The Effect of the Semantics of Diseases to Prevention Management Fees by using Propensity Score Matching

Ryosuke Matsuo
Faculty of Medicine
University of Miyazaki Hospital
Miyazaki City, Japan
ryosuke_matsuo@med.miyazaki-u.ac.jp

Taisuke Ogawa
Faculty of Medicine
University of Miyazaki Hospital
Miyazaki City, Japan
taisuke_ogawa@med.miyazaki-u.ac.jp

Kenji Araki
Faculty of Medicine
University of Miyazaki Hospital
Miyazaki City, Japan
taichan@med.miyazaki-u.ac.jp

***Abstract*—The principal benefit of clinical guidelines is to improve the quality of care received by patients. Although risk factors are defined in clinical guidelines, the semantics of diseases in terms of the weights of treatment is somehow indecipherable. We assume that there are different effects to the fulfillment of prevention management among the diseases' categories such as main diseases and sub diseases. To this end, we analyze the effect of the semantics of diseases that are risk factors when prevention management fees are calculated, by using propensity score matching (PSM). Our key idea to tackle the problem of identifying the important categories of diseases for an analytical purpose is to decompose diseases into the sub categories based on the importance to the fulfillment of prevention management and apply PSM to each of the sub categories for estimating the effects derived from odds ratios using observational data. In this paper, diseases are divided into three categories: main diseases, comorbidities and complications. As the case study, we focus on the pulmonary thromboembolism (PTE) where cancers are regarded as part of the risk factors. We use the diagnosis procedure combination data of 44,257 patients from Jan 1, 2014 to Mar 31, 2018 in University of Miyazaki Hospital. The odds ratios between disease and non-disease groups adjusted by PSM based on six covariates showed that the ranking of the significant diseases' categories at the calculation of the PTE's prevention management fee is followed by main diseases, complications and comorbidities.**

***Index Terms*—propensity score matching, semantics, diseases, prevention management fees**

I. Introduction

Clinical guidelines are "statements that include recommendations, intended to optimize patient care, that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options" [1]. The principal benefit is to improve the quality of care received by patients [2]. Promoting interventions of proved benefit and discouraging ineffective ones by guidelines have the potential to reduce morbidity and mortality [2].

However, as for risk factors in clinical guidelines, the semantics of diseases in terms of the weights of treatment is somehow indecipherable. For example, although one of the risk factors of the pulmonary thromboembolism (PTE) is defined as cancers [3], its category from the aspect of the priority of treatment such as main diseases or sub diseases that contain comorbidities and complications could not be explicitly identified. We assume that there are different effects to the fulfillment of prevention management among the categories of diseases. With the significant diseases' categories for a certain purpose are elucidated, the knowledge could be useful for analytical purposes such as the prediction of the fulfillment of the PTE's prevention management when patients are admitted to hospital.

In order to tackle the problem of identifying the important categories of diseases for an analytical purpose, our key idea is to decompose diseases into the sub categories based on the importance to the fulfillment of prevention management and apply a propensity score matching method to each of the sub categories for estimating the effects by using odds ratios from observational data.

Several propensity score analyses have been conducted in medicine [4]–[6], and Leeper et al. [7] considered the semantics of variables based on the relationship among variable names, the corresponding concepts, and the corresponding terms mentioned in clinical notes, by exploiting medical ontologies. However, there have been no studies that analyze the effect of the semantics of diseases in terms of the importance of an aspect under consideration by exploiting propensity score matching.

II. Objective

The objective of this paper is to analyze the effect of the semantics of diseases that are risk factors when prevention management fees are calculated, exploiting a propensity score matching method. As the case study, we focus on the pulmonary thromboembolism (PTE) to analyze the effect of the semantics of diseases for the prevention management fee of the PTE.

III. Methods

A. Data acquisition

The diagnosis procedure combination (DPC) data of 44,257 patients from Apr 1, 2014 to Mar 31, 2018 in University of

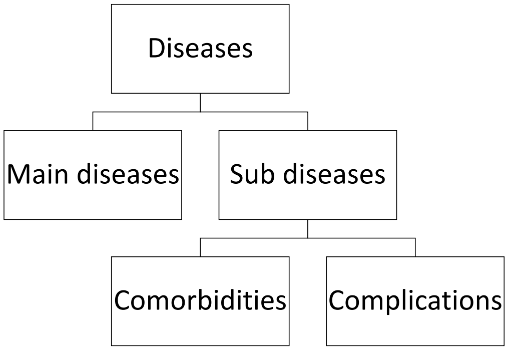

Fig. 1. The hierarchical structure of diseases' categories.

