การลดความเอนเอียงในการศึกษาเชิงสังเกตและการศึกษากึ่งทดลองโดยใช้วิธีคะแนนความโน้มเอียง

The Bias Reduction in Observational and Quasi-Experimental Studies by Using Propensity Score Method

นฤมล สุดใจ[1*] และ มณฑิรา ดวงสาพล[2]
Narumol Sudjai[1] and Monthira Duangsaphon[2]
[1]ภาควิชาศัลยศาสตร์ออร์โธปิดิคส์และกายภาพบำบัด คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล
[2]สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
[1]Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University
[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University
วันที่ส่งบทความ : 3 เมษายน 2563 วันที่แก้ไขบทความ : 10 กันยายน 2563 วันที่ตอบรับบทความ : 13 พฤศจิกายน 2563
Received: 3 April 2020, Revised: 10 September 2020, Accepted: 13 November 2020

บทคัดย่อ

ความเอนเอียงจากการเลือกตัวอย่างเป็นปัญหาเฉพาะในการศึกษาเชิงสังเกตและการศึกษากึ่งทดลองซึ่งเป็นเหตุให้ไม่สามารถทำการเปรียบเทียบระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมได้ เพื่อแก้ไขผลกระทบของความเอนเอียงจากการเลือกตัวอย่างนี้มีวิธีเชิงสถิติหลายวิธีที่ถูกนำเสนอ เช่น การวิเคราะห์แบบแบ่งชั้นภูมิ การวิเคราะห์หลายตัวแปรและ propensity score วิธี propensity score ถูกนำมาใช้อย่างกว้างขวางเพื่อลดความเอนเอียงจากการเลือกตัวอย่างนี้ โดยกำหนดให้ propensity score คือความน่าจะเป็นที่ผู้ป่วยจะได้รับทรีตเมนต์ภายใต้เงื่อนไขปัจจัยกวนต่าง ๆ และให้ปรับสมดุลระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมด้วย propensity score หลังจากนั้นจึงสามารถเปรียบเทียบผลลัพธ์ของการศึกษาใน 2 กลุ่ม โดยใช้การทดสอบเชิงสถิติที่เหมาะสมกับข้อมูลต่อไป ในบทความนี้ผู้นิพนธ์อธิบายหลักการพื้นฐานของ propensity score และแสดงให้เห็นถึงการนำไปใช้ผ่านตัวอย่างการศึกษาในอดีต

คำสำคัญ: คะแนนความโน้มเอียง ความเอนเอียงจากการเลือกตัวอย่าง การสุ่ม การศึกษาเชิงสังเกต การศึกษากึ่งทดลอง

Abstract

Selection bias is particular problem in observational and quasi-experimental studies, which it gives rise to noncomparability between treated and non-treated (or control) groups. To remedy the effect of selection bias, many statistical methods have been proposed such as stratified analysis, multivariate analysis, and propensity score. Propensity score method was widely used to reduce this problem. Propensity score is

[*]ที่อยู่ติดต่อ E-mail address: narumol.sud@mahidol.ac.th

defined as the probability of a subject to receive a treatment conditional on the confounding factors (covariates) and then provides to balance the covariates between the treated and control groups by using propensity score. After that we can be compared the outcome in two groups by using appropriate statistical tests for the data. In this article, we describe a basic principle of the propensity score and illustrate the uses through applied example in previous study.

Keywords: propensity score, selection bias, randomization, observational study, quasi-experimental study

1. บทนำ

การทดลองแบบสุ่มและมีกลุ่มควบคุม (randomized control trials: RCTs) ถูกพิจารณาให้เป็นมาตรฐาน (gold standard) สำหรับการออกแบบการวิจัย (research design) ในงานวิจัยทางด้านการแพทย์ สาธารณสุขและวิทยาศาสตร์สุขภาพ เมื่อผู้วิจัยต้องการประเมินประสิทธิผลของอิทธิพลทรีตเมนต์ (treatment effect) หรือ interventions เพราะการสุ่ม (randomization) เป็นการหลีกเลี่ยงความเอนเอียงในการเลือกตัวอย่างและลดปัจจัยกวน (confounding factors) ที่อาจจะส่งผลต่อผลลัพธ์ของการศึกษานั้น [1] แต่บางการศึกษามีข้อจำกัดที่จะทำการศึกษาแบบ RCT เช่น ข้อจำกัดเรื่องจริยธรรม การศึกษาวิจัยที่มีความเสี่ยงต่อผู้ป่วยสูง และงบประมาณที่ใช้สำหรับดำเนินการวิจัยมีไม่เพียงพอ เป็นต้น ดังนั้นการศึกษาเชิงสังเกตจึงถูกนำมาใช้เมื่อมีข้อจำกัดดังกล่าวข้างต้น แต่อย่างไรก็ตาม ผู้วิจัยต้องแน่ใจก่อนว่าลักษณะของหน่วยตัวอย่างในแต่ละกลุ่มที่นำมาเปรียบเทียบมีลักษณะที่เหมือนกันหรือคล้ายกันเมื่อเริ่มต้นการศึกษาจึงจะสามารถเปรียบเทียบกันได้ ในงานวิจัยทางด้านการแพทย์นั้นการศึกษาเชิงสังเกตถูกนำมาใช้มากกว่าการศึกษาเชิงทดลองเนื่องจากมีข้อจำกัด เช่น ขนาดตัวอย่างที่ทำการศึกษามีน้อย ข้อจำกัดด้านจริยธรรมมีน้อยกว่าและงบประมาณที่ใช้ต่ำกว่า เป็นต้น ซึ่งจะเห็นได้จากงานวิจัยทางด้านโรคกระดูกและข้อส่วนใหญ่เป็นการศึกษาเชิงสังเกตและการศึกษาแบบย้อนหลัง (retrospective study) [2]-[4]

สำหรับตัวแปรที่เป็นตัวกวน (confounders) ต้องมีความสัมพันธ์กับปัจจัยที่สนใจศึกษาและผลลัพธ์ของการศึกษานั้น ยกตัวอย่างเช่น ผู้วิจัยต้องการเปรียบเทียบระยะเวลาในการกลับมาผ่าตัดแก้ไขซ้ำ (time to revision) ในผู้ป่วยที่ได้รับการรักษาด้วยการผ่าตัดเปลี่ยนข้อสะโพกเทียม (total hip arthroplasty) ระหว่างการใช้อุปกรณ์ 2 แบบ คือแบบที่ 1 ผิวข้อสะโพกเทียมชนิดเซรามิกบนเซรามิก (ceramic-on-ceramic implant) และแบบที่ 2 ผิวข้อสะโพกเทียมมีส่วนประกอบของเซรามิกและพลาสติกโพลิเมอร์ (ceramic-on-polyethylene implant) ในการเลือกใช้อุปกรณ์ข้อสะโพกเทียมของแพทย์สำหรับการศึกษานี้ไม่ได้ทำการสุ่มแต่เป็นการให้ตามข้อบ่งชี้หรือตามความสมัครใจของผู้ป่วย ทำให้กลุ่มผู้ป่วยที่ได้รับการเลือกให้ใช้อุปกรณ์ข้อสะโพกเทียมแบบที่ 1 จะเป็นผู้ป่วยที่มีอายุน้อยกว่าอีกกลุ่มซึ่งใช้อุปกรณ์แบบที่ 2 ถ้าหากอายุมีความสัมพันธ์กับการกลับมาผ่าตัดแก้ไขซ้ำด้วยแล้วอายุจึงถือว่าเป็นตัวแปรกวน แต่ผู้วิจัยไม่ได้นำปัจจัยเรื่องอายุของผู้ป่วยนี้เข้ามารวมไว้ในการวิเคราะห์ข้อมูลส่งผลให้การประมาณค่าประสิทธิผลของอิทธิพลทรีตเมนต์ เช่น Odds ratio เอนเอียง การแก้ปัญหาเพื่อที่จะควบคุมตัวแปรกวนนี้มีหลายวิธี ซึ่งสามารถทำได้ทั้งในขั้นตอนของการออกแบบการวิจัยและการวิเคราะห์ข้อมูล ตัวอย่างวิธีควบคุมตัวแปรกวนที่ใช้ในขั้นตอนของการออกแบบการวิจัยคือ

