).

Model 2: Model 1 plus a fixed effect for country.

Model 3: Model 2 plus a fixed effect for GP gender and years of experience (<5, 5-10, 11-15, 16-20, 21-25, >25 years).

Model 4: Model 3 plus a fixed effect for percentage of patients in the GP's practice aged ≥65 years (<10%, 10-20%, 20-30%, >30%) and time since last diagnosis of subclinical hypothyroidism (<1 wk, 1 wk-1 mth, 1 mth-1 yr, 1-3 yrs, >3 yrs).

The parameter of interest was the variance of the random effect of the GP ("between-GP variance in preference"), which is calculated on a log odds scale. The interest lies in whether this variance decreases as country and characteristics of the GP are added to the model.

Deterministic monotonicity assumption

To investigate the monotonicity assumption we made a matrix plot[11] for each country, with cases 1 to 8 on the X-axis and the GPs, ordered from highest to lowest preference, on the Y-axis, the colour of each cell indicating whether levothyroxine was prescribed. This was used to visually examine whether the deterministic monotonicity assumption holds. GPs who did not complete the

survey were not included in these plots. eFigure 1 shows a matrix plot with the pattern expected if deterministic monotonicity holds completely: physicians with a certain preference always prescribe levothyroxine to those cases for which physicians with the same or a lower preference prescribe levothyroxine. (From these plots, which show the complete data pattern, it is also possible to derive whether deterministic monotonicity could hold for specific instruments such as treatment of the previous patient of the same GP.)

Stochastic monotonicity assumption

The exact formulation of the stochastic monotonicity assumption depends on the definition of the instrument. Because Small et al. discuss the stochastic monotonicity assumption in the context of a binary instrument, using treatment of the previous patient as an example, and because treatment of the previous patient is a frequently used physician's preference-based instrument, we evaluated whether stochastic monotonicity could hold for this instrument. Because all GPs were presented with all cases in the same order, we cannot use the true previous case as instrument. We therefore considered each other case as a potential previous patient -- i.e. for each case there were seven potential previous patients per GP. We denote the potential previous patient as the 'other patient'. Each possible index patient–other patient combination was classified according to the treatment of both patients and summed across GPs to a total per case (per country). For each case we calculated the probability of levothyroxine

treatment if the other patient received levothyroxine and if the other patient did not receive levothyroxine.

As a sensitivity analysis, we also assessed the stochastic monotonicity assumption for the proportion of all other cases the same GP decided to treat (although we note that Small et al. only discussed the stochastic monotonicity assumption with respect to a dichotomous instrument)[6]. We performed this analysis for the two countries with the largest number of responding GPs (The Netherlands and Switzerland).

Missing data

There was a technical problem in the electronic questionnaire sent to the Dutch GPs, resulting in 16 missing answers for case 6. Missing answers due to this technical problem were imputed, using logistic regression (10 imputations) with country, the answers for all other cases and characteristics of the GP as predictors.

Analyses were performed using Stata 12 (College Station, TX: StataCorp LP. 2011).

RESULTS

A total of 526 GPs from eight countries responded to the survey. eTable 1 lists the response rates per country. The overall response rate was 19% (526/2710) and ranged from 4% (New Zealand) to 41% (The Netherlands). The number of responding GPs ranged from 21 from Ireland to 262 from Switzerland. Table 2 shows the characteristics of the GPs. Of the 526 respondents, 468 (89%) answered all questions and 71% were male. The years of experience ranged from <5 years (8%) to >25 years (29%). Seventy percent of responding GPs had ≥20% patients aged 65 years and over in their practice and the vast majority (91%) had diagnosed a patient with subclinical hypothyroidism within the last year.

Variation in number of levothyroxine prescriptions

Figure 1 displays the distribution per country of the total number of cases for which the GP decided to start levothyroxine. There was substantial variation in this total within each country. The most frequent number of levothyroxine prescriptions was 4 for the UK, New Zealand, Ireland and Switzerland, 0 for The Netherlands and 8 for Germany.