Miyazaki Hospital were used. The DPC data include discharge abstract and administrative claims data [8].

The treatment assignment variable of propensity score matching as the dependent variable was each of the diseases' categories of cancers. Diseases were divided into main diseases and sub diseases in terms of the priority of treatment. The sub diseases were divided into comorbidities and complications, thus the total number of diseases' categories was three in this study. The hierarchical structure of the diseases' categories is represented by Fig.1. The covariates as the independent variables were age, male, female, BMI, smoking and cancer stage classification. The outcome variable was the prevention management fee of the PTE. In the DPC data, the six covariates and the treatment assignment variable were collected from the discharge abstract and the outcome variable was collected from administrative claims data.

B. Procedures

The all variables we used were binary where the value is 0 or 1. The six covariates were processed to transform the binary variables based on its risk other than the variables of sex by the conditions as shown in Table I. The value of age was 1 if the age is 65 years old or more that indicates elderly [9], 0 otherwise. Due to data we used, the value of male was 1 if the corresponding value of sex is 1, 0 otherwise. The value of female was 1 if the corresponding value of sex is 2, 0 otherwise. The value of the body mass index (BMI) was 1 if the BMI is 25 or more as overweight, 0 otherwise. The value of smoking was 1 if the brinkman index is 600 or more [10], 0 otherwise. The value of cancer was 1 if the cancer stage is 3 or 4, 0 otherwise.

C. Statistical analysis

In order to exclude the influence of the confounding in the comparison of outcomes between two groups derived from observational data, propensity score matching adjusted covariates of two groups, which are a group of patients who have cancers (disease group) and a group of patients who do not have cancers (non-disease group). Propensity score methods can reduce the effects of confounding by adjusting the distribution of observed covariates between two groups based on the propensity score [11]. The propensity score is defined as "the conditional probability of assignment to a particular treatment given a vector of observed covariates" [12].

A logistic regression model was performed to attain the propensity scores of each patient that are the probability of diseases in this study.

A nearest neighbor matching algorithm was employed to 1:1 matching of the propensity scores between the disease group and non-disease group. In the matching, a caliper of width was set to 0.2 of the standard deviation of the logit of the propensity score. We use a package MatchIt [13] to perform propensity score matching.

To measure the covariate balance, the standardized mean difference (SMD) was employed by using a package tableone [14]. A standard difference of less than 0.1 has been supported the assumption of balance between two groups [15].

We exploit odds ratios to estimate the effect of the semantics of diseases. The odds ratio (OR) is computed as follows

$$OR = \frac{p_1 / (1 - p_1)}{p_0 \,/ (1 - p_0)} \tag{1}$$

where p_1 and p_0 denote the probability of the fulfillment of the prevention management of the PTE in the disease group and non-disease group, respectively.

The c-statistic of the propensity score model was employed to confirm the discrimination power between the disease group and non-disease group. It is equal to the area under the receiver operating characteristic curve and it is derived from a package pROC [16].

These statistical analyses were executed on R [17].

IV. Results

For the study of main diseases, Table II indicated that the disease group and non-disease group are 12,583 and 31,674, respectively. By using propensity score matching, the matched samples of the disease group and non-disease group were

TABLE I
The conditions of the six covariates

Covariate name	Condition
age	1 if the age is 65 years old or more, 0 otherwise.
male	1 if the corresponding value of sex is 1, 0 otherwise.
female	1 if the corresponding value of sex is 2, 0 otherwise.
BMI	1 if the body mass index is 25 or more, 0 otherwise.
smoking	1 if the brinkman index is 600 or more, 0 otherwise.
cancer	1 if the cancer stage is 3 or 4, 0 otherwise.

