

การศึกษาความผิดพลาดของการใช้สถิติสำหรับการวิจัยทางวิสัญญีวิทยา

วรารณ ไวกกุล พ.บ.*,

โรจนรินทร์ โกมลหิรัญ ค.ม.

ภัทราภรณ์ สอนคำมี วท.ม.

Abstract: Statistics in Anesthesia Research: a review of most common pitfalls

Waraporn Waikakul M.D. Rojnarin Komonhirun M.Ed. Patthrapon Sonkhammee M.Sc.

*Department of Anesthesiology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok 10400, Thailand

วิจัยการศึกษา

โรคติดเชื้อและวิทยาการระบาด

นับตั้งแต่กลางคริสต์ศตวรรษที่ 18 เป็นต้นมา บุคลากรทางการแพทย์สาขาวิสัญญีวิทยาได้มีการพัฒนาวิธีการระงับความรู้สึกแบบทันสมัย (Modern anesthesia) ขึ้น¹ ศาสตร์ด้านวิสัญญีได้รับการพัฒนาอย่างต่อเนื่องจนกระทั่งปี ค.ศ. 1922 ได้มีการจัดตั้งสมาคมชื่อว่า International Anesthesia Research Society (IARS) ขึ้นที่ประเทศสหรัฐอเมริกา ซึ่งเป็นองค์กรไม่แสวงผลกำไรแห่งแรกที่จัดตั้งขึ้นเพื่อส่งเสริมการศึกษาและวิจัยเกี่ยวกับศาสตร์ทางวิสัญญีสมาคมแห่งนี้ได้ผลิตรายการวิชาการด้านวิสัญญีวิทยาลับแรกชื่อว่า Anesthesia & Analgesia² ประกอบกับในราวปี ค.ศ. 1930 วารสารวิชาการทางแพทย์เริ่มมีการนำความน่าจะเป็นทางสถิติ (Statistical probability) เข้ามาใช้ในการพิจารณานำเสนอรายงาน³ ดังนั้นจึง

เป็นเวลา 80 ปีเศษแล้ว ที่ข้อมูลทางสถิติถูกนำมาใช้ในการศึกษาวิจัยเพื่อพัฒนาความรู้ทางด้านวิสัญญีอย่างเป็นรูปธรรม

การใช้สถิติในการวิจัยทั่วไปนั้นถูกมองว่าเป็นเรื่องสำคัญมากขึ้นตั้งแต่การแพทย์เชิงประจักษ์ (Evidence - based medicine; EBM) เริ่มมีความสำคัญเด่นชัดในปี ค.ศ. 1996. Sacket et al.⁴ ได้ให้ความหมายของการแพทย์เชิงประจักษ์ไว้ว่าเป็นกระบวนการการใช้หลักฐานที่ดีที่สุดเท่าที่มีอยู่ในปัจจุบันจากการทำวิจัยทางคลินิก เพื่อการดูแลรักษาผู้ป่วย (The conscientious and judicious use of current best evidence from clinical care research in the management of individual patients) การใช้เพียงความรู้ในตำราจึงไม่อาจคาดเดาผลการรักษาที่แน่นอนที่สุดได้อีกต่อไป กระบวนการวิจัยและสถิติได้รับความนิยมนำมาใช้ในการใช้เป็น “หลักฐาน” เพื่อ

ประกอบการวินิจฉัยโรคและการรักษาผู้ป่วย ดังนั้น ความถูกต้องของสถิติที่ใช้ในการวิจัยจึงเป็นเรื่องที่สำคัญยิ่งที่จะนำไปสู่ผลสรุปของการวิจัยที่ถูกต้องได้

การที่ผู้วิจัยจะส่งต้นฉบับงานวิจัยแต่ละเรื่อง ไปที่วารสารส่วนใหญ่ นั้น จะต้องได้รับการตรวจสอบ ด้วยระบบ Peer - review โดยกองบรรณาธิการของวารสารเพื่อป้องกันความผิดพลาดของเนื้อหาอยู่แล้ว แต่ถึงกระนั้นก็อาจมีความผิดพลาดหลงเหลืออยู่ ซึ่งส่วนหนึ่งอาจเป็นความผิดพลาดที่เกิดจากการใช้สถิติที่ไม่ถูกต้องด้วย Yim KH et al.⁵ ได้ให้ข้อมูลว่า บุคลากรทางการแพทย์ด้านวิสัญญีอาจยังมีความรู้ทางสถิติไม่เพียงพอที่จะวิเคราะห์ข้อมูลได้ด้วยวิธีที่ถูกต้อง และได้ทำการวิจัยเพื่อสำรวจความผิดพลาดที่เกิดจากการใช้สถิติเพื่อการวิจัยสำหรับผลงานวิจัยที่ตีพิมพ์ในวารสาร Korean Journal of Pain (KJP) ซึ่งเป็นการให้ข้อมูลป้อนกลับวารสารเพื่อที่จะนำไปพัฒนารูปแบบการวิเคราะห์ข้อมูลทางสถิติให้ถูกต้องต่อไป เขาได้รวบรวมผลงานวิจัยตีพิมพ์ซึ่งไม่รวม case reports และบทความจากบรรณาธิการ จากวารสาร Korean Journal of Pain (KJP) ฉบับตั้งแต่ปี ค.ศ. 2004 - 2008 และบันทึกข้อมูลเกี่ยวกับสถิติที่ใช้และความผิดพลาดที่เกิดขึ้นเอาไว้ และได้แบ่งรูปแบบของความผิดพลาดที่เกิดจากการใช้สถิติไว้ได้อย่างน่าสนใจ โดยแบ่งเป็น 2 ประเภท คือ ความผิดพลาดที่เกิดจากการละเลย (Errors of omission) และความผิดพลาดที่เกิดจากการใช้งาน (Errors of commission) ความผิดพลาดที่เกิดจากการละเลย เช่น การไม่อธิบายถึงข้อมูลเบื้องต้น หรือค่าสถิติที่ใช้ ส่วนความผิดพลาดที่เกิดจากการใช้งาน เช่น การใช้สถิติแบบอิงพารามิเตอร์ในข้อมูลที่ต้องใช้สถิติแบบไม่อิงพารามิเตอร์ หรือการใช้ Independent t - test มาวิเคราะห์ความแตกต่างของข้อมูลสองกลุ่ม ทั้งที่เป็นข้อมูลที่มีการวัดก่อน - หลังในคนกลุ่มเดียวกัน เป็นต้น ในบทความ