Association between GP characteristics and treatment preference

Table 3 displays results of the mixed-effects logistic regression used to investigate the effect of GP characteristics on levothyroxine prescription. Country explained some of the variance in levothyroxine prescription between

GPs, as shown by the reduction in between-GP variance from 9.9 (95% CI 8.0;12) to 8.3 (6.7;10) after adding a fixed effect for country. Adding GP characteristics (Model 3) resulted in a very small reduction in between-GP variance in treatment to 8.2 (6.6;10). Adding time since last subclinical hypothyroidism diagnosis and the proportion of patients aged 65 years and over (Model 4) resulted in a similarly small reduction. There was therefore still substantial variation in levothyroxine prescription among GPs after adjusting for all available patient and doctor characteristics.

Deterministic monotonicity assumption

Figure 2 shows matrix plots per country of the treatment decisions for each case by each GP. GPs are ordered from highest (eight cases treated) to lowest preference (0 cases treated). The prescription patterns of the UK (Figure 2B) only showed a single violation of deterministic monotonicity: the GP who prescribed levothyroxine to five cases treated case 2 while the GP who prescribed levothyroxine to six cases did not treat case 2. There were more violations of deterministic monotonicity in the other countries. Treating all cases with a TSH of 15 mU/L was a common pattern in the UK, the Netherlands, New Zealand, Switzerland, and Ireland. For example, 75 of 89 GPs who treated four cases in Switzerland decided to initiate levothyroxine in cases 3, 4, 7, and 8. In both the Netherlands and Switzerland, most GPs with a lower preference treated (one or more) cases with a high TSH only and most

GPs with a higher preference treated at least the high TSH cases. However, there was not a consistent pattern regarding the 5th, 6th,or 7th case treated, or the 1st, 2nd,or 3rd case treated within those with a TSH of 15 mU/L. Prescribing patterns in Germany differed from those in other countries: many GPs (25 of 55) treated all cases with levothyroxine, and for the other GPs the prescribing patterns were less consistent.

Stochastic monotonicity assumption

Table 4 displays the probability of levothyroxine prescription per case, dependent on treatment of a different patient of the GP. The probability of levothyroxine prescription was higher if the other patient was prescribed levothyroxine for nearly all cases in all countries. Exceptions were case 1 in the UK and in New Zealand, for whom treatment probability did not differ depending on the other patient's treatment. Importantly, there were no cases for whom the probability of levothyroxine was higher if the other patient did not receive levothyroxine, i.e. the instrument was related to treatment in the same direction for all cases in all countries. The instrument strength (the difference between the probability of the index patient receiving levothyroxine if the other patient received levothyroxine and the probability of the index patient receiving levothyroxine if the other patient did not receive levothyroxine) varied across cases within each country. For example, in the Netherlands, it varied from 20% (case 1) to 47% (case 4).

The sensitivity analysis in which we evaluated the stochastic monotonicity assumption for a continuous instrument (the proportion of all other cases treated) showed violations of this assumption (eTable 2). Although for both countries the probability of treatment increased as the value of the instrument increased for all cases, it did not increase monotonically. Specifically, the probability of treatment was higher if 3/7 other cases were treated than if 4/7 other cases were treated.

DISCUSSION

This survey study showed marked within-country variation amongst GPs in their tendency to treat patients with subclinical hypothyroidism with levothyroxine. Presenting the same cases to all GPs ensured that observed differences in prescribing behaviour truly reflect differences in preference, rather than differences in case-mix. The existence of underlying relative preference for levothyroxine treatment for subclinical hypothyroidism patients amongst GPs as a "pseudo-random" phenomenon is further supported by the very limited decrease in between-GP variance in levothyroxine prescription after adjusting for GP characteristics. Even country explained a relatively small amount of the variation: the within-country variation is considerable compared to between-country differences.

The minimal amount of between-GP variance in levothyroxine prescription explained by GP characteristics within countries is reassuring with regard to main IV assumptions. If GP gender and years of experience were related to relative preference for levothyroxine, this would threaten the validity of the exclusion restriction assumption: years of experience in particular may affect the prognosis of subclinical hypothyroidism patients through other ways than levothyroxine prescription. If the proportion of older patients were related to preference for levothyroxine this would threaten the validity of the independence assumption: the baseline prognosis of patients would then differ according to GP's preference. With regard to the independence assumption, it is important to make the distinction between physician's preference as assessed in this survey and physician's preference as it is typically used as IV in observational studies. A measure of preference based on previous patients of the physician is typically used: the treatment of these previous patients is determined both by the underlying preference of the physician and by characteristics of these patients.[2] Physicians with the same underlying preference (i.e. who would give the same responses to our survey questions) can have a different case-mix of patients, and an estimate of their preference based on treatment of these patients would then differ. Although the assumption of no confounding seems to hold for underlying preference in our survey data, it may well be violated in observational data if measures of preference based on treatment of previous patients are used, due to confounding by case-mix. This

issue of confounding of instruments based on prescribing history was also discussed by Swanson et al. [2]