ต้องพิจารณาเกณฑ์การคัดเลือกและคัดออกของผู้ร่วมวิจัยอย่างเข้มงวดและต้องระบุให้ชัดเจนหรืออาจจะพิจารณาทำการจับคู่ เป็นต้น ส่วนวิธีควบคุมตัวแปรกวนในขั้นตอนของการวิเคราะห์ข้อมูลอาจจะใช้วิธี stratified analysis, multivariable analysis และ propensity score เป็นต้น

วิธี propensity score นี้ถูกเสนอขึ้นใช้เป็นครั้งแรกใน ปี ค.ศ. 1983 โดย Rosenbaum และ Rubin [5] เพื่อเป็นทางเลือกหนึ่งในการประเมินประสิทธิผลของอิทธิพลทรีตเมนต์หรือ interventions ในกรณีที่การศึกษานั้นไม่สามารถทำการสุ่มได้ ปัจจุบันได้มีการนำวิธีการวิเคราะห์ข้อมูลด้วย propensity score มาใช้ในงานวิจัยทางด้านการแพทย์เพิ่มมากขึ้นเรื่อย ๆ [6]-[7] โดยหลักการของ propensity score method คือนำเอาปัจจัยกวนต่าง ๆ ที่มีผลต่อการตัดสินใจให้ทรีตเมนต์หรือ interventions มาคำนวณเป็นคะแนน 0 ถึง 1 โดยที่ตัวอย่างผู้ป่วยที่มีคะแนนสูงหมายถึงมีโอกาสที่จะได้รับทรีตเมนต์หรือ interventions มาก นั่นหมายความว่าหากเรานำผู้ป่วยที่มี propensity score เท่ากันหรือใกล้เคียงกันมาเปรียบเทียบก็จะใกล้เคียงกับหลักการสุ่ม นั่นคือผู้ป่วยทุกรายมีโอกาสได้รับ assign ทรีตเมนต์หรือ interventions พอ ๆ กันนั่นเอง

ในบทความนี้ผู้นิพนธ์จะกล่าวถึงหลักการพื้นฐานของ propensity score ขั้นตอนการวิเคราะห์ข้อมูลทางสถิติโดยใช้ propensity score การคำนวณ propensity score และ propensity score method การวิเคราะห์ข้อมูลทางสถิติโดยใช้ propensity score ด้วยโปรแกรม R พร้อมทั้งยกตัวอย่างงานวิจัยที่นำการวิเคราะห์ข้อมูลด้วยวิธี propensity score matching ไปใช้

2. เนื้อหา

2.1 แนวคิดของ propensity score (e_i)

Propensity score คือความน่าจะเป็นที่ผู้ป่วยจะได้รับทรีตเมนต์เมื่อ $\underset{\sim}{x}_i$ คือตัวแปรร่วมหรือปัจจัยกวนต่าง ๆ ก่อนการศึกษาวิจัย (baseline covariates) ซึ่งตัวแปรร่วมเหล่านี้มีความสัมพันธ์กับปัจจัยที่สนใจศึกษาและผลลัพธ์ของการศึกษานั้น โดยสามารถเขียนในรูปแบบสัญลักษณ์ได้ดังนี้

$$e_i = P(Z_i = 1 \mid \underset{\sim}{x}_i) \quad ; i = 1, 2, 3, ..., n$$

และ $$P(Z_1 = z_1, ..., Z_n = z_n \mid x_1, ..., x_n) = \prod_{i=1}^{n} \left\{ e_i^{z_i} \left(1 - e_i\right)^{1-z_i} \right\}$$

เมื่อ e_i คือ propensity score ($0 < e_i < 1$)

Z_i คือตัวแปร treatment assignment

z_i คือค่าของตัวแปร treatment assignment หน่วยที่ i ; $i = 1, 2, 3, ..., n$ ซึ่งมีค่าได้เพียง 2 ค่า โดยที่ $z_i = 1$ แทนหน่วยตัวอย่างที่ได้รับทรีตเมนต์และ $z_i = 0$ แทนหน่วยตัวอย่างที่ไม่ได้รับทรีตเมนต์

$\underset{\sim}{x}_i$ คือเวกเตอร์ขนาด $1 \times (p+1)$ ที่มีสมาชิกในแถวที่ i ของเมทริกซ์ $\underset{\sim}{X}$

$$\underset{\sim}{x}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}_{1 \times (p+1)}$$

$\underset{\sim}{X}$ คือเมทริกซ์ของตัวแปรร่วมที่มีขนาด $n \times (p+1)$

p คือจำนวนตัวแปรร่วม

n คือขนาดตัวอย่าง

สำหรับการศึกษาแบบ RCT ค่า e_i จะเป็นค่าคงที่ ซึ่งปกติแล้ว $e_i = 0.5$ ถ้าหน่วยตัวอย่างถูกแบ่งออกเป็น 2 กลุ่ม เท่า ๆ กัน

2.2 ขั้นตอนการวิเคราะห์ข้อมูลทางสถิติโดยใช้ propensity score มีดังนี้

ขั้นตอนที่ 1 พิจารณาเลือกตัวแปรกวนที่มีความสัมพันธ์กับปัจจัยที่สนใจศึกษาและผลลัพธ์ของการศึกษานั้น

ขั้นตอนที่ 2 คำนวณค่า propensity score

ขั้นตอนที่ 3 นำเอาข้อมูลทั้ง 2 กลุ่ม คือกลุ่มทรีตเมนต์และกลุ่มควบคุมมาปรับสมดุลโดยเลือกใช้ propensity score method วิธีใดวิธีหนึ่งจาก 4 วิธี คือ matching, stratification, covariate adjustment และ inverse probability of treatment weighting (IPTW) ซึ่งทั้ง 4 วิธีนี้ทำภายใต้ propensity score ที่คำนวณได้จากขั้นตอนที่ 2

ขั้นตอนที่ 4 ประเมินความสมดุลของ baseline characteristics (หรือเรียกว่า baseline covariates) ระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมว่ามีการกระจายเหมือนกันหรือไม่ ในการประเมินเบื้องต้นให้เปรียบเทียบความเหมือนหรือคล้ายกันของ baseline covariates ระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมด้วยค่าเฉลี่ยหรือค่ามัธยฐานเมื่อตัวแปรร่วมเป็นข้อมูลเชิงปริมาณ ส่วนตัวแปรร่วมที่เป็นข้อมูลเชิงกลุ่มให้เปรียบเทียบการกระจายระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุม นอกจากนี้สามารถใช้ standardized difference ในการประเมินความสมดุลของ baseline covariates ระหว่าง 2 กลุ่มได้ดังนี้