นี้ผู้เขียนขอแบ่งประเภทของความผิดพลาดของการใช้สถิติในงานวิจัยที่เคยพบ ในรูปแบบเดียวกันดังนี้

ความผิดพลาดที่เกิดจากการละเลย (Errors of omission)

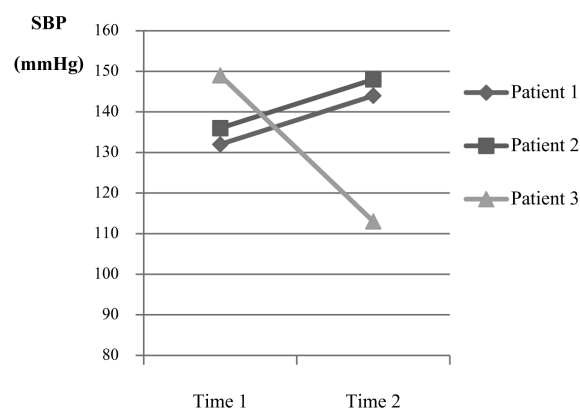
การวิเคราะห์ข้อมูลในแต่ละครั้งนั้น นอกจากเราจำเป็นต้องเชื่อมั่นในผลการวิเคราะห์แล้ว ยังจำเป็นต้องต้องทำให้ผู้อ่านมั่นใจได้ด้วยว่าสถิติที่เรานำมาใช้วิเคราะห์ข้อมูลนั้นถูกต้องและเหมาะสม การเขียนสรุปข้อมูลไปเลยโดยไม่ใช้ค่าสถิติที่จำเป็น ยืนยันทั้ง ๆ ที่ควรจะใช้มัน ถึงแม้จะทำให้รวดเร็ว แต่ก็ยังเป็นสิ่งที่ผิด และดูเหมือนจะเป็นการเจตนามุสา ดังมีคำกล่าวสนับสนุนของ Mosteller ที่ว่า “ในขณะที่มันง่ายที่จะโกหกด้วยค่าสถิติ แต่มันง่ายยิ่งกว่าที่จะโกหกโดยไม่ใช้มัน” (While it is easy to lie with statistics, it is even easier to lie without them)⁶ ผู้เขียนได้รวบรวมความผิดพลาดที่เกิดจากการละเลยในการใช้สถิติสำหรับงานวิจัยทางวิสัญญีวิทยาที่เคยพบไว้ 5 ข้อ พร้อมทั้งรายละเอียดและตัวอย่าง ดังนี้

1. รายงานผลเฉพาะค่าเฉลี่ยรวมต่อกลุ่ม ในข้อมูลแบบจับคู่ (Paired data)

ข้อมูลชุดใด ๆ ที่มาจากผู้ป่วยคนเดียวกัน เรียกว่าข้อมูลจับคู่ หรือข้อมูลที่ไม่เป็นอิสระจากกัน ในแต่ละกลุ่มของผู้ป่วยที่บันทึกข้อมูลเดิมแต่วัดซ้ำหลายครั้งนั้น ก็อาจมีค่าสถิติที่เปลี่ยนแปลงแตกต่างกันไประหว่างช่วงเวลาได้ หรือแม้แต่ค่าสถิติภายในกลุ่มของผู้ป่วยใด ๆ เองนั้นก็อาจแปรผันไประหว่างช่วงเวลาได้เช่นกัน อย่างไรก็ตามความแปรผันเหล่านี้ถูกซ่อนไว้ด้วยการรายงานเฉพาะค่าเฉลี่ยของกลุ่ม (group mean) ผู้อ่านรายงานจึงไม่อาจพบความขัดแย้งใด ๆ ระหว่างเวลาที่ไม่ได้รายงานตัวเลขไว้ เมื่อพิจารณาตัวอย่างในแผนภาพที่ 1 จะพบว่าข้อมูลนี้สามารถรายงานได้ทั้ง ‘มีการลดลงของค่าเฉลี่ย (mean) ของความดันเลือดซิสโตลิกจาก Time 1 ไป

Time 2' หรือ 'มีการเพิ่มขึ้นของค่าความดันเลือดซิสโตลิกที่วัด ในผู้ป่วยจำนวนสองในสามคน ในขณะที่มีการลดลงของค่าดังกล่าวในผู้ป่วยเพียงหนึ่งคน' จะเห็นว่าการรายงานผลทั้งสองรูปแบบนี้ถูกต้องในทางปฏิบัติ แต่สามารถนำไปสู่การสรุปผลที่แตกต่างกัน ดังนั้นในการรายงานผลการศึกษาคจากข้อมูลแบบจับคู่ ผู้วิจัยจึงควรรายงานค่าสถิติของช่วงเวลานั้น ๆ อีกทั้งควรใช้สถิติทดสอบที่ถูกต้องสำหรับข้อมูลแบบจับคู่ ในการวิเคราะห์ความแตกต่างด้วย

	Patient 1	Patient 2	Patient 3
Time 1	132	136	149
Time 2	144	148	113
Differences	12	12	-36
Average of differences	-4		



แผนภาพที่ 1 การนำเสนอข้อมูลแบบจับคู่ (Paired data) ของความดันเลือดซิสโตลิก ที่มีการวัดซ้ำสองครั้ง ในผู้ป่วยจำนวน 3 ราย

2. รายงานเฉพาะค่า p (p - value) ในผลการวิจัย

ค่า p หรือค่าความน่าจะเป็น (probability) มักจะถูกตีความผิดอยู่เสมอ และถึงแม้บางครั้งตีความถูกแล้วก็ยังมิชอบเขตการตีความที่ต้องนำมาพิจารณา

ในการรายงานผลการวิเคราะห์ทางสถิติที่ดีนั้นควร มีการรายงานวิธีการวัดหรือการประมาณค่า มีการรายงานค่าทางสถิติเช่นค่าเฉลี่ย รวมถึงรายงานความแม่นยำ (precision) ของค่านั้น ๆ ด้วย ขอให้พิจารณาตัวอย่างการรายงานข้อมูลทางสถิติเหล่านี้