The preference patterns observed within the six countries deviated in varying degrees from the pattern expected if the deterministic monotonicity assumption would hold. The violation of the deterministic monotonicity assumption in this survey with relatively simple case descriptions indicates it is unlikely to hold for physician's preference as an instrument in true prescription data. For a dichotomous instrument, the bias in the local average treatment effect estimate caused by violation of deterministic monotonicity depends on the proportions of compliers and defiers and the difference in treatment effects for compliers and defiers.[9] For a multi-levelled or continuous instrument, the bias caused by violation of the deterministic monotonicity assumption will be determined by analogous factors: i.e. the severity and pattern of the deviation from monotonicity, and the level of heterogeneity of treatment effects. In our example, heterogeneity is most likely to exist according to TSH levels, but looking at TSH only, there is relatively little violation of deterministic monotonicity.

In these data, the stochastic monotonicity assumption was not falsified when treatment of a different patient of the same GP was used as an instrument. However, in the sensitivity analysis using the proportion of all other patients of

the same GP treated as an instrument, the data were not compatible with the stochastic monotonicity assumption for that instrument. This may be due to the specific setting of the study: a certain proportion of other patients treated often corresponds to a certain pattern of specific cases treated in these data. Overall, these results suggest that the stochastic monotonicity assumption may be plausible for physician's preference-based IV studies, depending on how the instrument is defined. Estimates of preference based on a larger number of previous patients may be more likely in general to violate stochastic monotonicity, because the probability of treatment must increase monotonically across all levels of these instruments for all strata of patients.

The effect estimate under the stochastic monotonicity assumption is not the local average treatment effect but the strength-of-IV-weighted treatment effect, a generalisation of the local average treatment effect with a similar interpretation.[6] There has recently been discussion on the usefulness of the local average treatment effect. It centres around the question of whether the treatment effect for the compliers is a relevant effect,[12;13] particularly because we cannot identify who the compliers are.[12] The strength-of-IV-weighted treatment effect has similar drawbacks to the local average treatment effect: the interpretation is difficult, since it is a weighted average of effects in strata which we cannot identify and for which we do not know the weights.

The existing survey data used for this study provided a unique opportunity to investigate the assumptions underlying the use of physician's preference as an IV, but also presented some limitations. One limitation is the low response rate, which may have affected our results in various ways. Responding GPs may be more aware of guidelines and more alike in their prescription patterns: i.e. the deterministic monotonicity assumption could be violated to a greater extent in the entire GP population. There may have been more 'random' variation in answers if all GPs had responded (i.e. if underlying preference is a stronger determinant of treatment in the respondents than in GPs overall). This would have reduced the overall strength of GP's preference as an instrument. However, we would not expect it to affect the validity of the stochastic monotonicity assumption for treatment of one other case as the instrument: we do not expect such vastly different patterns among non-respondents that treatment of a particular case would be inversely related to treatment of a different case.

All GPs were presented with the cases in the same order. Random ordering of the cases per GP would have been preferable for assessing preference in the context of an IV. It would have enabled us to use a true 'previous case' for the evaluation of the stochastic monotonicity assumption. Furthermore, the ordering of the cases may have had some influence on answers given for specific cases.

