Causal inference based lifestyle coaching system for thyroid disease patients when lifestyle variables are continuous

Yeongho Lee [1†] Joonhyung Kim [2] Kyubo Shin [2] Minhyun Kang [1]
Youngin Kwon [1] Gi-Soo Kim* [1,3] Jae Hoon Moon* [2,4]

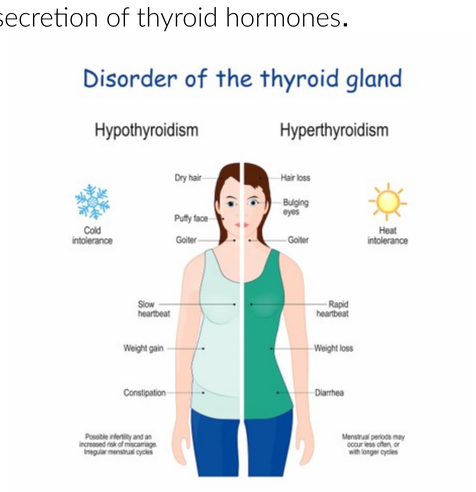
[1]Artificial Intelligence Graduate School, Ulsan National Institute of Science and Technology [2]R&D division, THYROSCOPE INC.

[3]Department of Industrial Engineering, Ulsan National Institute of Science and Technology

[4]Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul National University Bundang Hospital

Introduction

Thyroid dysfunction is a common chronic disease that can be caused by either too much or too little secretion of thyroid hormones.



Disorder of thyroid gland - https://rockymountaindiabetes.com/

- As thyroid dysfunction is a chronic disease, it requires regular monitoring and treatment. Currently, thyroid dysfunction is monitored through blood sampling. So, the patient is required to visit the hospital regularly.
- Regular hospital visits for blood sampling and diagnosis increase the cost of medical care. This can be a burden to patients.

In this work, we provide a means for self-management of thyroid dysfunction. We expect that the self-management system will help reduce unnecessary visits to the hospital and related costs.

Objective

Thyroid dysfunction can be caused by lifestyle factors. For example, healthy lifestyle habits such as regular meals and adequate sleep can help reduce the risk of thyroid dysfunction. The purpose of this study is **to estimate the causal relationship between lifestyle and symptoms of thyroid dysfunction patients and to coach their lifestyle.** If a lifestyle coaching system is introduced, thyroid dysfunction patients can improve their lifestyle and control their symptoms without visiting the hospital too much.

Causal Inference

The reason for using causal inference in building lifestyle coaching systems is that general predictive models do not clearly distinguish the causal relationship between cause and effect as they do not adjust for confounding variables properly.

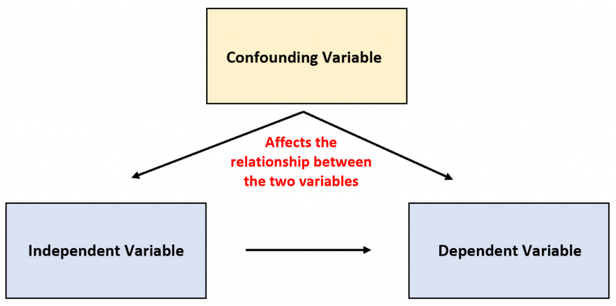


Figure 1. Confounding variable created by statology - www.statology.org

To properly adjust for confounding variables, we will use propensity score methods.

Propensity Score

The propensity score is defined as the conditional probability density of a treatment given confounding variables. Propensity scores, via their balancing property, help to isolate the effect of a treatment from other differences that may exist between treatment and control groups.

Problem Setting

- N : number of patients, N_i : number of i^{th} patient's data
- T_{ij} : treatment (lifestyle habit) value of i^{th} patient at j^{th} time point
- $\mathbf{X}_{ij}$: covariate vector of i^{th} patient at j^{th} time point
- Y_{ij}: symptom value (outcome) of the i^{th} patient at j^{th} time point
- $Y_{ij}(t)$: *potential outcome* that would have been observed if the i^{th} patient had a lifestyle value t at j^{th} time point ($Y_{ij}(t) = Y_{ij}$ if $T_{ij} = t$)