ก. 'ความปวดของผู้ป่วยลดลงหลังจากให้ยา celecoxib อย่างมีนัยสำคัญทางสถิติ' ประโยคนี้ขาดการกล่าวถึงปริมาณการลดลงของคะแนนความปวด วิธีการวัดความปวด และค่าแสดงระดับนัยสำคัญทางสถิติ ในกรณีนี้ผู้อ่านอาจสรุปความหมายเองว่างานวิจัยนี้สนับสนุนให้ใช้ celecoxib

ข. 'คะแนนความปวดวัดโดย Visual analog scale: VAS for pain ของผู้ป่วยลดลงหลังจากให้ยา celecoxib อย่างมีนัยสำคัญทางสถิติ (p < 0.05)' ประโยคนี้ไม่ได้กล่าวถึงปริมาณการลดลงของคะแนนความปวด และค่า p ที่ได้ อาจมีค่าถึง 0.049 ซึ่งใกล้เคียง 0.05 อย่างมากก็เป็นได้ ข้อสรุปนี้จึงยังไม่อาจให้ความหมายใด ๆ มากพอในการนำไปใช้ได้

ค. 'คะแนนความปวดวัดโดย Visual analog scale: VAS for pain ของผู้ป่วยลดลงหลังจากให้ยา celecoxib โดยเฉลี่ย 4.1 อย่างมีนัยสำคัญทางสถิติ (p = 0.023)' รูปแบบนี้นิยมเขียนกันโดยทั่วไป ซึ่งให้รายละเอียดของข้อมูลที่วัดครบทั้งสองกลุ่ม แต่อย่างไรก็ตามยังขาดการกล่าวถึงความแม่นยำ (precision) อันจะเป็นการบ่งบอกถึงคุณค่าของข้อมูลที่จะนำไปใช้

ง. 'คะแนนความปวดวัดโดย Visual analog scale: VAS for pain ของผู้ป่วยลดลงหลังจากให้ยา celecoxib โดยเฉลี่ย 4.1 อย่างมีนัยสำคัญทางสถิติ (95% CI 0.03 - 8.2; p = 0.023)' ค่า 95% CI หรือค่าช่วงความเชื่อมั่น 95% ที่แสดงนี้บ่งบอกว่า หากใช้ celecoxib ในผู้ป่วย 100 คน ที่มีลักษณะเหมือนในรายงานแล้ว ผู้ป่วยจำนวน 95 คนจาก 100 คน จะมีคะแนนความปวดลดลงเป็นไปได้ตั้งแต่ 0.03 ไป

จนถึง 8.2 คะแนน ซึ่งแน่นอนว่าคะแนนความปวดเมื่อลดลงเพียง 0.03 นั้นไม่นับสำคัญทางคลินิก ดังนั้นหากการศึกษาภายหลังจะอ้างอิงตัวเลขผลการศึกษานี้แล้ว ผลการศึกษานี้จะมีความสำคัญน้อยกว่าการศึกษาที่ได้คะแนนความปวดช่วง 95% CI ทั้งหมดอยู่ในปริมาณที่มีนัยสำคัญทางคลินิกอย่างแน่นอน

3. ไม่มีหรือไม่กล่าวถึงวิธีการคำนวณขนาดของตัวอย่างที่ใช้ในการวิจัย

เนื่องด้วยประชากรเป้าหมายนั้นมีจำนวนมาก การเก็บข้อมูลจากประชากรทั้งหมดจึงใช้เวลานาน และสิ้นเปลืองทรัพยากร ดังนั้นนักวิจัยจึงควรใช้การกำหนดขนาดของกลุ่มตัวอย่างที่เหมาะสม ซึ่งจำนวนนี้ควรเพียงพอที่จะเป็นตัวแทนของประชากร, เพียงพอกับงบประมาณที่ได้รับ, คำนึงถึงความเสี่ยงของผู้ป่วยและมากพอในทางสถิติเพื่อให้ผลการวิจัยถูกต้องตามเป้าหมายของความมีนัยสำคัญทางสถิติ (statistical significant) หากว่าขนาดตัวอย่างที่ไม่เพียงพอจะนำมาอ้างอิงทางสถิติ (underpowered) แล้วนั้น ผลการวิจัยจะขาดหลักฐานและไม่สามารถสรุปผลแน่ชัดจากสถิติที่ใช้ได้ ดังคำกล่าวว่า ‘The absence of evidence’ is not the ‘evidence of absence’ ขณะเดียวกันหากขนาดตัวอย่างมีมากเกินไปก็จะขัดต่อหลักจริยธรรมการวิจัยในมนุษย์ที่ว่า ประโยชน์ที่ได้จากงานวิจัยนั้นไม่คุ้มค่ากับความเสี่ยงที่ผู้ป่วยจะได้รับหรือทรัพยากรที่จะต้องเสียไปด้วย โดยปกติแล้วผู้วิจัยคำนวณขนาดของตัวอย่างที่เพียงพอโดยพิจารณาจากคำถามวิจัย (research question) และรูปแบบการวิจัย (study design) โดยใช้สูตรคำนวณซึ่งค้นหาได้ไม่ยากจากแหล่งอ้างอิงที่เกี่ยวข้อง อย่างไรก็ตามความคุ้มค่าของจำนวนตัวอย่างที่ได้เทียบกับทรัพยากรอื่น ๆ นั้นยังเป็นเรื่องที่ได้แย้งตกลงกันได้เป็นราย ๆ ไป ผู้เขียน

ขอยกตัวอย่างการพิจารณาขนาดตัวอย่างสำหรับการวิจัยครั้งนี้ ‘ในการวิจัยเพื่อเปรียบเทียบประสิทธิภาพของการระงับปวดซึ่งวัดจากปริมาณความต้องการมอร์ฟีน (Morphine requirement) ใน 24 ชั่วโมงหลังผ่าตัดของผู้ป่วยสองกลุ่ม กลุ่มแรกได้รับ Ketamine และกลุ่มที่สองได้รับ Thiopentone นั้น จากการศึกษาก่อนหน้านี้ในกลุ่มอื่นพบว่าปริมาณความต้องการมอร์ฟีนรวม (Morphine consumption) ใน 24 ชั่วโมงหลังผ่าตัดของสองกลุ่มนี้แตกต่างกันร้อยละ 40 เมื่อใช้ค่าส่วนเบี่ยงเบนมาตรฐานของปริมาณความต้องการมอร์ฟีนรวมจากฐานข้อมูลของหน่วยระงับปวด ณ โรงพยาบาลนั้น ๆ คือ 25 mg แล้ว กลุ่มตัวอย่างกลุ่มละ 21 คน เพียงพอที่จะนำมาวิเคราะห์ความแตกต่าง เมื่อกำหนดให้ power of test เป็นร้อยละ 80 และ type I error rate เป็น 0.05’