By evaluating the stochastic monotonicity assumptions across these eight patient types (strata) in the survey, we considered the characteristics that define these patient types, i.e. age, vitality status and TSH levels, to be a sufficient set of measured and unmeasured common causes of treatment and the outcome. While this may hold for the simplified survey data, this is unlikely to be a sufficient set in a true patient population. We were therefore only able to evaluate the stochastic monotonicity assumption for the simplified setting of the survey. Related to this, the fictitious cases in the survey were not intended to represent any particular population of subclinical hypothyroidism patients for whom we may want to estimate the effect of levothyroxine treatment. Rather, the survey was designed in such a manner that characteristics which were thought to be important in the treatment decision varied among the cases. The cases were intended to represent a well-known clinical decision problem: whether to treat subclinical hypothyroidism. In this sense estimating a treatment effect for this group would be of potential interest, although the types of subclinical hypothyroidism patients represented by the cases are limited. For example, the cases were all women and there was no variation in the symptoms with which they presented.

Findings which may be of interest to clinicians are that we can distinguish several groups of factors which are related to the decision whether to treat a patient with subclinical hypothyroidism: characteristics of the patient, country

(and its guidelines), and GP's preference. In this setting of treatment of subclinical hypothyroidism, the lack of stringent guidelines leaves substantial room for GP's preference to play a role in treatment decisions. While this would provide an opportunity to utilize this variation in an IV study of the effect of treatment of subclinical hypothyroidism, the ultimate aim of such a study would paradoxically be to reduce this preference-based variation through the development of evidence-based guidelines.

In conclusion, our study supports the existence of physician's preference as a determinant in treatment decisions. Little of the variation in preference was explained by characteristics of the GP or their patient population, indicating that main IV assumptions may be plausible for physician's treatment preferences. The deterministic monotonicity assumption did not hold and will generally not be plausible for physician's preference as an instrument. The stochastic monotonicity assumption may be plausible, depending on how the instrument is defined.

References

(1) Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.

(2) Swanson SA, Miller M, Robins JM, Hernán MA. Definition and Evaluation of the Monotonicity Condition for Preference-based Instruments. *Epidemiology* 2015.

(3) Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.

(4) Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 1994;62:467-475.

(5) Small DS, Tan Z. A stochastic monotonicity assumption for the instrumental variables method. Working Paper, Department of Statistics, University of Pennsylvania, 2007.

(6) Small DS, Tan Z, Lorch SA, Brookhart MA. Instrumental variable estimation when compliance is not deterministic: the stochastic monotonicity assumption. 2014. arXiv: 1407.7308v2

(7) Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007;3: Article 14.

(8) den Elzen WP, Lefebre-van de Fliert AA, Virgini V et al. International variation in GP treatment strategies for subclinical hypothyroidism in older adults: a case-based survey. *Br J Gen Pract* 2015;65:e121-e132.

(9) Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 1996;91:444-455.

(10) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537-554.

(11) *PLOTMATRIX: Stata module to plot values of a matrix as different coloured blocks.* [Version S439602. Boston College Department of Economics; 2004.

(12) Swanson SA, Hernán MA. Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation. *Stat Sci* 2014;29:371-374.

(13) Imbens GW. Instrumental Variables: An Econometrician's Perspective. *Stat Sci* 2014;29:323-358.

Figure legends

Figure 1. Distribution per participating country of the number of cases for which a GP would prescribe levothyroxine.

A. The Netherlands (n=117) *B. United Kingdom (n=21)* *C. New Zealand (n=25)*

D. Ireland (n=15) *E. Switzerland (n=235)* *F. Germany (n=55)*

Figure 2. Matrix plots of the prescription patterns of the GPs within each country. GPs are ordered from highest to lowest preference, with their response for each case indicated by the colour of the cell (yes: dark-grey, no: light-grey, missing: mid-grey). GPs with equal preferences were ordered according to their preferences for case 1(first yes, then no) through to 8, and subsequently by their identification-number (if all answers were equal).

A. The Netherlands (n=117) *B. United Kingdom (n=21)* *C. New Zealand (n=25)*

D. Ireland (n=15) *E. Switzerland (n=235)* *F. Germany (n=55)*

Table 1. Age, vitality status and thyroid stimulating hormone (TSH) of the eight cases in the survey.

Case	Age	Vitality status	TSH (mU/L)
1	70	Vital	6
2	70	Vulnerable	6
3	70	Vital	15
4	70	Vulnerable	15
5	85	Vital	6
6	85	Vulnerable	6
7	85	Vital	15
8	85	Vulnerable	15

Adapted from Den Elzen et al, British Journal of General Practice 2015.

Table 2. Characteristics of participating general practitioners (GPs).

