วารสารวิชาการสาธารณสุข
ปีที่ 27 ฉบับที่ 2 มีนาคม - เมษายน 2561

Journal of Health Science
Vol. 27 No. 2, March - April 2018

บทความพิเศษ *Special Article*

เทคนิคการจัดการ Missing Data ในงานวิจัยด้านวิทยาศาสตร์สุขภาพและสาธารณสุข

ระพีพงศ์ สุพรรณไชยมาตย์ พ.บ., อ.ว. (เวชศาสตร์ป้องกัน, สาธารณสุขศาสตร์), Ph.D. (Public Health and Policy)*, **

** สำนักงานพัฒนานโยบายสุขภาพระหว่างประเทศ กระทรวงสาธารณสุข*

*** โรงพยาบาลบ้านไผ่ อำเภอบ้านไผ่ จังหวัดขอนแก่น*

การขาดหายไปของข้อมูล หรือ missing data เป็นปัญหาที่พบได้บ่อยในงานวิจัยทางการแพทย์และสาธารณสุข การมี missing data ทำให้ผลการวิเคราะห์ข้อมูลขาด statistical power และเสี่ยงต่อการเกิดอคติ การวิเคราะห์เฉพาะข้อมูลที่มีอยู่ (completed case analysis of observed data) ทำให้สูญเสียความสามารถการเป็นตัวแทนของประชากรตั้งต้น แม้เป็นปัญหาที่พบได้บ่อย แต่งานวิจัยในปัจจุบันมักละเลยที่จะรายงานการจัดการ missing data จากรายงานของ Horton และ Switzer พบว่ามีเพียงร้อยละ 8 ของงานวิจัยที่ตีพิมพ์ในวารสาร New England Journal of Medicine ในปี พ.ศ.2547-2548 ที่แสดงผลการจัดการ missing data[1]

Missing data มีสามรูปแบบ ได้แก่

1) Missing completely at random (MCAR)

2) Missing at random (MAR)

3) Missing not at random (MNAR)

โดยมีรายละเอียดของแต่ละประเภท[2] ดังนี้

MCAR หมายถึง การไม่มีความแตกต่างอย่างเป็นระบบระหว่าง missing data กับ observed data อาทิ ข้อมูลความดันโลหิตของผู้ป่วยขาดหายไปเนื่องจาก sphygmomanometer เสีย ในทางคณิตศาสตร์สามารถเขียน MCAR ได้เป็น

$$P(R=1| X,Y) = P(R=1)$$

ซึ่ง $P(R=1|X,Y)$ หมายถึง probability function ของการที่จะมี missing data ภายใต้เงื่อนไขของค่าตัวแปรต้นและตัวแปรตาม มีค่าเท่ากับ probability function ของการที่จะมี missing data โดยไม่กำหนดเงื่อนไขใด ๆ หรืออีกนัยหนึ่งคือ probability function ของการมี missing data เป็นอิสระจากทั้งตัวแปรต้นและตัวแปรตาม

MAR หมายถึง การมีความแตกต่างอย่างเป็นระบบระหว่าง missing data กับ observed data ซึ่งความแตกต่างนี้อธิบายได้ด้วยความแตกต่างของตัวแปรร่วม แต่เป็นอิสระจากตัวแปรตาม หรือก็คือ

$$P(R=1| X,Y) = P(R=1| X)$$

อาทิ ผู้วิจัยเก็บข้อมูลความดันโลหิตของผู้ป่วยนอกที่มารับการรักษาในเวลาราชการ ต่อมาพบว่า ข้อมูลความดันโลหิตของผู้ป่วยมักขาดหายไปในผู้ป่วยวัยทำงาน เนื่องจากผู้ป่วยกลุ่มนี้ไม่สามารถลางานเพื่อมาเข้ารับการตรวจที่โรงพยาบาลได้ เมื่อวิเคราะห์เฉพาะ observed data ความดันโลหิตของกลุ่มตัวอย่างอาจสูงกว่าความดันโลหิตของประชากร

MNAR หมายถึงการมีความแตกต่างอย่างเป็นระบบระหว่าง missing data กับ observed data ซึ่งอธิบายได้ด้วยความแตกต่างของตัวแปรตาม หรือก็คือ

$$P(R=1| X,Y) = P(R=1| X)$$

หรืออีกนัยหนึ่งหมายถึง ค่าของตัวแปรตามมีผลต่อการมีหรือไม่มี missing data ในระบบ อาทิ ผู้วิจัยพบว่าข้อมูลความดันโลหิตของผู้ป่วยส่วนหนึ่งขาดหายไป เนื่องจากผู้ป่วยที่มีความดันโลหิตสูงมากไม่แข็งแรงพอที่จะเดินทางมาโรงพยาบาล ดังนั้นเมื่อวิเคราะห์เฉพาะ observed data ความดันโลหิตของกลุ่มตัวอย่างจึงต่ำกว่าความดันโลหิตของประชากร