4. ไม่มีการสุ่ม (randomization)

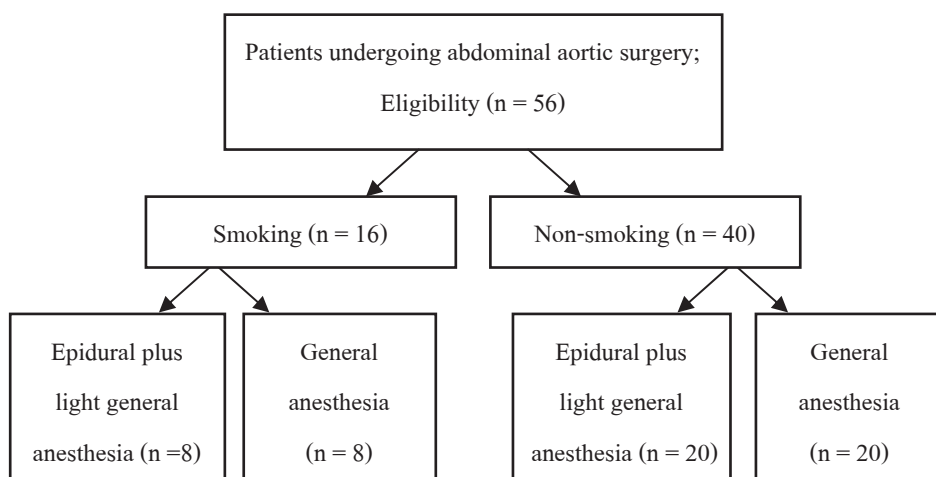
การทดลองทางคลินิกส่วนใหญ่เน้นจุดประสงค์หลักคือการเปรียบเทียบผลของการรักษา หากว่ากลุ่มทดลองมีลักษณะอื่นใดที่แตกต่างกันนอกเหนือจากความแตกต่างที่มาจากวิธีการรักษาแล้ว การเปรียบเทียบใด ๆ ระหว่างกลุ่มอาจจะเกิดความเอนเอียง (bias) ขึ้นได้ เมื่อปัจจัยใด ๆ ของผู้ป่วยมีความสัมพันธ์ทางสถิติกับค่าที่ต้องการศึกษา ดังนั้นการทำให้กลุ่มทดลองมีลักษณะทั้งที่ทราบและไม่ทราบที่คล้ายคลึงกันจึงมีความสำคัญ เนื่องจากปัจจัยที่ทราบได้นั้นผู้วิจัยอาจสามารถปรับค่าได้จากวิธีการทางสถิติ แต่ปัจจัยที่ไม่ทราบนั้นทำไม่ได้ จึงเป็นที่มาของวิธีการสุ่มตัวอย่าง (randomization) เพื่อทำให้เกิดความสมดุล (balance) ของลักษณะต่าง ๆ ภายในผู้ป่วยระหว่างกลุ่มทดลองได้

การวิจัยคลินิกทางด้านวิสัญญีบางเรื่องใช้กลุ่มตัวอย่างในการเปรียบเทียบจำนวนน้อย แต่ใน

กลุ่มอาจมี prognostic factors หรือตัวแปรร่วม (covariates) อยู่เป็นจำนวนมาก การสุ่มตัวอย่างเข้ากลุ่มจึงควรปรับให้เหมาะสมเพื่อที่จะทำให้ตัวอย่างแต่ละกลุ่มมีลักษณะใกล้เคียงกันมากที่สุดในการหาผลของการรักษาที่ได้ การสุ่มแบบหนึ่งที่นิยมใช้คือการสุ่มแบบแบ่งชั้น (Stratified random sampling) ซึ่งเป็นวิธีที่มีประโยชน์มากที่จะลดอิทธิพลของตัวแปรกวน (confounding factor) โดยการทำให้ตัวแปรกวนที่ทราบนั้น กลายเป็นเกณฑ์ของการแบ่งกลุ่ม ตัวแปรสำคัญที่มักนำมาสุ่มแบบแบ่งชั้นก็คือ เพศ อายุผู้ป่วย ความเสี่ยงต่าง ๆ การสูบบุหรี่ การติดเครื่องดมแอลกอฮอล์ (แล้วแต่กรณี) หรือหากเป็นการศึกษาแบบสหสถาบัน (multicenter trial) ก็แบ่งด้วยสถาบันนั่นเอง ดังจะเห็นจากตัวอย่างในแผนภาพที่ 2 ที่แสดงถึงการแบ่งการมีอยู่และไม่มี

ตัวแปรกวน (การสูบบุหรี่และไม่สูบ) ของผู้ป่วยออกเป็นบล็อก (block) และแต่ละบล็อกก็จะถูกสุ่มเข้ากลุ่ม (allocation) แยกกัน ดังนั้นจำนวนตัวอย่างที่เท่ากันของแต่ละกลุ่มก็จะได้รับอิทธิพลของตัวแปรกวนที่เท่า ๆ กันด้วย

แผนภาพที่ 2 ในการศึกษาผลของการทำ Epidural anesthesia เทียบกับ General anesthesia ที่มีต่อการเกิดภาวะแทรกซ้อนของผู้ป่วยที่ได้รับการผ่าตัด Abdominal aortic surgery นั้น เมื่อพบว่าสถานะการสูบบุหรี่ในปัจจุบันเป็นตัวแปรกวน ผู้วิจัยจึงกระจายให้กลุ่ม Epidural anesthesia และกลุ่ม General anesthesia ได้รับจำนวนผู้ป่วยที่สูบบุหรี่และไม่สูบบุหรี่จำนวนเท่า ๆ กัน รวมเป็นกลุ่มละ 28 คนตามสัดส่วน



5. ไม่ทดสอบความเป็นปกติของโค้งการแจกแจงข้อมูลแบบต่อเนื่อง (continuous data)