GP characteristics	No. (%) Total n=526
Country	
The Netherlands	129 (25)
United Kingdom	22 (4)
New Zealand	31 (6)
Ireland	21 (4)
Switzerland	262 (50)
Germany	61 (12)
Male	373 (71)
Experience as a GP (years)	
<5	41 (8)
5-10	70 (13)
11-15	90 (17)
16-20	82 (16)
21-25	88 (17)
>25	155 (29)
Patients aged 65 years and over in GP practice (%)	
<10	35 (7)
10-20	122 (23)
20-30	188 (36)
>30	181 (34)
Time since last subclinical hypothyroidism diagnosis	
<1 week	76 (14)
1 week-1 month	194 (37)
1 month-1 year	211 (40)

1-3 years	27 (5)
>3 years	18 (3)

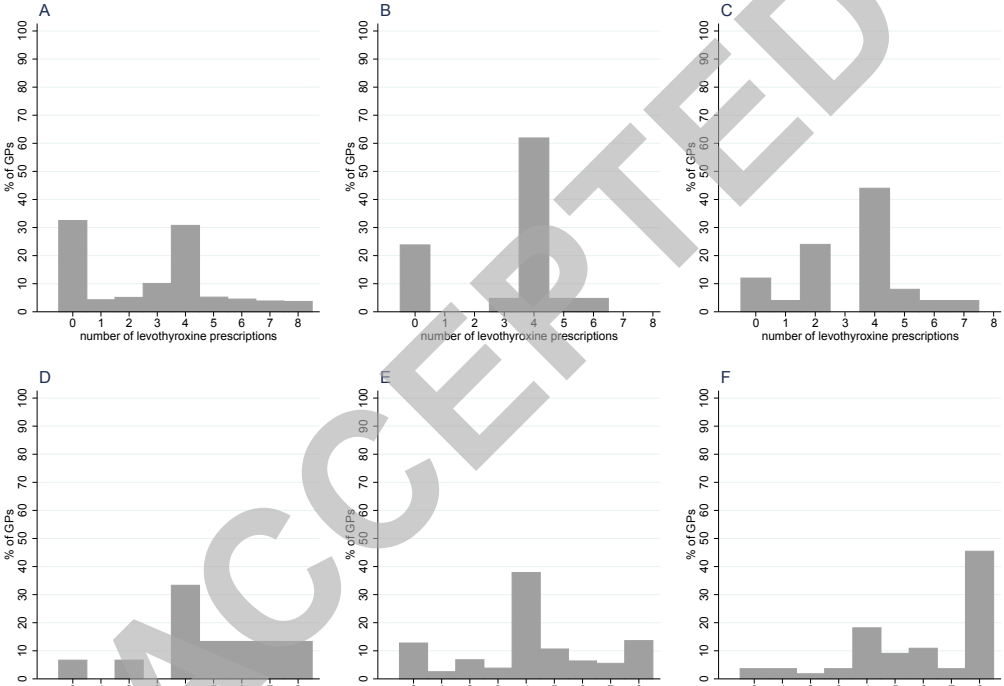
Table 3 Between general practitioner (GP) variance in treatment