TABLE II
The sample sizes of main diseases' study

	Non-disease group	Disease group
All	31674	12583
Matched	10093	10093
Unmatched	21581	2490

10,093. The unmatched samples of the two groups were 2,490 and 21,581, respectively. Other than the variables of sex, the SMD of each covariate was less than 0.1, as shown in Table III. Before propensity score matching, the odds ratio between the two groups was 1.4, 95 % CI 1.34 - 1.47, P $<$ 0.001 (written as "unadjusted" in Table IV). After propensity score matching, the adjusted odds ratio between the two groups was 1.41, 95 % CI 1.33 - 1.5, P $<$ 0.001 (written as "Adjusted by propensity score matching" in Table IV). The c-statistic was 0.7059, 95% CI 0.7003-0.7115.

For the study of comorbidities, Table V indicated that the disease group and non-disease group are 4,692 and 39,565, respectively. By using propensity score matching, the matched samples of the disease group and non-disease group were 4,692. The unmatched sample of the non-disease group was 34,873. There was no unmatched sample of the disease group. The SMD of all covariates was less than 0.1, as shown in Table VI. Before propensity score matching, the odds ratio between the two groups was 1.19, 95 % CI 1.11 - 1.27, P $<$ 0.001 (written as "unadjusted" in Table VII). After propensity score matching, the adjusted odds ratio between the two groups was 1.09, 95 % CI 1.0 - 1.2, P $<$ 0.1 (written as "Adjusted by propensity score matching" in Table VII). The c-statistic was 0.6504, 95% CI, 0.6419-0.6589.

For the study of complications, Table VIII indicated that the disease group and non-disease group are 1,697 and 42,560, respectively. By using propensity score matching, the matched samples of the disease group and non-disease group were 1,697. The unmatched sample of the non-disease group was 40,863. There was no unmatched sample of the disease group. The SMD of all covariates was less than 0.1, as shown in Table IX. Before propensity score matching, the odds ratio between the two groups was 1.14, 95 % CI 1.02 - 1.27, P $<$ 0.05 (written as "unadjusted" in Table X). After propensity score matching, the adjusted odds ratio between the two groups was 1.26, 95 % CI 1.08 - 1.48 , P $<$ 0.01 (written as "Adjusted by propensity score matching" in Table X). The c-statistic was 0.6248, 95% CI, 0.611-0.6386.

TABLE III
THE SMD OF MAIN DISEASES' STUDY

	Stratified by main disease		
	0	**1**	**SMD**
n	**10093**	**10093**	
age = 1 (%)	6744 (66.8)	6640 (65.8)	0.022
male = 1 (%)	5413 (53.6)	6024 (59.7)	0.122
female = 1 (%)	4680 (46.4)	4069 (40.3)	0.122
BMI = 1 (%)	2679 (26.5)	2512 (24.9)	0.038
smoking = 1 (%)	2015 (20.0)	2241 (22.2)	0.055
cancer = 1 (%)	210 (2.1)	210 (2.1)	< 0.001

TABLE IV
THE ODDS RATIOS OF MAIN DISEASES' STUDY

Model	**Odds ratio (95 % CI)**	**P**
Unadjusted	1.4 (1.34 - 1.47)	< 0.001
Adjusted by propensity score matching	1.41 (1.33 - 1.5)	< 0.001

TABLE V
THE SAMPLE SIZES OF COMORBIDITIES' STUDY

	Non-disease group	**Disease group**
All	39565	4692
Matched	4692	4692
Unmatched	34873	0

TABLE VI
THE SMD OF COMORBIDITIES' STUDY

	Stratified by comorbidity		
	0	**1**	**SMD**
n	**4692**	**4692**	
age = 1 (%)	3127 (66.6)	3061 (65.2)	0.030
male = 1 (%)	2642 (56.3)	2756 (58.7)	0.049
female = 1 (%)	2050 (43.7)	1936 (41.3)	0.049
BMI = 1 (%)	1025 (21.8)	1011 (21.5)	0.007
smoking = 1 (%)	867 (18.5)	1009 (21.5)	0.076
cancer = 1 (%)	868 (18.5)	868 (18.5)	< 0.001