ในการทดสอบทางสถิติกับข้อมูลแบบต่อเนื่องนั้น การเลือกใช้สถิติทดสอบเป็นสิ่งที่สำคัญมาก การทดสอบทางสถิติในข้อมูลแบบต่อเนื่องนั้นมีสถิติที่ใช้อยู่หลากหลาย ขึ้นอยู่กับจำนวนตัวแปร

จำนวนตัวอย่างและลักษณะการแจกแจงแบบเป็นโค้งปกติของข้อมูล ด้วยเหตุนี้ลักษณะการแจกแจงแบบเป็นโค้งปกติของข้อมูลจึงเป็นสิ่งจำเป็นอย่างหนึ่งที่จะต้องนำมาพิจารณา

การทดสอบความเป็นโค้งปกติทำได้หลายวิธีการทดสอบ Goodness of fit ก็เป็นอีกวิธีหนึ่งที่นิยม

ใช้ทดสอบ สถิติทดสอบที่นิยมใช้ก็คือ Kolmogorov-Smirnov test แต่สำหรับข้อมูลที่มีจำนวนไม่เกิน 50 จะใช้ Shapiro - Wilk test แทน ซึ่งหากผลออกมาได้ค่า p มากกว่าหรือเท่ากับ 0.05 แล้ว ผู้วิจัยจะใช้สถิติทดสอบแบบอิงพารามิเตอร์ แต่หากได้ค่า p น้อยกว่า 0.05 แล้ว ผู้วิจัยควรใช้การแปลงข้อมูล (data transformation) หรือใช้สถิติทดสอบแบบไม่อิงพารามิเตอร์แทน

โดยปกติแล้วการทดสอบข้อมูลเชิงตัวเลข (Numerical data) จะใช้สถิติแบบอิงพารามิเตอร์ ส่วนข้อมูลอันดับ (ordinal data) จึงจะใช้สถิติแบบไม่อิงพารามิเตอร์ แต่รูปแบบการนำเสนอข้อมูลนั้นมีประเด็นต้องพิจารณาคือ ข้อมูลบางประเภทนั้นมีลักษณะการแจกแจงที่ต่อเนื่องในทางทฤษฎี (theoretical continuous) เช่น คะแนนความปวด คะแนนความพึงพอใจ เป็นต้น ข้อมูลเหล่านี้ถึงแม้จะเป็นข้อมูลอันดับ แต่การนำเสนอก็ควรนำเสนอในลักษณะของข้อมูลต่อเนื่องด้วย โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีขนาดใหญ่⁷

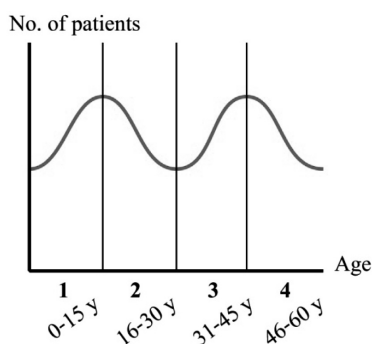
ความผิดพลาดที่เกิดจากการใช้งาน (Errors of commission)

ผลการวิจัยที่ตีพิมพ์ออกมานั้นจะมีบุคลากรทางการแพทย์ซึ่งอ่านและนำข้อมูลไปใช้ในการตัดสินใจรักษาผู้ป่วยหรือนำไปอ้างอิงเพื่อศึกษาต่อยอด ดังนั้นหากผู้วิจัยคำนวณข้อมูลทางสถิติด้วยวิธีที่ไม่ถูกต้องแล้ว ก็อาจจะทำให้ผลการวิจัยไม่ตรงกับสิ่งที่เกิดขึ้นจริงๆ ได้ ผู้เขียนได้รวบรวมความผิดพลาดที่เกิดจากการใช้สถิติสำหรับงานวิจัยทางสัตวณูวิทยาที่เคยพบไว้ 7 ข้อ ดังนี้

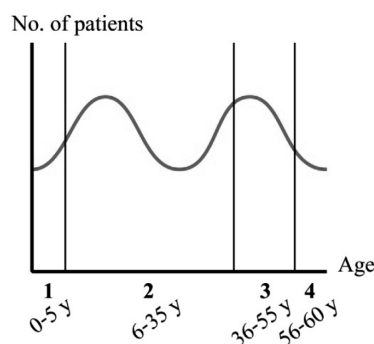
1. แบ่งกลุ่มข้อมูลต่อเนื่อง (continuous data) เป็นข้อมูลอันดับ (ordinal data) โดยมีได้อธิบายถึงเหตุผลหรือวิธีการแบ่งกลุ่ม

ผู้วิจัยบางท่านอาจทำให้การวิเคราะห์ข้อมูลแบบต่อเนื่องมีความง่ายขึ้นโดยการแบ่งกลุ่มข้อมูลต่อเนื่องนั้นให้เป็นข้อมูลอันดับ เช่น ดัชนีมวลกาย ซึ่งมีหน่วยวัดเป็น กก./ม.² ก็อาจแบ่งกลุ่มข้อมูลเป็นกลุ่มผอมมาก ผอม ปกติ อ้วน หรืออ้วนมาก เป็นต้น การลดระดับของการวัดแบบนี้จะลดความแม่นยำของค่าที่วัดแน่นอน เนื่องจากการลดความแปรปรวน (variability) ของข้อมูล ผู้วิจัยควรให้เหตุผลว่าเหตุใดจึงลดระดับของการวัด มีการใช้ขอบเขตใดในการแบ่งกลุ่ม และผู้วิจัยมีการป้องกันความเอนเอียงในการแบ่งกลุ่มที่อาจเกิดขึ้นอย่างไร