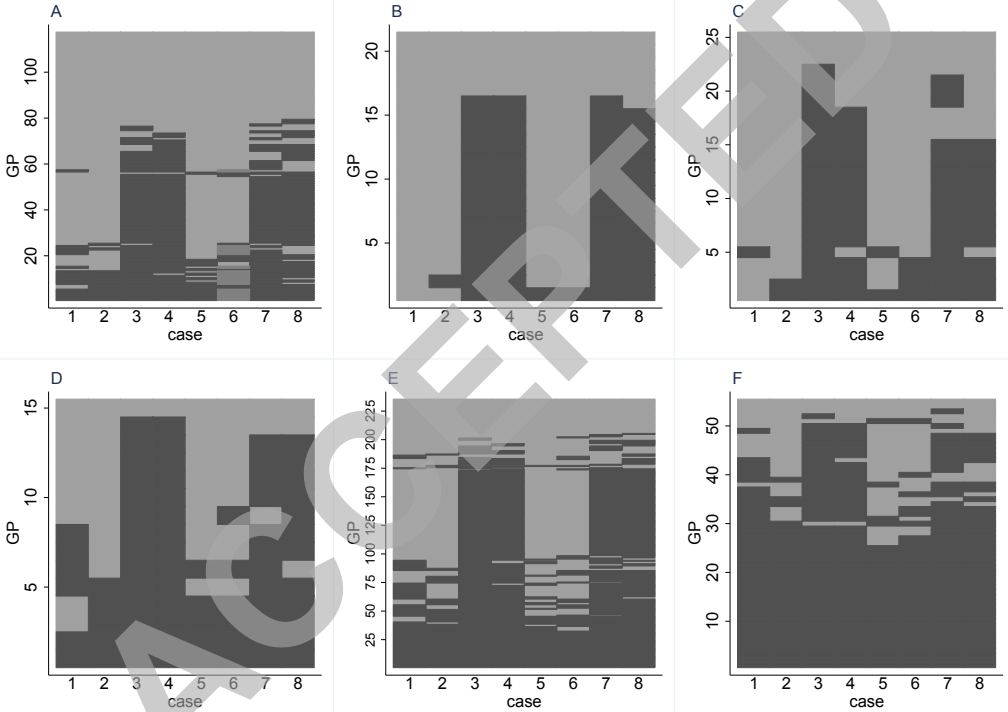
Model	Between GP variance (95% CI)
1: Random effect for GP; fixed effect for age, TSH and vitality status of case	9.9 (8.0;12)
2: Model 1 + fixed effect for country	8.3 (6.7;10)
3: Model 2 + fixed effect for gender and years of experience	8.2 (6.6;10)
4: Model 3 + fixed effect for time since last diagnosis of subclinical hypothyroidism and proportion of patients aged 65 years and over	8.0 (6.5;9.9)

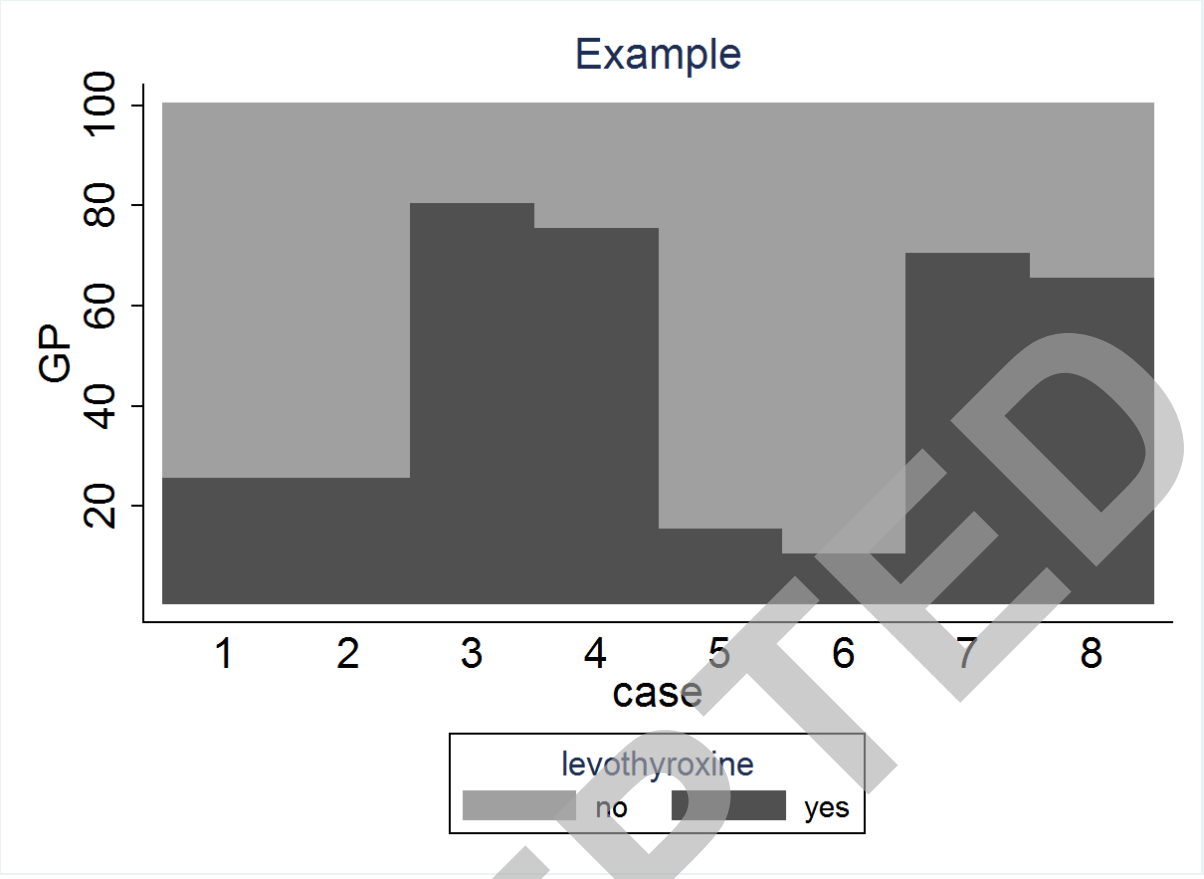
Table 4. Probability (%) of levothyroxine dependent on treatment of a different case by the same general practitioner (GP)