TABLE VII
THE ODDS RATIOS OF THE COMORBIDITIES' STUDY

Model	**Odds ratio (95 % CI)**	**P**
Unadjusted	1.19 (1.11 - 1.27)	< 0.001
Adjusted by propensity score matching	1.09 (1.0 - 1.2)	< 0.1

In these three studies (main diseases, comorbidities and complications), before the propensity score matching, the probabilities of the fulfillment of the prevention management of the PTE in the disease group and non-disease group were 29.71 % vs 23.14 % (the difference 6.58 %, 95 % CI 5.66 - 7.5, P $<$ 0.001), 27.98 % vs 24.66 % (the difference 3.33 %, 95 % CI 1.99 - 4.7, P $<$ 0.001) and 27.4 % vs 24.91 % (the difference 2.49 %, 95 % CI 0.38 - 4.7, P $<$ 0.05), respectively. After the propensity score matching, the probabilities of the fulfillment of the prevention management of the PTE in the disease group and non-disease group were 31.48 % vs 24.52 % (the difference 6.96 %, 95 % CI 5.72 - 8.19, P $<$ 0.001), 27.98 % vs 26.24 % (the difference 1.75 %, 95 % CI -0.05 - 3.55, P $<$ 0.1) and 27.4 % vs 22.98 % (the difference 4.42 %, 95 % CI 1.5 - 7.33, P $<$ 0.01), respectively.

V. DISCUSSION

The results by using propensity score matching suggest that the significant ranking of the diseases' categories at the calculation of the prevention management fee of the PTE based on the adjusted odds ratios is followed by main diseases (1.41, 95 % CI 1.33 - 1.5, P $<$ 0.001), complications (1.26, 95 % CI 1.08 - 1.48 , P $<$ 0.01) and comorbidities (1.09, 95 % CI 1.0 - 1.2, P $<$ 0.1). From the aspect of the weights of treatment, the results of the adjusted odds ratios of the studies of main diseases and sub diseases that compose of comorbidities and complications could reasonably consider that main diseases are more important than sub diseases. By reducing the influence of the confounding based on propensity

TABLE VIII
THE SAMPLE SIZES OF THE COMPLICATIONS' STUDY

	Non-disease group	Disease group
All	42560	1697
Matched	1697	1697
Unmatched	40863	0

TABLE IX
THE SMD OF THE COMPLICATIONS' STUDY

	Stratified by complication		
	0	1	SMD
n	1697	1697	
age = 1 (%)	1149 (67.7)	1115 (65.7)	0.043
male = 1 (%)	950 (56.0)	975 (57.5)	0.030
female = 1 (%)	747 (44.0)	722 (42.5)	0.030
BMI = 1 (%)	373 (22.0)	389 (22.9)	0.023
smoking = 1 (%)	389 (22.9)	360 (21.2)	0.041
cancer = 1 (%)	228 (13.4)	262 (15.4)	0.057

TABLE X
THE ODDS RATIOS OF THE COMPLICATIONS' STUDY

Model	Odds ratio (95 % CI)	P
Unadjusted	1.14 (1.02 - 1.27)	< 0.05
Adjusted by propensity score matching	1.26 (1.08 - 1.48)	< 0.01

score matching, the adjusted odds ratio of the comorbidities' study decreased, by contrast, the adjusted odds ratio of the complications' study increased. As shown by the adjusted odds ratios of the studies of complications and comorbidities (1.26 vs 1.09), it is assumed that the patient's condition with comorbidities is more stable than the patient's condition with complications.

The knowledge derived from this study or the estimated effects of the three categories of diseases by the adjusted odds ratios could be useful for predicting the fulfillment of the prevention management of the PTE when patients are admitted to hospital for the improvement of hospital management. The semantic analysis has potential to apply to different case studies as well as different categories of variables.

Regarding the evaluation of propensity score matching in those analyses, the results of the SMD were mostly less than 0.1 other than the SMD of sex in the case of the main diseases' study, thus the covariates were acceptably balanced between the disease group and non-disease group. Furthermore, as the results of the c-statistic of the propensity score models for the studies of three categories of diseases were more than 60 %, the two groups were tolerably discriminated by the propensity scores. Therefore, propensity score matching for the three categories of diseases was well performed by the evaluation measures of the SMD and the c-statistic.