ในแผนภาพที่ 3 แสดงถึงความเอนเอียงในการแบ่งกลุ่มช่วงอายุ (Age range) เพื่อที่จะสรุปผลการวิจัยในเชิงความสัมพันธ์ระหว่างอายุกับปริมาณความต้องการมอร์ฟีน (Morphine requirement) ใน 24 ชม. หลังผ่าตัด อันเนื่องการแบ่งกลุ่มช่วงอายุแบบแผนภาพที่ 3.1 นั้นกำหนดให้ช่วงแต่ละช่วงมีความกว้าง (range) เท่า ๆ กัน จาก Trend line แสดงให้เห็นว่าปริมาณข้อมูลดูเหมือนจะมีการกระจายอยู่เท่า ๆ กัน ส่วนในแผนภาพที่ 3.2 นั้นกำหนดให้ช่วงอายุที่ 2 มีความกว้างมากกว่าช่วงอื่น ส่งผลให้ข้อมูลส่วนใหญ่มารวมตัวกันตกอยู่ในช่วงที่ 2 ซึ่งดูเหมือนว่าจงใจจะทำเพื่อประโยชน์ในทางใดทางหนึ่ง ดังนั้นในทางที่ถูกต้องแล้วควรแบ่งช่วงเท่า ๆ กัน หรือแบ่งช่วงตามขอบเขตทางทฤษฎีที่มีหลักฐานอ้างอิง



(3.1) กำหนดให้ช่วงแต่ละช่วงมีความกว้าง (range) เท่า ๆ กัน



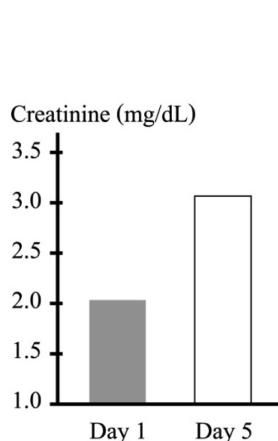
(3.2) กำหนดให้ช่วงแต่ละช่วงมีความกว้าง (range) ไม่เท่ากัน

แผนภาพที่ 3 ความเอนเอียงในการแบ่งกลุ่มช่วงอายุ (Age range)

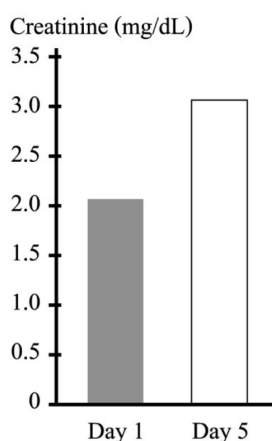
2. เลือกใช้การนำเสนอแผนภาพหรือกราฟที่สื่อความหมายไม่ตรงกับข้อมูลที่จะนำเสนอ

มนุษย์เราจดจำภาพที่มองเห็นได้มากกว่าตัวอักษรหรือตัวเลข เพราะเหตุนี้ผู้วิจัยจึงควรให้ความสำคัญกับการเลือกนำเสนอแผนภาพหรือกราฟ (figure, graph) ให้ตรงกับข้อมูล ตัวอย่างเช่น ในแผนภาพที่ 4.1 ซึ่งแสดงค่า creatinine ในเลือดผู้ป่วยวิกฤต กราฟที่ปรากฏในภาพนี้อาจชี้นำไปได้ว่าในวันที่ 1 (Day 1) ผู้ป่วยมีค่า creatinine ในเลือดน้อยกว่าวันที่ 5 (Day 5) อยู่ครึ่งต่อครึ่ง อย่างไรก็ตาม

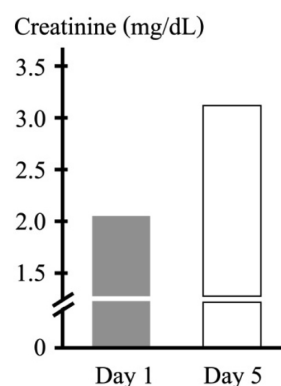
นั่นคือการสรุปที่ผิดพลาดเนื่องจากแท่งของกราฟไม่ได้มีค่าเริ่มต้นที่ 0 ส่วนในแผนภาพที่ 4.2 ซึ่งมีค่าเริ่มต้นที่ 0 นั้นมองคล้ายว่าในวันที่ 1 ผู้ป่วยมีค่า creatinine สองในสามเท่าของวันที่ 5 เลยทีเดียว ซึ่งหมายความว่าค่าที่ได้ไม่ได้เพิ่มขึ้นครึ่งต่อครึ่ง ทางที่ดีนั้นหากไม่ต้องการให้แท่งกราฟสูงเกินไปก็ควรให้แท่งกราฟมีการแยกส่วนกันดังในแผนภาพที่ 4.3 เพื่อแสดงให้เห็นอย่างชัดเจนว่าแกนของกราฟมีการขยับระยะและเริ่มต้นที่ค่า 0



(4.1) แกนของกราฟไม่ได้เริ่มต้นที่ค่า 0



(4.2) แกนของกราฟเริ่มต้นที่ค่า 0

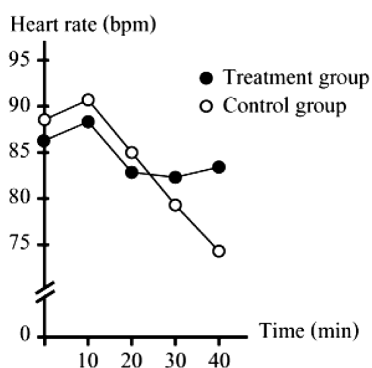


(4.3) แกนของกราฟเริ่มต้นที่ค่า 0 แต่มีการแยกส่วนเพื่อย่นระยะมิให้แท่งกราฟสูงเกินไป

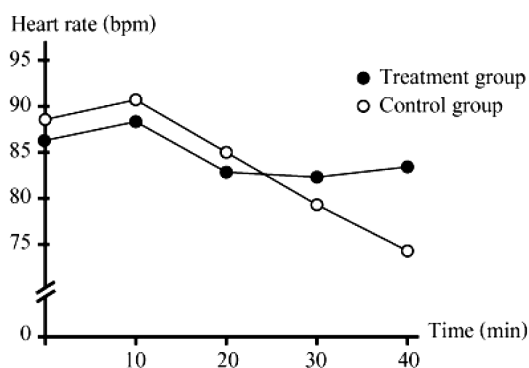
แผนภาพที่ 4 ตัวอย่างการนำเสนอแผนภาพหรือกราฟจากแกนหลายรูปแบบอาจส่งผลให้การสื่อความหมายคลาดเคลื่อน

นอกจากนี้ยังมีปัญหาเรื่องการที่แกนกราฟมีความกว้างและแคบต่างกัน กราฟที่ถูกบีบในแกนใดแกนหนึ่งจนเกินไปเนื่องจากพื้นที่แคบหรือสาเหตุอื่น ๆ นั้น อาจทำให้เส้นกราฟมีความชัน (slope) ที่แตกต่างกันได้ ส่งผลต่อภาพความสัมพันธ์ระหว่างกราฟสองเส้นที่คนดูด้วยสายตา ถึงแม้ว่าจะเป็น