สำหรับตัวแปรร่วมที่เป็นข้อมูลเชิงปริมาณค่า standardized difference คือ

$$d = \frac{\left(\bar{x}_{treatment} - \bar{x}_{control}\right)}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

เมื่อ $\bar{x}_{treatment}$ คือค่าเฉลี่ยตัวอย่างของตัวแปรร่วมในกลุ่มทรีตเมนต์

$\bar{x}_{control}$ คือค่าเฉลี่ยตัวอย่างของตัวแปรร่วมในกลุ่มควบคุม

$s^2_{treatment}$ คือความแปรปรวนตัวอย่างของตัวแปรร่วมในกลุ่มทรีตเมนต์

$s^2_{control}$ คือความแปรปรวนตัวอย่างของตัวแปรร่วมในกลุ่มควบคุม

สำหรับตัวแปรร่วมที่เป็นข้อมูลเชิงกลุ่มค่า standardized difference คือ

$$d = \frac{\left(\hat{p}_{treatment} - \hat{p}_{control}\right)}{\sqrt{\frac{\left[\hat{p}_{treatment}(1-\hat{p}_{treatment})\right] + \left[\hat{p}_{control}(1-\hat{p}_{control})\right]}{2}}}$$

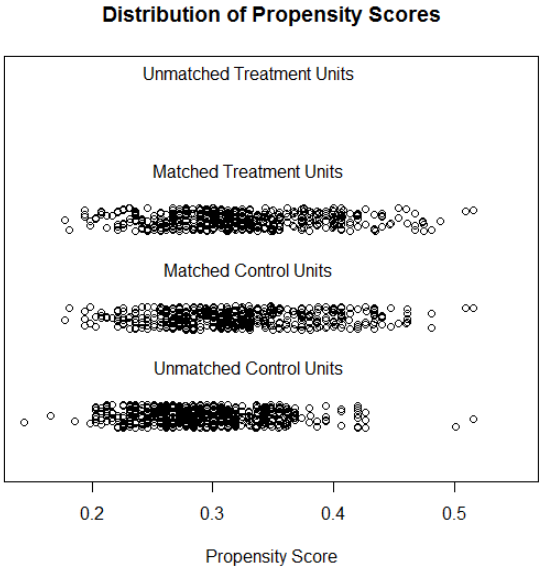
เมื่อ $\hat{p}_{treatment}$ คือ ความชุก (prevalence) ของ dichotomous variable ในกลุ่มทรีตเมนต์

$\hat{p}_{control}$ คือ ความชุกของ dichotomous variable ในกลุ่มควบคุม

กรณีที่ตัวแปรร่วมเป็นข้อมูลเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า (multilevel categorical variables) สามารถหาค่า standardized difference โดยการจัดให้เป็น binary indicator variable [8]

การประเมินความสมดุลของ baseline covariates ด้วยค่า standardized difference (d) นี้จะพิจารณาค่า d ที่มีค่ามากกว่า 0.1 ซึ่งโดยปกติแล้วจะถือว่า baseline covariates ระหว่าง 2 กลุ่ม ไม่สมดุลกัน [9]

เมื่อปรับสมดุลแล้วข้อมูลจะมีลักษณะการกระจายดังตัวอย่างในรูปที่ 1 ดังนี้



รูปที่ 1. แสดงลักษณะของข้อมูลก่อนและหลังปรับสมดุล baseline covariates ด้วยกราฟจุด (dot plot) ของ propensity score ในกลุ่มตัวอย่างทั้ง 2 กลุ่ม

ขั้นตอนที่ 5 วิเคราะห์ข้อมูลทางสถิติเพื่อประเมินประสิทธิผลของอิทธิพลทรีตเมนต์หรือ interventions ด้วยสถิติทดสอบที่เหมาะสม

2.3 การคำนวณ Propensity score (e_i)

วิธีการคำนวณที่ใช้ในการประมาณค่า propensity score มีหลายวิธี เช่น ตัวแบบการถดถอยลอจิสติก (logistic regression model), tree-based methods และ neural networks เป็นต้น แต่โดยส่วนใหญ่แล้วจะใช้วิธีตัวแบบการถดถอยลอจิสติก [10] ดังนั้นผู้นิพนธ์ขอยกตัวอย่างการคำนวณ propensity score โดยใช้การวิเคราะห์การถดถอยลอจิสติกแบบสองกลุ่ม (binary logistic regression analysis) เมื่อกำหนดให้ treatment assignment เป็นตัวแปรตาม (dependent variable: Z_i) ซึ่งมีค่าได้เพียง 2 ค่า คือ $z_i = 1$ แทนหน่วยตัวอย่างที่ได้รับทรีตเมนต์ และ $z_i = 0$ แทนหน่วยตัวอย่างที่ไม่ได้รับทรีตเมนต์ ส่วนตัวแปรกวนทั้งหมดให้เป็นตัวแปรร่วม โดยตัวแบบการถดถอยลอจิสติกที่นำมาใช้ในการประมาณค่า propensity score คือ

$$\ln\left(\frac{e_i}{1-e_i}\right) = \ln\left(\frac{P(Z_i=1|\underset{\sim}{x}_i)}{1-P(Z_i=1|\underset{\sim}{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

และ $e_i = P(Z_i = 1|\underset{\sim}{x}_i)$; $i = 1, 2, 3, \ldots, n$

จะได้ว่า $\hat{e}_i = \frac{1}{1+e^{-(\underset{\sim}{x}_i\underset{\sim}{\hat{\beta}})}}$

โดยที่ $\hat{e}_i$ คือค่าประมาณของ propensity score หน่วยที่ $i\,;\, i=1,2,3,...,n$

$\underset{\sim}{\hat{\beta}}$ คือเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่มีขนาด $(p+1)\times 1$

ตัวอย่างที่ 1 เมื่อมีตัวแปรร่วม จำนวน 3 ตัวแปร คือ เพศ อายุ และโรคประจำตัว โดยใช้วิธีประมาณค่า propensity score ด้วย logistic regression model มีดังนี้

กำหนดให้ $e_i = P(Z_i = 1 | \underset{\sim}{x}_i)\;\; ; i=1,2,3,...,n$

เมื่อ Z_i คือตัวแปร treatment assignment

z_i คือค่าของตัวแปร treatment assignment หน่วยที่ $i\,;\, i=1,2,3,...,n$ ซึ่งมีค่าได้เพียง 2 ค่า โดยที่ $z_i = 1$ แทนหน่วยตัวอย่างที่ได้รับทรีตเมนต์และ $z_i = 0$ แทนหน่วยตัวอย่างที่ไม่ได้รับทรีตเมนต์

x_{1i} = เพศ (Gender) โดยที่ 0 = เพศชาย และ 1 = เพศหญิง

x_{2i} = อายุ (Age) มีหน่วยเป็นปี

x_{3i} = โรคประจำตัว (Underlying disease: UD) โดยที่ 0 = ไม่มีโรคประจำตัว และ 1 = มีโรคประจำตัว