Hyperthyroidism data						
USER ID	X^1	X^2	$\cdots$	X^P	Treatment	Outcome
user 1	X^1_{11}	X^2_{11}	$\cdots$	X^P_{11}	30	$Y(30)$
user 1	X^1_{12}	X^2_{12}	$\cdots$	X^P_{12}	45	$Y(45)$
user 2	X^1_{21}	X^2_{21}	$\cdots$	X^P_{21}	60	$Y(60)$
user 2	X^1_{22}	X^2_{22}	$\cdots$	X^P_{22}	90	$Y(90)$
user 3	X^1_{31}	X^2_{31}	$\cdots$	X^P_{31}	30	$Y(30)$
⋮	⋮	⋮	⋯	⋮	⋮	⋮

Estimation target: the target estimand is $\mathbb{E}[Y(t)]$ for each value t of the lifestyle habit variable, where the expectation is taken over all observations from all patients. $\mathbb{E}[Y(t)]$ is the expected symptom value when all patients show the same value (t) of lifestyle variable.

Methodology We denote the generalized propensity score (GPS) for continuous treatment T as $r(t,x) = f_{T|X}(t|x)$. We also assume the ignorability assumption (Rosenbaum and Rubin, 1983), i.e., $Y(t) \perp\!\!\!\perp T|X$. A known property is that when GPS is fixed, the distribution of confounding variables and treatment are independent($X \perp\!\!\!\perp I(T = t)|r(t,X)$). Due to this property and the ignorability assumption, we can show that the potential outcome and treatment are independent when the GPS is fixed ($Y(t) \perp\!\!\!\perp I(T = t)|r(t,X)$). This property plays a crucial role in estimating $\mathbb{E}[Y(t)]$. According to the derivation below, $\mathbb{E}[Y(t)]$ can be estimated by modeling the symptom value using the observed data only.

$$\mathbb{E}[Y(t)] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i(t)|r(t,x_i)] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i(t)|T_i = t, r(t,x_i)] \ (\because Y(t) \perp\!\!\!\perp I(T=t)|r(t,x))$$
$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i(T_i)|T_i = t, r(t,x_i)] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i|T_i = t, r(t,x_i)] \qquad (*)$$

GPS estimation and estimation of $\mathbb{E}[Y(t)]$

Lifestyle habit variables considered in this study are continuous variables (ex. average amount of consumed alcohol over a month, average number of smoked cigarettes over a month, average sleeping time over a month etc) Therefore, we estimate the **Generalized Propensity Score(GPS).**

We assume the GPS is a normal density.

$$f_\theta(T_{ij}|\mathbf{X}_{ij}) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2\sigma^2}(T_{ij} - \mathbf{X}_{ij}^T\beta)^2\right)$$

- To estimate the parameters β and σ^2, we use the fact that when data are weighted by the inverse of GPS, the treatment and covariates should have covariance 0. (T^*_{ij} and $\mathbf{X^*_{ij}}$ are centered and normalized values of T_{ij} and $\mathbf{X}_{ij}$(Fong, Hazlett and Imai, 2018).)

$$\mathbb{E}\left(\frac{f(T^*_{ij})}{f_\theta(T^*_{ij}|\mathbf{X^*_{ij}})}T^*_{ij}\mathbf{X^*_{ij}}\right) = \int\left\{\int\frac{f(T^*_{ij})}{f_\theta(T^*_{ij}|\mathbf{X^*_{ij}})}T^*_{ij}\,dF(T^*_{ij}|\mathbf{X^*_{ij}})\right\}\mathbf{X^*_{ij}}dF(\mathbf{X^*_{ij}}) \qquad (1)$$
$$= \mathbb{E}(T^*_{ij})\mathbb{E}(\mathbf{X^*_{ij}}) = 0.$$

$$\frac{f(T^*_{ij})}{f_\theta(T^*_{ij}|\mathbf{X^*_{ij}})} = \sigma exp\left[\frac{1}{2\sigma^2}(T^*_{ij} - \mathbf{X^{*T}_{ij}}\beta)^2 - \frac{T^{*2}_{ij}}{2}\right]$$

We find $\theta = (\beta, \sigma)$ that makes the covariance between the treatment and covariate 0 when weighted by the inverse GPS. We use the estimation function below, where $\mathbf{F}_\theta$ is the empirical version of the covariance (1).