ข้อมูลชุด เดียวกันหรือใกล้เคียงกัน ยกตัวอย่างเช่น กราฟแสดงอัตราการเต้นของหัวใจ (Heart rate) ของผู้ป่วยที่วัดตามคาบเวลา 10, 20, 30 และ 40 นาทีหลังเริ่มผ่าตัด ของผู้ป่วยสองกลุ่มทดลอง ดังแผนภาพที่ 5



(5.1) กราฟเส้นตรง ทรงแคบ



(5.2) กราฟเส้นตรง ทรงกว้าง

แผนภาพที่ 5 ตัวอย่างการนำเสนอกราฟเส้นตรงที่ได้จากข้อมูลชุดเดียวกัน แต่พื้นที่กราฟมีความกว้างแคบต่างกัน ซึ่งอาจสื่อความหมายได้ต่างกัน

ไม่มีความชัดเจนระหว่างประสิทธิผล (effectiveness) กับประสิทธิภาพ (efficacy) ในงานวิจัย

การศึกษาเชิงอธิบาย (Explanatory study) นั้นส่วนใหญ่จะมุ่งหาประสิทธิภาพ (efficacy) ของกระบวนการหรือวิธีการเพื่อให้ได้ข้อมูลจำเพาะของสิ่งหรือโรคต่างๆ การศึกษาข้อมูลประเภทนี้ทำได้ดีที่สุดเมื่อสภาพแวดล้อมหรือสถานที่ศึกษาถูกคัดเลือกหรือจำกัดอยู่ในรูปแบบที่แน่นอน ข้อมูลที่ได้จะมีความแม่นยำที่จะตอบคำถามวิจัยที่เฉพาะ แต่อาจนำไปประยุกต์ใช้ในชีวิตจริงไม่ได้ เพราะในความเป็นจริงสภาพแวดล้อมไม่ได้ถูกจำกัดไว้แบบนั้น การศึกษาเชิงสัมฤทธิ์ผลนิยม (Pragmatic study)

ส่วนใหญ่เป็นไปเพื่อการตัดสินใจจากผลลัพธ์หรือประสิทธิผล (effectiveness) ที่เกิดขึ้น การศึกษานี้จะจัดขึ้นในสภาพการณ์และเงื่อนไขปกติ ซึ่งอาจถูกรบกวนด้วยปัจจัยหลากหลายและอาจลดทอนความแม่นยำของการศึกษาลง แต่จะได้กลุ่มศึกษาที่มีลักษณะหลากหลายและสามารถนำค่าที่ได้ไปประยุกต์ใช้จริงได้อย่างดี มีงานวิจัยจำนวนมากที่ทำการศึกษาทั้งสองรูปแบบ แต่กระนั้นควรคำนึงถึงข้อมูลของการวิจัยที่ได้มาด้วยว่าควรนำไปบรรยายและสรุปความอย่างไร ตัวอย่างลักษณะของการศึกษาทั้งสองประเภทแสดงในตารางที่ 1

Table 1 ตัวอย่างความแตกต่างของลักษณะของการศึกษาทั้งสองประเภท ทั้งเชิงอรรถอธิบายและสัมฤทธิ์ผล
นิยม

	เชิงอรรถอธิบาย (Explanatory study)	เชิงสัมฤทธิ์ผลนิยม (Pragmatic study)
Eligibility	Men and women aged ≥ 18 years with a diagnosis of chronic low-back pain (≥ 10 weeks duration), required regular use of analgesics (≥ 4 days/week), experienced moderate to severe pain at baseline visit (NRS-pain score of ≥ 4)	Men and women aged ≥ 18 years with a diagnosis of chronic low-back pain
Design	Blinded and/or Placebo controlled	Blinded and/or Placebo controlled
Treatment	A non - steroidal anti - inflammatory drug vs Tramadol, controlled by researcher	A non - steroidal anti - inflammatory drug vs Tramadol (and other analgesics), controlled by patients
Outcomes	Change in NRS score after 2 hours	Lower incidence of treatment emergent AEs after 3 months follow up
Relevance to practice	Indirect: Tramadol has better analgesic potency than the non - steroidal anti - inflammatory drug	Direct: Tramadol is a more effective medication for chronic low - back pain than the non -steroidal anti - inflammatory drug but equally safe

ทั้งนี้ผู้วิจัยอาจเห็นว่าการที่จะพิจารณาว่างานวิจัยของตนเองนั้นควรออกแบบในเชิงอรรถอธิบายหรือสัมฤทธิ์ผลนิยมนั้นเป็นเรื่องยาก ในปัจจุบันมีผู้ค้นคว้าและจัดทำเกณฑ์จำแนกงานวิจัยในกรณีนี้อยู่จำนวนหนึ่ง ตัวอย่างเช่น Pragmatic – explanatory continuum indicator summary (PRECIS)⁸ ซึ่งเป็นเครื่องมือที่ช่วยให้ผู้วิจัยประเมินว่างานวิจัยของตนเองนั้นควรออกแบบไปในแนวทางใด โดยพิจารณาจากจุดมุ่งหมายของงานวิจัย (study purpose)

3. ใช้ค่าสถิติพรรณนาไม่ถูกต้อง

ค่าสถิติพรรณนา (descriptive statistics) ที่ใช้เป็นประจำในข้อมูลแบบต่อเนื่อง (continuous data)

ก็คือค่าเฉลี่ย (mean) และส่วนเบี่ยงเบนมาตรฐาน (standard deviation; SD) อันที่จริงแล้วค่าเหล่านี้ใช้อธิบายได้แค่ข้อมูลที่มีการแจกแจงแบบเป็นโค้งปกติ (normal distribution) เท่านั้น ในโค้งการแจกแจงแบบปกติ ข้อมูลร้อยละ 68 จะตกอยู่ในช่วง ± 1 SD ข้อมูลร้อยละ 95 จะตกอยู่ในช่วง ± 2 SD และข้อมูลร้อยละ 99 จะตกอยู่ในช่วง ± 3 SD แต่ในการแจกแจงที่ไม่เป็นโค้งปกติจะไม่เป็นเช่นนั้น ซึ่งหมายความว่าค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานจะไม่มีความสัมพันธ์กับรูปแบบของการแจกแจงข้อมูล ในขณะที่การใช้ค่าสถิติเป็นค่ามัธยฐาน (median) หรือค่ากลางของข้อมูลซึ่งก็คือเปอร์เซ็นต์ไทล์ที่ 50 ค่าพิสัย (range)