จะได้ว่า $\hat{e}_i = \frac{1}{1+e^{-(\underset{\sim}{x}_i\underset{\sim}{\hat{\beta}})}} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 Gender_i+\hat{\beta}_2 Age_i+\hat{\beta}_3 UD_i)}}$

สมมติว่า ค่าประมาณพารามิเตอร์ $\underset{\sim}{\hat{\beta}}$ โดยใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum likelihood method) คือ $\hat{\beta}_0$ = -3.95, $\hat{\beta}_1$ = 0.69, $\hat{\beta}_2$ = 0.03 และ $\hat{\beta}_3$ = 0.32

ดังนั้น $\hat{e}_i = \frac{1}{1+e^{-[(-3.95)+(0.69*Gender_i)+(0.03*Age_i)+(0.32*UD_i)]}}$

เมื่อหน่วยตัวอย่างคนที่ 1 เป็นเพศหญิง อายุ 87 ปี และมีโรคประจำตัว จะได้ว่า

$$\hat{e}_1 = \frac{1}{1+e^{-[(-3.95)+(0.69*1)+(0.03*87)+(0.32*1)]}} = 0.42$$

ดังนั้น ค่าประมาณ propensity score ของหน่วยตัวอย่างที่ 1 มีค่าเท่ากับ 0.42

2.4 Propensity score method

Propensity score method มี 4 วิธี ที่ถูกนำมาใช้ในงานวิจัยคือ matching, stratification, covariate adjustment และ inverse probability of treatment weighting (IPTW) สำหรับ 3 วิธีแรกนำเสนอโดย Rosenbaum และ Rubin [5] ส่วนวิธีที่ 4 IPTW นำเสนอโดย Rosenbaum [11] การเลือกใช้ propensity score method ควรเลือกตามการประมาณค่าเป้าหมายที่สนใจศึกษา (an estimand of interest) ซึ่งขึ้นอยู่กับคำถามวิจัยและประชากรเป้าหมาย การประมาณค่าเป้าหมายโดยทั่วไปคือผลกระทบโดยเฉลี่ยของทรีตเมนต์ต่อผู้ป่วยที่ได้รับทรีตเมนต์หรือ interventions ที่สนใจเท่านั้น (the average effect of the treatment on the treated: ATT) และผลกระทบโดยเฉลี่ย

ของทรีตเมนต์หรือ interventions ต่อผู้ป่วยที่เข้าร่วมการศึกษาทั้งหมดซึ่งหมายถึงกลุ่มทรีตเมนต์และกลุ่มควบคุม (the average treatment effect: ATE) สำหรับการประมาณค่าแบบ ATE จะเหมาะสมถ้าทุก ๆ ทรีตเมนต์สามารถนำมาใช้กับผู้ป่วยทุกคนได้ ในขณะที่การประมาณค่าแบบ ATT จะเหมาะสมเมื่อผู้ป่วยมีลักษณะที่ควรจะได้รับทรีตเมนต์ที่สนใจศึกษาเท่านั้น หลังจากที่ตัดสินใจเลือกการประมาณค่าเป้าหมายแบบ ATT หรือ ATE แล้วจึงพิจารณาเลือกวิธี propensity score method เพื่อปรับสมดุลของ baseline covariates ต่อไป โดยแต่ละวิธีมีรายละเอียดดังนี้

1) วิธี Matching

วิธีนี้สามารถใช้ประมาณค่าแบบ ATT เท่านั้น ซึ่งเหมาะสำหรับการศึกษาที่จำนวนขนาดตัวอย่างของกลุ่มควบคุมมีมากกว่ากลุ่มที่ได้รับทรีตเมนต์ วิธีนี้เป็นวิธีที่นิยมนำมาใช้ในงานวิจัยทางด้านการแพทย์ โดยวิธีการจับคู่นั้นจะพิจารณาจากค่าประมาณของ propensity score ที่มีค่าเท่ากันหรือใกล้เคียงกันมากที่สุด ซึ่งส่วนใหญ่จะใช้การจับคู่แบบหนึ่งต่อหนึ่ง (one-to-one matching หรือเรียกว่า pair matching) โดยที่แต่ละคู่ประกอบไปด้วยผู้ป่วยที่ได้รับทรีตเมนต์และไม่ได้รับทรีตเมนต์ แต่บางการศึกษาอาจจะเลือกใช้แบบอื่นได้ เช่น one-to-many matching (1:M) ขึ้นอยู่กับความเหมาะสมของแต่ละการศึกษา สำหรับผู้ป่วยที่ไม่สามารถจับคู่ได้จะถูกตัดออกไป จะเห็นได้ว่า ข้อเสียของวิธี matching นี้คือต้องใช้ขนาดตัวอย่างใหญ่ หลังจากที่จับคู่เรียบร้อยแล้วจึงทำการเปรียบเทียบด้วยสถิติทดสอบที่เหมาะสมกับข้อมูลต่อไป

2) วิธี Stratification (หรือ Subclassification)

วิธีนี้สามารถใช้ประมาณค่าได้ทั้งแบบ ATT และ ATE โดยแยกผู้ป่วยออกเป็นชั้นภูมิ (stratum) ซึ่งในการแบ่งชั้นภูมินั้นจะใช้ propensity score เป็นเกณฑ์ในการพิจารณา การแบ่งจะจัดให้ชั้นภูมิเดียวกันมี propensity score เท่ากันหรือใกล้เคียงกันรวมทั้งจำนวนผู้ป่วยในแต่ละชั้นภูมิเหมือนกันด้วย โดยทั่วไปจะแบ่งชั้นภูมิออกเป็น 5 กลุ่มเท่า ๆ กัน โดยใช้ควอนไทล์ (quantiles) ของ propensity score ซึ่งจะสามารถกำจัดความเอนเอียงที่มาจากตัวกวนที่วัดได้ (measured confounders) ถึง 90% [9] หลังจากที่แบ่งชั้นภูมิเสร็จแล้วให้ทำการเปรียบเทียบผลลัพธ์ของการศึกษาระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมในแต่ละชั้นภูมิต่อไป

3) วิธี Covariate Adjustment

วิธีนี้จะใช้ propensity score เป็นตัวแปรร่วมในการปรับความแตกต่างเมื่อเริ่มต้นการศึกษา (baseline differences) โดยใช้ตัวแบบการถดถอยเชิงเส้นหรือตัวแบบการถดถอยลอจิสติก การประเมินประสิทธิผลของอิทธิพลทรีตเมนต์ถูกหาโดยใช้ค่าประมาณสัมประสิทธิ์การถดถอยจากตัวแบบที่สร้างขึ้น ซึ่งมีตัวแปร treatment assignment และ propensity score เป็นตัวแปรอิสระ (หรือเรียกว่า covariates) โดย treatment effect (τ) สามารถประมาณค่าได้ดังนี้

$$\hat{\tau} = (\bar{Y}_{treatment} - \bar{Y}_{control}) - \hat{\beta}(\bar{X}_{treatment} - \bar{X}_{control})$$