Case	Country																	
	Netherlands (n=117)			UK (n=21)			New Zealand (n=25)			Ireland (n=15)			Switzerland (n=235)			Germany (n=55)		
	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ
1	7	27	21	0	0	0	4	4	0	27	47	20	17	45	28	50	89	39
2	4	27	23	4	6	2	3	14	11	11	45	34	10	41	31	19	73	54
3	44	90	46	64	100	36	82	100	18	84	100	16	67	96	30	76	97	20
4	45	92	47	64	100	36	60	83	23	84	100	16	62	93	31	68	95	27
5	4	27	22	2	8	5	5	12	6	11	45	34	7	36	29	15	64	50
6	6	26	20	2	8	5	9	25	16	22	50	28	11	38	27	23	72	49
7	39	85	46	64	100	36	63	89	26	65	90	25	61	90	29	65	93	28
8	40	78	38	59	94	35	45	74	29	70	87	17	61	90	30	46	87	41

Percentage of yes answers per case within each country, dependent on the treatment of a different case (the 'other patient') by the same GP. Each other answer of the same GP was used as an 'other patient'. Treatment of the 'previous patient' is indicated by – (no levothyroxine) and + (levothyroxine). The columns indicate the following (in %): – : Pr[D=1|Z=0]; +: Pr[D=1|Z=1]; Δ: Pr[D=1|Z=1]-Pr[D=1|Z=0].







eFigure 1. Example of a matrix plot showing prescription patterns which would fulfil the monotonicity assumption. GPs are ordered from highest to lowest preference, with their response for each case indicated by the colour of the cell (yes: dark-grey, no: light-grey).

eTable 1. Response rates per country and overall

Country	Responses	Surveys sent out	Response rate (%)
The Netherlands	129	315	41
United Kingdom	22	178	34
New Zealand	31	850	4
Ireland	21	150	14
Switzerland	262	1086	25
Germany	61	178	34
Total	526	2710	19

Adapted from Den Elzen et al, British Journal of General Practice 2015.

eTable 2. Probability of levothyroxine dependent on treatment of all other cases by the same general practitioner (GP).

Case	Country															
	The Netherlands (n=117)								Switzerland (n=235)							
	0/7	1/7	2/7	3/7	4/7	5/7	6/7	7/7	0/7	1/7	2/7	3/7	4/7	5/7	6/7	7/7
1	0	0	13	25	4	47	73	67	0	33	7	53	15	48	62	86
2	0	0	0	12	1	42	81	100	0	14	12	22	6	37	73	94
3	5	50	75	92	76	100	100	100	6	67	47	98	96	100	100	100
4	0	44	86	100	86	100	84	98	0	45	39	98	88	88	100	100
5	0	0	0	8	8	46	62	87	0	0	6	27	5	13	48	94
6	0	0	0	21	10	28	46	80	3	17	12	36	6	26	53	89
7	3	33	64	87	69	93	100	100	6	64	18	92	80	93	93	100
8	5	50	70	87	58	75	79	80	3	58	36	95	74	88	100	100

Percentage of yes answers per case within each country, dependent on the treatment of the other cases of the same GP. The column headings indicate the proportion of the other patients treated.