ปัญหาที่พบใน MCAR คือการสูญเสีย statistical power แต่การคำนวณเฉพาะ observed data ไม่ได้ก่อปัญหาในเชิงอคติมากนัก ขณะที่ MNAR เป็นปัญหาที่ก่ออคติมากที่สุด ซึ่งไม่อาจแก้ไขได้โดยเทคนิคทางสถิติทั่วไป ต้องอาศัยการปรับรูปแบบของงานวิจัยหรือเก็บข้อมูลใหม่ ปัญหาที่พบได้บ่อยที่สุดจึงเป็น MAR และในหลายกรณีผู้วิจัยไม่สามารถลงเก็บข้อมูลซ้ำได้ ผลลัพธ์ที่ได้จึงมีโอกาสเกิดอคติได้สูง จึงมีการพัฒนาเทคนิคทางสถิติเพื่อที่จะแทนค่า missing data เทคนิคเหล่านั้น ได้แก่

1. Single imputation คือ การเติม missing data ด้วยค่าคาดประมาณค่าใดค่าหนึ่ง ซึ่งการหาค่าคาดประมาณนั้น ทำได้หลายวิธี อาทิ

   i. Last observation carried forward (LOCF) คือ การแทนค่า missing data ด้วยค่า observed data ที่ติดกัน วิธีนี้มักใช้ในข้อมูลที่มีการวัดซ้ำ (repeated measurement) โดยตั้งอยู่บนสมมติฐานว่า ข้อมูลที่ขาดหายไปมีค่าใกล้เคียงกับข้อมูลสุดท้ายที่มีอยู่

   ii. Mean substitution (MS) คือ การแทนค่า missing data ด้วยค่าเฉลี่ยของ observed data วิธีนี้หากเป็น MAR มีข้อดี คือ ลดอคติ แต่มีผลเสียคือ (1) ทำให้ความแปรปรวน (uncertainty) ของตัวแปรตามลดลงและทำให้ไม่สามารถคงคุณลักษณะของประชากรได้ และ (2) ทำให้ความสัมพันธ์ (correlation) ระหว่างตัวแปรต้นและตัวแปรตามลดลง

   iii. Regression imputation (RI) คือ การแทนค่า missing data ด้วยค่าคาดการณ์ (predicted value) ที่ได้จากสมการถดถอยเชิงเส้น โดยใช้ตัวแปรต้นของสมการคือ ตัวแปรร่วม (confounders) ที่ปรากฏในชุดข้อมูลทั้งหมด ข้อดีคือ ลดอคติ แต่มีผลเสีย คือ ค่าคาดการณ์ไม่ได้รวม error term ของสมการเข้าไป ทำให้เกิดปัญหา over-fitting และทำให้ความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามสูงเกินจริง

2. Multiple imputation (MI) มีหลักการคล้าย RI แต่แทนที่จะแทนค่า missing data ด้วยชุดข้อมูลเดียว แต่แทนค่า missing data ด้วยข้อมูลหลายชุด (สมมติ 100 datasets) ซึ่งทั้ง 100 datasets นี้ได้จากการสุ่มค่าที่เป็นไปได้ทั้งหมด (sampling based on predictive distribution) ตามหลักสถิติของ Bayes จากนั้นจึงรวมค่าคาดการณ์จากทั้ง 100 datasets นี้เข้าด้วยกัน และค่าความแปรปรวนของค่าคาดการณ์คำนวณจากความแปรปรวนภายในแต่ละ dataset และความแปรปรวนระหว่าง dataset วิธีนี้มีข้อดีคือ ลดอคติและช่วยรักษา uncertainty ของข้อมูลได้ดีกว่าวิธี RI ข้อเสียคือ ใช้การคำนวณที่ซับซ้อน แต่ปัจจุบันมีการพัฒนาโปรแกรมสถิติที่สามารถคำนวณวิธีนี้ได้สะดวกขึ้น เช่น โปรแกรม R และ Stata

ทั้งนี้ยังมีการพัฒนาวิธีอื่น ๆที่คล้ายหรือต่อยอดจาก MI ซึ่งไม่ได้ลงรายละเอียดในที่นี้ อาทิ multiple imputation by chained equations (MICE) และ maximum likelihood estimation (ML)

สรุป

Missing data เป็นปัญหาสำคัญในงานวิจัยทางการแพทย์และสาธารณสุข การแก้ปัญหา missing data ผู้วิจัยควรพิจารณาว่า missing data ที่พบเป็นประเภทอะไร ถ้าเป็น MCAR มักไม่ทำให้ผลการศึกษามีอคติ ถ้าเป็น MNAR ต้องแก้ปัญหาด้วยการปรับระเบียบวิธีวิจัย ในทาง

ปฏิบัติปัญหาที่พบมากที่สุด คือ MAR ซึ่งสามารถแก้ไขด้วยวิธีทางสถิติ วิธีที่ก่ออคติน้อยที่สุด คือ RI และ MI แต่ MI มีข้อดีมากกว่าในการรักษาความแปรปรวนของข้อมูลได้ ทำให้คงคุณลักษณะของประชากรหลักได้

เอกสารอ้างอิง

1. Horton NJ, Switzer SS. Statistical Methods in the Journal. New England Journal of Medicine 2005;353:1977-9.
2. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.