จะเห็นได้ว่า วิธีนี้จะใช้ค่า propensity score จริง ๆ ในขณะที่ 2 วิธีที่กล่าวมาแล้วข้างต้นใช้ค่าประมาณของ propensity score เพื่อทำการจับคู่หรือแบ่งชั้นภูมิภายใต้ค่า propensity score ที่เหมือนหรือใกล้เคียงกันมากที่สุด

4) วิธี Inverse Probability of Treatment Weighting (IPTW)

วิธีนี้สามารถใช้ประมาณค่าได้ทั้งแบบ ATT และ ATE ซึ่งใช้วิธีการถ่วงน้ำหนักที่แตกต่างกัน ในการถ่วงน้ำหนักระหว่างกลุ่มทรีตเมนต์และกลุ่มควบคุมนั้นจะต้องสะท้อนถึงประชากรที่ทำการศึกษาซึ่งเป็นหลักการที่เหมือนกับการสุ่มตัวอย่างแบบถ่วงน้ำหนักในงานวิจัยเชิงสำรวจ โดย

กำหนดให้ Z_i คือตัวแปรที่แสดงว่าผู้ป่วยคนที่ i ได้รับทรีตเมนต์หรือไม่ และ e_i คือ propensity score ซึ่งเราจะใช้ e_i มาเป็นตัวถ่วงน้ำหนัก ในการถ่วงน้ำหนักนั้นสามารถกำหนดเป็น

$$w_i = \left(\frac{Z_i}{e_i}\right) + \left(\frac{1-Z_i}{1-e_i}\right) \; ; i = 1, 2, 3, ..., n$$

จะเห็นได้ว่า การถ่วงน้ำหนักของผู้ป่วยแต่ละคนเท่ากับความน่าจะเป็นผกผันของการได้รับทรีตเมนต์ที่ผู้ป่วยได้รับจริง ๆ ซึ่งหน่วยตัวอย่างจะถูกถ่วงน้ำหนักด้วย $\frac{1}{e_i}$ สำหรับกลุ่มที่ได้รับทรีตเมนต์ และ $\frac{1}{1-e_i}$ สำหรับกลุ่มควบคุม

วิธี IPTW นี้ถูกเสนอขึ้นใช้เป็นครั้งแรกโดย Rosenbaum [11] ในปี ค.ศ. 1987 ต่อมาได้มีการพัฒนาวิธีการถ่วงน้ำหนักสำหรับการประมาณค่าเป้าหมายแบบ ATE ดังนี้

$$w_{i,ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i Y_i}{e_i}\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-Z_i)Y_i}{1-e_i}\right) ; i = 1, 2, 3, ..., n$$

เมื่อ Y_i คือตัวแปรผลลัพธ์ของหน่วยตัวอย่างที่ i

สำหรับวิธีการถ่วงน้ำหนักเพื่อประมาณค่าเป้าหมายแบบ ATT คือ

$$w_{i,ATT} = Z_i + \left(\frac{(1-Z_i)e_i}{1-e_i}\right) ; i = 1, 2, 3, ..., n$$

และ $$w_{i,ATC} = \frac{Z_i(1-e_i)}{e_i} + (1-Z_i) \; ; i = 1, 2, 3, ..., n$$

เมื่อ $w_{i,ATC}$ คือวิธีการถ่วงน้ำหนักสำหรับการประมาณค่าผลกระทบโดยเฉลี่ยของทรีตเมนต์ต่อผู้ป่วยที่ไม่ได้รับทรีตเมนต์ (the average effect of treatment in the controls: ATC) [10]

ข้อดีของวิธี IPTW นี้คือยังคงเก็บข้อมูลผู้ป่วยไว้ได้ทั้งหมดและลดความเอนเอียง (bias) ได้มากกว่าวิธี stratification และ covariate adjustment แต่การถ่วงน้ำหนักอาจจะไม่ถูกต้องหรือไม่เสถียรถ้าความน่าจะเป็นที่จะได้รับทรีตเมนต์ต่ำมาก ๆ (หมายถึงค่าประมาณของ propensity score มีค่าเข้าใกล้ศูนย์ [9]-[10]

3. การวิเคราะห์ข้อมูลทางสถิติโดยใช้ propensity score ด้วยโปรแกรม R

ตัวอย่างที่ 2 ต้องการเปรียบเทียบอัตราการเกิดภาวะแทรกซ้อนหลังผ่าตัด (overall complication) ระหว่างกลุ่มตัวอย่างผู้ป่วยที่มีน้ำหนักตัวปกติ (normal group) และผู้ป่วยที่มีน้ำหนักตัวมากกว่าปกติ (obese group) ว่าแตกต่างกันหรือไม่ โดยข้อมูลพื้นฐานประกอบด้วย เพศ (sex) อายุ (age) โรคประจำตัว (underlying disease) และจำนวนวันที่ผู้ป่วยพักรักษาตัวอยู่ในโรงพยาบาล (LOS)

สำหรับชุดคำสั่งที่ใช้ในการวิเคราะห์ข้อมูลด้วยโปรแกรม R เวอร์ชั่น 3.6.3 มีขั้นตอนดังนี้ (ตัวอักษรหลังเครื่องหมาย '#' เป็นคำอธิบายไม่ใช่ชุดคำสั่งที่ให้โปรแกรมทำงาน)

ขั้นตอนที่ 1 : ให้โปรแกรมอ่านข้อมูลชื่อ non-obese.txt มาสู่โปรแกรม R ใน data fame ชื่อ q แล้วเปลี่ยนชื่อใหม่เป็น df.normal หลังจากนั้นกำหนดให้ชุดข้อมูลนี้เป็นข้อมูลตัวอย่างของกลุ่ม Normal

```
q=read.table("f://non-obese.txt",dec = ".", header = TRUE) # import data
df.normal <- data.frame(q)                # defines the dataset
df.normal$Sample <- as.factor('Normal')    # Normal group
```

ขั้นตอนที่ 2 : ให้โปรแกรมอ่านข้อมูลชื่อ obese.txt มาสู่โปรแกรม R ใน data fame ชื่อ w แล้วเปลี่ยนชื่อใหม่เป็น df.obese หลังจากนั้นกำหนดให้ชุดข้อมูลนี้เป็นข้อมูลตัวอย่างของกลุ่ม Obese

```
w=read.table("f://obese.txt",dec = ".", header = TRUE) # import data
df.obese <- data.frame(w)                # defines the dataset
df.obese$Sample <- as.factor('Obese')       # Obese group
```

ขั้นตอนที่ 3 : รวมข้อมูลตัวอย่างของทั้ง 2 ชุด เข้าไว้ด้วยกัน โดย package ที่จำเป็นต้องใช้คือ wakefield เมื่อรวมข้อมูลตัวอย่างของทั้ง 2 ชุด เข้าด้วยกันแล้วจึงกำหนดให้ชุดข้อมูลนี้มีชื่อว่า mydata

```
# Merging the dataframes
library(wakefield)
mydata <- rbind(df.normal, df.obese)
mydata$Group <- as.logical(mydata$Sample == 'Normal')
```