$$\mathbf{F}_\theta = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{j=1}^{N_i}\begin{pmatrix}\frac{1}{\sigma^2}(T^*_{ij} - \mathbf{X^{*T}_{ij}}\beta)^2 - 1 \\ \sigma exp\left[\frac{1}{2\sigma^2}(T^*_{ij} - \mathbf{X^{*T}_{ij}}\beta)^2 - \frac{T^{*2}_{ij}}{2}\right]T^*_{ij}\mathbf{X^*_{ij}}\end{pmatrix}, \hat{\theta} = \underset{\theta}{\operatorname{argmin}}||\mathbf{F}_\theta||_2$$

This estimation procedure allows to find value of θ that satisfy the balancing property of GPS. When balancing property holds, the (*) equation is established, which allows causal inference.

- The outcome values are modeled with treatment and GPS, interactions, and square terms.

$$\hat{\mathbb{E}}[Y(t)] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{E}[Y_{ij}|T_{ij} = t, r(t,\mathbf{X}_{ij})]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{j=1}^{N_i}\left(\hat{\alpha_0} + \hat{\alpha_1}t + \hat{\alpha_2}r(t,\mathbf{X}_{ij}) + \hat{\alpha_3}tr(t,\mathbf{X}_{ij}) + \hat{\alpha_4}r(t,\mathbf{X}_{ij})^2 + \hat{\alpha_5}t^2\right)$$

Experiment and Result

- Curve of $\hat{\mathbb{E}}[Y(t)]$ according to value t

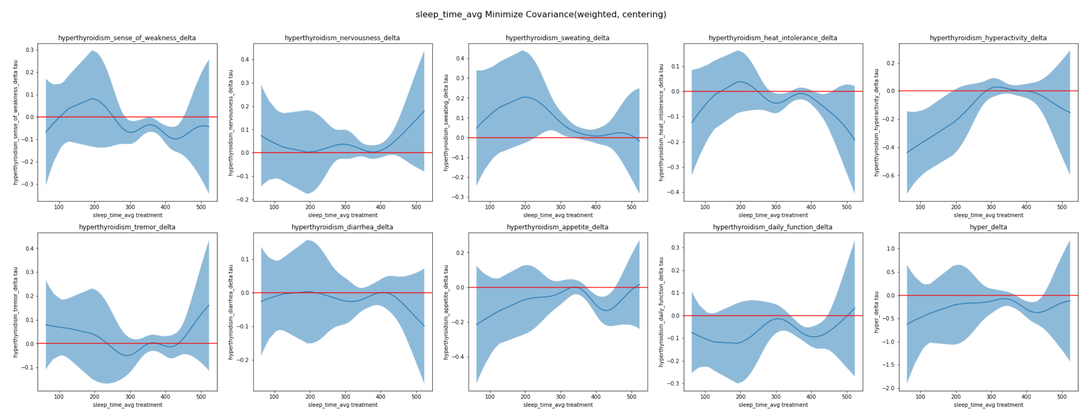
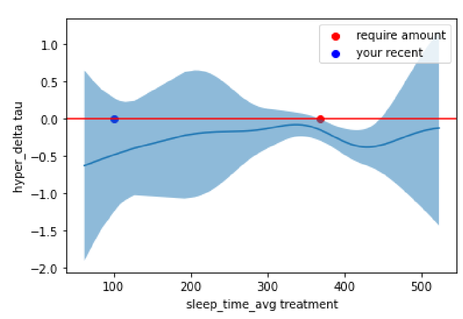


Figure 2. Treatment : sleep time, Method : Minimize Cov(Treatment,Covariates)

Figure 2 shows the curve of $\hat{\mathbb{E}}[Y(t)]$ according to the amount of sleeping time t for 10 different symptoms. The confidence bands are obtained by bootstrapping. Red line is the zero line. If $\hat{\mathbb{E}}[Y(t)]$ is negative on an interval of t, it means that symptoms are relieved when patients sleep for the amount of time in that interval.
For example, according to the first graph, the "sense of weakness" symptom is relieved when the patients sleep between 60 and 530 minutes.

- Recommendation Example



Using the graphs above, we can provide a specific coaching to patients so as they reduce the thyroid dysfunction symptoms. For example, if the patient is currently sleeping 100 minutes (blue dot), we can recommend him/her to sleep 260 minutes more since the confidence band starts to drop below the zero line at 360 minutes of sleep.