A limitation of propensity score matching is that unobserved variables can not be included in propensity score analyses. Another limitation is that propensity score matching brought about a number of unmatched patients in the two groups. For example, in the case of the comorbidities' study and the complications' study, the unmatched patients of the disease groups were 34,873 and 40,863, respectively.

VI. Conclusion

This study analyzed the effect of the semantics of diseases in terms of the weights of treatment in which the semantic types of diseases are main diseases, comorbidities and complications to prevention management fees, by using propensity score matching. The results of the adjusted odds ratios showed that the ranking of significant diseases' categories at the calculation of the prevention management fee of the PTE is followed by main diseases, complications and comorbidities. The acquired knowledge could be useful for building a prediction model of patients who are admitted to hospital for the improvement of hospital management. The semantic analysis of variables in terms of the importance of an aspect under consideration by exploiting propensity score matching could be applied to different case studies as well as different categories of variables.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K09948.

References

[1] E. Steinberg, S. Greenfield, D. M. Wolman, M. Mancher, R. Graham *et al.*, *Clinical practice guidelines we can trust*. National Academies Press, 2011.

[2] S. H. Woolf, R. Grol, A. Hutchinson, M. Eccles, and J. Grimshaw, "Potential benefits, limitations, and harms of clinical guidelines," *BMJ*, vol. 318, no. 7182, pp. 527–530, 1999.

[3] JCS Joint Working Group, "Guidelines for the diagnosis, treatment and prevention of pulmonary thromboembolism and deep vein thrombosis (JCS 2009)," *Circulation Journal*, vol. 75, no. 5, pp. 1258–1281, 2011.

[4] V. H. Thourani, S. Kodali, R. R. Makkar, H. C. Herrmann, M. Williams, V. Babaliaros, R. Smalling, S. Lim, S. C. Malaisrie, S. Kapadia *et al.*, "Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis," *The Lancet*, vol. 387, no. 10034, pp. 2218–2225, 2016.

[5] D. Pincus, B. Ravi, D. Wasserstein, A. Huang, J. M. Paterson, A. B. Nathens, H. J. Kreder, R. J. Jenkinson, and W. P. Wodchis, "Association between wait time and 30-day mortality in adults undergoing hip fracture surgery," *JAMA*, vol. 318, no. 20, pp. 1994–2003, 2017.

[6] R. J. Desai, M. Mahesri, Y. Abdia, J. Barberio, A. Tong, D. Zhang, P. Mavros, S. C. Kim, and J. M. Franklin, "Association of osteoporosis medication use after hip fracture with prevention of subsequent nonvertebral fractures: An instrumental variable analysis," *JAMA Network Open*, vol. 1, no. 3, p. e180826, 2018.

[7] N. J. Leeper, A. Bauer-Mehren, S. V. Iyer, P. LePendu, C. Olson, and N. H. Shah, "Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes," *PloS one*, vol. 8, no. 5, p. e63499, 2013.

[8] H. Yasunaga, H. Matsui, H. Horiguchi, K. Fushimi, and S. Matsuda, "Application of the diagnosis procedure combination (DPC) data to clinical studies," *Journal of UOEH*, vol. 36, no. 3, pp. 191–197, 2014.

[9] WHO, "Proposed working definition of an older person in africa for the mds project: Definition of an older or elderly person," *http://www. who. int/healthinfo/survey/ageingdefnolder/en/*.

[10] N. Miyatake, J. Wada, Y. Kawasaki, K. Nishii, H. Makino, and T. Numata, "Relationship between metabolic syndrome and cigarette smoking in the japanese population," *Internal Medicine*, vol. 45, no. 18, pp. 1039–1043, 2006.

[11] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011.

[12] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[13] D. E. Ho, K. Imai, G. King, and E. A. Stuart, “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference,” *Political Analysis*, vol. 15, no. 3, pp. 199–236, 2007.

[14] K. Yoshida and J. Bohn, “Package ‘tableone’ for R,” https://cran.r-project.org/web/packages/tableone/tableone.pdf.

[15] S.-L. T. Normand, M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil, “Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores,” *Journal of Clinical Epidemiology*, vol. 54, no. 4, pp. 387–398, 2001.

[16] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/