ขั้นตอนที่ 4 : ก่อนทำ propensity score matching ให้วิเคราะห์ข้อมูลทางสถิติเพื่อเปรียบเทียบข้อมูลพื้นฐานและอัตราการเกิดภาวะแทรกซ้อนหลังผ่าตัดระหว่างกลุ่มตัวอย่างทั้ง 2 กลุ่ม โดย package ที่จำเป็นต้องใช้คือ pacman, kableExtra และ tableone

```
library(pacman)
library(tableone)
pacman::p_load(tableone)
table1 <- CreateTableOne(vars = c('SEX', 'AGE', 'Underlying', 'LOS',
                                  'OverallCompilaction'), data = mydata,
                                  factorVars = 'SEX', strata = 'Sample')
table1 <- print(table1, printToggle = FALSE, noSpaces = TRUE)
library(kableExtra)
table1 %>%
kable(caption = 'Table 1: Comparison of unmatched samples') %>%
kable_styling()%>%
footnote(general = "A p-value of 0.05 or less was considered as significally
         significant")
```

ขั้นตอนที่ 5 : ทำการจับคู่ข้อมูลด้วย propensity score matching สำหรับ package ที่จำเป็นต้องใช้คือ MatchIt โดยใช้ตัวแปร SEX AGE และ Underlying ในการคำนวณค่า propensity score แล้วจึงทำการจับคู่แบบ 1:1

```
library(MatchIt)
match.it <- matchit(Group ~ SEX + AGE + Underlying, data = mydata,
                    method="nearest", ratio=1)  # Matching the samples
```

ขั้นตอนที่ 6 : หลังจากทำ propensity score matching แล้วให้วิเคราะห์ข้อมูลทางสถิติเพื่อเปรียบเทียบข้อมูลพื้นฐานและอัตราการเกิดภาวะแทรกซ้อนหลังผ่าตัดระหว่างกลุ่มตัวอย่างทั้ง 2 กลุ่ม

```
df.match <- match.data(match.it)[1:ncol(mydata)]
rm(df.normal, df.obese)
pacman::p_load(tableone)
table2 <- CreateTableOne(vars = c('SEX', 'AGE', 'Underlying', 'LOS',
                                  'OverallCompilaction'), data = df.match,
                                  factorVars = 'SEX', strata = 'Sample')
table2 <- print(table2, printToggle = FALSE, noSpaces = TRUE)
kable(table2[,1:3], align = 'c',
      caption = 'Table 2: Comparison of matched samples')%>%
      kable_styling()%>%
footnote(general = "A p-value of 0.05 or less was considered as significally
         significant")
```

จะได้ผลลัพธ์ดังนี้

ตารางที่ 1. Comparison of unmatched samples

Table 1: Comparison of unmatched samples

	Normal	Obese	p
n	489	1093	
SEX = M (%)	72 (14.7)	143 (13.1)	0.423
AGE (mean (SD))	70.36 (7.83)	69.02 (7.60)	0.001
Underlying = Yes (%)	362 (74.0)	889 (81.3)	0.001
LOS (mean (SD))	7.46 (4.76)	7.51 (3.47)	0.799
OverallCompilaction = Yes (%)	26 (5.3)	81 (7.4)	0.154

Note:
A p-value of 0.05 or less was considered as significally significant

ตารางที่ 2. Comparison of matched samples

Table 2: Comparison of matched samples

	Normal	Obese	p
n	489	489	
SEX = M (%)	72 (14.7)	64 (13.1)	0.518
AGE (mean (SD))	70.36 (7.83)	70.38 (7.61)	0.970
Underlying = Yes (%)	362 (74.0)	363 (74.2)	1.000
LOS (mean (SD))	7.46 (4.76)	7.64 (4.12)	0.532
OverallCompilaction = Yes (%)	26 (5.3)	44 (9.0)	0.035

Note:
A p-value of 0.05 or less was considered as significally significant

จากตารางที่ 1 ก่อนทำ propensity score matching จะเห็นได้ว่า ข้อมูลพื้นฐาน ได้แก่ อายุเฉลี่ยและสัดส่วนของผู้ป่วยที่มีโรคประจำตัวระหว่างกลุ่มตัวอย่างผู้ป่วยที่มีน้ำหนักตัวปกติและกลุ่มตัวอย่างผู้ป่วยที่มีน้ำหนักตัวมากกว่าปกติแตกต่างกันอย่างมีนัยสำคัญทางสถิติ (p-value = 0.001 และ 0.001 ตามลำดับ) และเมื่อเปรียบเทียบอัตราการเกิดภาวะแทรกซ้อนหลังผ่าตัดพบว่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value = 0.154)

และจากตารางที่ 2 หลังทำ propensity score matching จะเห็นได้ชัดเจนว่า ข้อมูลพื้นฐาน ได้แก่ อายุเฉลี่ยและสัดส่วนของผู้ป่วยที่มีโรคประจำตัวระหว่างกลุ่มตัวอย่างผู้ป่วยที่มีน้ำหนักตัวปกติและกลุ่มตัวอย่างผู้ป่วยที่มีน้ำหนักตัวมากกว่าปกติแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value = 0.970 และ 0.999 ตามลำดับ) และเมื่อเปรียบเทียบอัตราการเกิดภาวะแทรกซ้อนหลังผ่าตัดพบว่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ (p-value = 0.035)

4. การนำ propensity score ไปใช้ในงานวิจัย

ปัจจุบันได้มีการนำ propensity score มาใช้ในงานวิจัยทางด้านการแพทย์เพิ่มมากขึ้น [6]-[7] และหนึ่งในวิธีที่นิยมนำมาใช้วิเคราะห์ข้อมูลมากที่สุดคือ propensity score matching ในที่นี้ผู้นิพนธ์จะยกตัวอย่างของงานวิจัยที่นำวิธีนี้มาใช้ในการวิเคราะห์ข้อมูลดังนี้

ตัวอย่างที่ 3 งานวิจัยของ Fukuda และคณะ [12] เป็นการศึกษาแบบย้อนหลัง มีวัตถุประสงค์เพื่อศึกษาผลกระทบของเทคนิคการให้ยาระงับความรู้สึก (general anesthesia หรือ spinal anesthesia) ที่มีผลต่อความสามารถในการทำกิจวัตรประจำวัน (activity daily life: ADL) 5 ด้าน ได้แก่ การรับประทานอาหาร การแต่งตัว การใช้ห้องสุขา การอาบน้ำ และการเดินในผู้ป่วยสูงอายุที่กระดูกสะโพกหักและได้รับการรักษาด้วยการผ่าตัดที่นอนพักรักษาตัวอยู่ที่โรงพยาบาลและหลังจากกลับไปพักฟื้นที่บ้าน โดยที่ general anesthesia คือการระงับความรู้สึกแบบทั่วร่างกาย และ spinal anesthesia หรือ subarachnoid block คือการฉีดยาชาเข้าไปในช่อง subarachnoid จะออกฤทธิ์ที่ spinal nerve และ dorsal ganglion ทำให้มีอาการชาและหย่อนกล้ามเนื้อในบริเวณที่ถูก block นั้น ซึ่งเป็นการระงับความรู้สึกเฉพาะส่วน (regional anesthesia) ผลการศึกษาพบว่า จากจำนวนผู้ป่วย

ทั้งหมด 12,342 ราย เป็นผู้ป่วยที่ได้รับ general anesthesia จำนวน 6,918 (56.1%) ราย และผู้ป่วยที่ได้รับ spinal anesthesia จำนวน 5,424 (43.9%) ราย ก่อนที่จะทำ propensity score matching เห็นได้ชัดว่า มี 18 ปัจจัยจากทั้งหมด 23 ปัจจัย (baseline factors) ที่มีความแตกต่างกันระหว่างกลุ่มผู้ป่วยที่ได้รับ general anesthesia และ spinal anesthesia อย่างมีนัยสำคัญทางสถิติ แต่หลังจากที่ทำการจับคู่ด้วย propensity score matching แบบ 1:1 จะเหลือผู้ป่วยในแต่ละกลุ่มเพียง 3,949 ราย ผลจากการทดสอบความแตกต่างของ baseline factors ทั้งหมด 23 ปัจจัย ระหว่างกลุ่มผู้ป่วยที่ได้รับ general anesthesia และ spinal anesthesia พบว่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติและเมื่อทดสอบความสัมพันธ์ระหว่างความสามารถในการทำกิจวัตรประจำวัน (ADL scores) 4 ด้าน ได้แก่ การรับประทานอาหาร การแต่งตัว การอาบน้ำและการเดินหลังจากกลับไปพักฟื้นที่บ้านกับเทคนิคการให้ยาระงับความรู้สึกพบว่าไม่มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ แต่ความสามารถในการทำกิจวัตรประจำวันในการใช้ห้องสุขาหลังจากกลับไปพักฟื้นที่บ้านมีความสัมพันธ์กับเทคนิคการให้ยาระงับความรู้สึกอย่างมีนัยสำคัญทางสถิติ โดยกลุ่มผู้ป่วยที่ได้รับ spinal anesthesia มีจำนวนหรือเปอร์เซนต์ที่ไม่สามารถช่วยเหลือตนเองในการใช้ห้องสุขาสูงกว่ากลุ่มผู้ป่วยที่ได้รับ general anesthesia อย่างมีนัยสำคัญทางสถิติ (spinal vs. general: 40.3% vs. 36.7% ; p-value < 0.003) ในการศึกษานี้มีผู้ป่วยที่เสียชีวิต จำนวน 174 ราย ซึ่งจะถูกคัดออกจากการวิเคราะห์ข้อมูลโดยเป็นผู้ป่วยในกลุ่มที่ได้รับ general anesthesia จำนวน 109 (1.6%) ราย และกลุ่มที่ได้รับ spinal anesthesia จำนวน 65 (1.2%) ราย และอัตราการตาย (mortality rates) ของผู้ป่วย 2 กลุ่มนี้แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value = 0.078) และจากการวิเคราะห์ข้อมูลด้วย multivariate logistic regression analysis พบว่าเทคนิคการให้ยาระงับความรู้สึกไม่มีความสัมพันธ์กับการแย่ลงของความสามารถในการทำกิจวัตรประจำวันทั้ง 5 ด้าน ได้แก่ การรับประทานอาหาร การแต่งตัว การใช้ห้องสุขา การอาบน้ำและการเดินหลังจากทำผ่าตัดรักษากระดูกสะโพกหัก และการไม่ให้ยาต้านอักเสบชนิดไม่ใช่สเตียรอยด์ (NSAIDs) มีความสัมพันธ์กับการแย่ลงของความสามารถในการทำกิจวัตรประจำวันทั้ง 5 ด้าน นอกจากนี้ การที่อายุเพิ่มขึ้น เป็นเพศชาย มี Modified Charlson Comorbidity Index หรือ CCI scores สูง ๆ เป็นโรคทางจิตเวช (psychiatric disease) และนอนพักรักษาตัวอยู่ที่โรงพยาบาลไม่นาน ปัจจัยเหล่านี้มีความสัมพันธ์กับการแย่ลงของความสามารถในการทำกิจวัตรประจำวัน 4 ด้าน ได้แก่ การรับประทานอาหาร การแต่งตัว การใช้ห้องสุขาและการเดิน จากการศึกษาของ Fukuda และคณะ [12] เห็นได้ชัดว่า มีข้อจำกัดที่ไม่สามารถควบคุมปัจจัยกวนก่อนเริ่มต้นการศึกษาได้ เนื่องจากเป็นการศึกษาแบบย้อนหลัง ปัญหาที่เกิดจาก selection bias จึงส่งผลให้ผลลัพธ์ของการศึกษาคลาดเคลื่อนไป ซึ่งจะเห็นได้จากผลลัพธ์ก่อนและหลังทำ propensity score matching นั้นไม่เป็นไปในทิศทางเดียวกัน ดังนั้นการพิจารณาเลือกใช้วิธีการเชิงสถิติที่เหมาะสมกับข้อมูลจึงเป็นสิ่งที่สำคัญมากเพราะจะนำไปสู่การสรุปผลลัพธ์ทางคลินิกที่ถูกต้อง

ตัวอย่างที่ 4 งานวิจัยของ Lim และคณะ [13] เป็นการศึกษาแบบไปข้างหน้า (prospective study) มีวัตถุประสงค์เพื่อหาว่าผลการรักษาผู้ป่วยที่ทำผ่าตัดแก้ไขภาวะนิ้วหัวแม่เท้าเอียง (hallux valgus) ระหว่างผู้ป่วยเพศชายและเพศหญิงมีความแตกต่างกันหรือไม่ โดยข้อมูลพื้นฐานประกอบด้วย อายุ ดัชนีมวลกาย (กก./ม^2) คะแนนความปวด AOFAS scores และ SF-36 scores ผลการศึกษาพบว่า จากจำนวนผู้ป่วยทั้งหมด 438 ราย เป็นผู้ป่วยชาย จำนวน 26 (5.9%) ราย และผู้ป่วยหญิง จำนวน 412 (94.1%) ราย ก่อนทำ propensity score matching เห็นได้ชัดว่า กลุ่มตัวอย่างผู้ป่วยชายมีอายุเฉลี่ยเท่ากับ 49.8±17.5 ปี ซึ่งน้อยกว่ากลุ่มตัวอย่างผู้ป่วยหญิงที่มีอายุเฉลี่ยเท่ากับ

55.2±11.8 ปี อย่างมีนัยสำคัญทางสถิติ (p-value = 0.03) และ SF-36 subscale: SF-36 role physical ก่อนผ่าตัดของกลุ่มตัวอย่างผู้ป่วยชาย (37.5±43.7) มีค่าต่ำกว่ากลุ่มตัวอย่างผู้ป่วยหญิง (59.2±42.6) อย่างมีนัยสำคัญทางสถิติ (p-value = 0.01) แต่หลังจากทำการจับคู่ด้วย propensity score matching แบบ 1:2 จะเหลือจำนวนผู้ป่วยหญิงเพียง 52 ราย ผลจากการทดสอบความแตกต่างของข้อมูลพื้นฐานก่อนผ่าตัดรวมทั้งความรุนแรงของภาวะนิ้วหัวแม่เท้าเอียงผิดรูประหว่างกลุ่มตัวอย่างผู้ป่วยชายและกลุ่มตัวอย่างผู้ป่วยหญิงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value > 0.05) และผลการรักษาหลังผ่าตัดได้ 2 ปี ซึ่งประเมินจากค่าคะแนนความปวด AOFAS scores และ SF-36 scores ระหว่างกลุ่มตัวอย่างผู้ป่วยชายและกลุ่มตัวอย่างผู้ป่วยหญิงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value > 0.05) ยกเว้น SF-36 subscale: SF-36 general health ของกลุ่มตัวอย่างผู้ป่วยชาย (68.7±20.6) มีค่าต่ำกว่ากลุ่มตัวอย่างผู้ป่วยหญิง (79.3±17.8) อย่างมีนัยสำคัญทางสถิติ (p-value = 0.02)

ตัวอย่างที่ 5 งานวิจัยของ Guo และคณะ [14] เป็นการศึกษาเชิงสังเกต มีวัตถุประสงค์เพื่อประเมินผลการรักษา ภาวะแทรกซ้อนระหว่างการผ่าตัด และอัตราการรอดชีวิตในผู้ป่วยสูงอายุที่มีอายุมากกว่า 90 ปี ซึ่งมีกระดูกหักบริเวณ interthrochanteric และได้รับการรักษาด้วย intramedullay fixation เปรียบเทียบกับผู้ป่วยที่มีอายุน้อยกว่า 90 ปี ผลการศึกษาพบว่า จากจำนวนผู้ป่วยทั้งหมด 2,242 ราย เป็นผู้ป่วยที่มีอายุ < 90 ปี (กลุ่ม A) จำนวน 2,036 (90.8%) ราย และผู้ป่วยที่มีอายุ ≥ 90 ปี (กลุ่ม B) จำนวน 206 (9.2%) ราย ก่อนทำ propensity score matching เห็นได้ชัดว่า ที่ baseline characteristics มีอยู่ 7 ตัวแปรจากทั้งหมด 12 ตัวแปร ที่มีความแตกต่างกันระหว่างกลุ่ม A และ B อย่างมีนัยสำคัญทางสถิติ (p-value < 0.05) แต่หลังจากที่ทำการจับคู่ด้วย propensity score matching แบบ 1:1 จะเหลือผู้ป่วยในแต่ละกลุ่มเพียง 192 ราย ผลจากการทดสอบความแตกต่างของ baseline characteristics ทั้งหมด 12 ตัวแปร ระหว่างกลุ่ม A และ B พบว่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value > 0.05) และเมื่อเปรียบเทียบผลการรักษา ภาวะแทรกซ้อนระหว่างการผ่าตัด และอัตราการรอดชีวิตพบว่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (p-value > 0.05)

5. สรุปและข้อเสนอแนะ

ความเอนเอียงจากการเลือกตัวอย่างในการศึกษาเชิงสังเกตและการศึกษากึ่งทดลองอาจจะนำไปสู่การประมาณค่าหรือการสรุปผลการศึกษาที่คลาดเคลื่อนไปเนื่องจากหน่วยตัวอย่างแต่ละหน่วยมีลักษณะที่ไม่เหมือนกันเมื่อเริ่มต้นการศึกษา ทางเลือกหนึ่งที่สามารถนำมาใช้เพื่อแก้ปัญหาดังกล่าวนี้ คือ multivariable regression models แต่มีข้อจำกัดในเรื่องของขนาดตัวอย่าง (the number of events per covariates) ที่ต้องมีมากเพียงพอสำหรับการวิเคราะห์ แต่บางการศึกษามีตัวแปรกวนจำนวนมากส่งผลให้ขนาดตัวอย่างต่อตัวแปรร่วมมีน้อย (small sample size) ดังนั้น การนำ propensity score มาใช้เป็นเครื่องมือในการปรับความเอนเอียงจากการเลือกตัวอย่างจึงมีประโยชน์มากในการวิเคราะห์ข้อมูล ซึ่งวิธี propensity score นี้ไม่มีข้อจำกัดดังที่กล่าวมาทั้งหมดข้างต้น แม้ว่าวิธีนี้จะสามารถนำมาใช้จัดการตัวแปรกวนได้ แต่ก็ยังมีข้อด้อยตรงที่ไม่สามารถแสดงและวัดผลได้สำหรับปัจจัยกวนที่ไม่ทราบ (unobserved factors) ดังนั้น หากผู้วิจัยสามารถทำการศึกษาวิจัยแบบ RCT ได้ย่อมเป็นทางเลือกที่ดีกว่าเพราะไม่มีปัญหาเรื่องความเอนเอียงจากการเลือกตัวอย่าง

เอกสารอ้างอิง (References)

[1] Schulz, K. and Grimes, D. 2019. Essential concepts in clinical research. Randomised controlled trials and observational epidemiology. 2nd ed, Sydney: Elsevier, China.

[2] Hanzlik, S., Mahabir, R.C., Baynosa, R.C. and Khiabani, K.T. 2009. Levels of evidence in research published in The Journal of Bone and Joint Surgery (American Volume) over the last thirty years. *J Bone Joint Surg Am*, 91(2), 425-428.

[3] Cunningham, B.P., Harmsen, S., Kweon, C., Patterson, J., Waldrop, R., McLaren, A., and McLemore, R. 2013. Have levels of evidence improved the quality of orthopaedic research?. *Clin Orthop Relat Res*, 471(11), 3679-3686.

[4] Parsa, A. and Ebrahimzadeh, M. H. 2017. Level of Evidence in the Archivs of Bone and Joint Surgery Journal. *Arch Bone Jt Surg*, 5(6), 347-350.

[5] Rosenbaum, P.R. and Rubin, D.B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

[6] Yao, X.I., Wang, X., Speicher, P.J., Hwang, E.S., Cheng, P., Harpole, D.H., Berry, M.F., Schrag, D. and Pang, H.H. 2017. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. *J Natl Cancer Inst*, 109(8), djw323, doi: 10.1093/jnci/djw323 (1-9).

[7] Borah, B. J., Moriarty, J. P., Crown, W. H. and Doshi, J. A. 2014. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go?. *J Comp Eff Res*, 3(1), 63-78.

[8] Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*, 28(25), 3083-3107.

[9] Benedetto, U., Head, S. J., Angelini, G. D. and Blackstone, E. H. 2018. Statistical primer: propensity score matching and its alternatives. *Eur J Cardiothorac Surg*, 53(6), 1112-1117.

[10] Austin, P. C. 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, 46(3), 399-424.

[11] Rosenbaum, P. R. 1987. Model-Based Direct Adjustment. *Journal of the American Statistical Association*, 82(398), 387-394.

[12] Fukuda, T., Imai, S., Nakadera, M., Wagatsuma, Y. and Horiguchi, H. 2018. Postoperative daily living activities of geriatric patients administered general or spinal anesthesia for hip fracture surgery: A retrospective cohort study. *J Orthop Surg (Hong Kong)*, 26(1), 2309499017754106, doi: 10.1177/2309499017754106 (1-9).

[13] Lim, W. S. R., Liow, M. H. L., Rikhraj, I. S., Goh, G. S. and Koo, K. 2019. The effect of gender in hallux valgus surgery. A Propensity score matched study. *Foot Ankle Surg*, 25, 670-673.

[14] Guo, J., Wang, Z., Fu, M., Di, J., Zha, J., Liu, J., Zhang, G., Wang, Q., Chen, H., Tang, P., Hou, Z. and Zhang, Y. 2020. Super elderly patients with intertrochanteric fractures do not predict worse outcomes and higher mortality than elderly patients: a propensity score matched analysis. *Aging*, 12(13), 13583–13593.