หรือผลต่างระหว่างค่าสูงสุดและค่าต่ำสุดของข้อมูล และค่าพิสัยควอไทล์ (interquartile range) หรือผลต่างระหว่างเปอร์เซ็นไทล์ที่ 75 และ 50 จะมีความเหมาะสมมากกว่า

ถึงแม้ว่าค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานจะสามารถคำนวณได้จากข้อมูลอย่างน้อยที่สุดเพียงสองค่า แต่ค่าเหล่านี้ก็ไม่สามารถนำมาใช้กับข้อมูลจำนวนน้อยขนาดนั้นได้ดีนัก ในเมื่อข้อมูลทางชีวภาพส่วนใหญ่ไม่ได้มีการแจกแจงแบบโค้งปกติ ดังนั้นการทดสอบการแจกแจงของข้อมูลก่อนจึงมีความสำคัญมาก ทั้งนี้เพื่อที่จะได้เลือกใช้สถิติที่เหมาะสมมาอธิบายข้อมูลดังกล่าว

4. มีความสับสนระหว่าง “นัยสำคัญทางสถิติ” (statistical significance) กับ “ความสำคัญทางคลินิก” (clinical importance)

ในทางสถิตินั้น การเปรียบเทียบระหว่างกลุ่มที่มีจำนวนตัวอย่างมากแม้ว่าจะมีความแตกต่างของค่าที่วัดเพียงน้อยนิดซึ่งไม่มีความหมายใด ๆ แต่ก็อาจมีนัยสำคัญทางสถิติได้ ในทางกลับกันการเปรียบเทียบระหว่างกลุ่มที่มีจำนวนตัวอย่างน้อยแม้ว่าจะมีความแตกต่างมากจนมีความสำคัญทางคลินิก แต่ก็อาจไม่มีนัยสำคัญทางสถิติ ยกตัวอย่างเช่น การศึกษาปัจจัยทำนายอย่างหนึ่งในผู้ป่วยโรคไตระยะสุดท้าย ถ้าหากว่ามีผู้ป่วยเพียง 1 คน รอดชีวิต ผู้วิจัยก็ถือว่ามีความสำคัญทางคลินิก ถึงแม้ว่าอัตราการรอดชีวิต (survival rate) จะไม่มีนัยสำคัญทางสถิติก็ตาม

อย่างไรก็ดี ความสำคัญทางคลินิกเป็นค่าที่สามารถนำไปใช้ในการตัดสินใจว่าจะเลือกหรือไม่เลือกวิธีการรักษานั้น ๆ แต่ถึงแม้ว่าค่านี้จะมีความสำคัญอย่างไร ปัจจุบันก็ยังไม่สามารถตั้งกฎเกณฑ์

ที่แน่นอนที่จะกำหนดค่า cut - off ที่ถือว่ามีความสำคัญในทางคลินิกได้ ตัวอย่างค่าทางสถิติที่นิยมนำมาใช้พิจารณาความสำคัญทางคลินิก แต่ก็ยังไม่มีการกำหนดตัวเลขอ้างอิงในการใช้พิจารณาที่แน่นอน ตัวเลขที่ได้จะเป็นไปในเชิงเปรียบเทียบกับค่าอื่นเท่านั้น เช่น relative risk (RR) และ number needed to treat (NNT) เป็นต้น

5. การเลือกใช้การวิเคราะห์แบบ Per protocol กับ Intention to treat ไม่ถูกต้อง

ปัญหาในการวิจัยทางคลินิกที่พบบ่อยคือ การขาดการติดตามผลจนครบตามกำหนด (loss to follow up) ทำให้ข้อมูลจากการวิจัยได้มาไม่สมบูรณ์ ซึ่งมีผลกระทบคือ ขนาดตัวอย่างจะลดลง ผลที่ได้มีอำนาจการทดสอบน้อยลง หรืออาจทำให้ไม่พบความแตกต่างของข้อมูลระหว่างกลุ่มทดลองและกลุ่มเปรียบเทียบได้ ถ้าความเป็นจริงแล้วข้อมูลเป็นผลเสียแต่ไม่ถูกนำมาวิเคราะห์จะทำให้เกิดการสรุปผลผิดพลาด โดยให้ค่าการประมาณผลการทดสอบดีกว่าความเป็นจริง (over - estimated) ในทางกลับกันถ้าข้อมูลเป็นผลดี แต่ไม่นำมาวิเคราะห์จะทำให้เกิดการสรุปผลผิดพลาด โดยให้ค่าการประมาณผลการทดสอบด้อยกว่าความเป็นจริง (under - estimated)

ปัญหาข้อมูลผลการวิจัยไม่สมบูรณ์ อาจเกิดขึ้นได้จากสาเหตุ 3 อย่าง ดังนี้

(1) ผู้ป่วยสูญหายจากการติดตาม (loss to follow-up) เป็นปัญหาสำคัญมากในทางคลินิก เพราะทำให้ไม่ทราบว่าผลการวิจัยที่สนใจเกิดขึ้นหรือไม่ โดยเฉพาะการสูญหายเกิดขึ้นในสัดส่วนที่ไม่เท่ากัน หรือแม้แต่การสูญหายเท่ากันทั้งสองกลุ่ม ถ้าเหตุผลการสูญหายต่างกัน ก็จะทำให้มีปัญหาในการแปลผลและสรุปผลการวิจัยได้มากขึ้น

(2) ผู้ป่วยมีเหตุให้ออกจากการวิจัยก่อนกำหนด (drop out) หรือเสียชีวิต

(3) มีการละเมิดต่อแบบแผนการวิจัย (protocol violation) พบได้ 3 กรณี คือ 3.1) การสลับกลุ่มผู้ป่วยหลังจากแบ่งกลุ่มแล้ว (switched group) 3.2) ไม่ปฏิบัติตามคำแนะนำของแพทย์ (non compliance) และ 3.3) การดำเนินการอื่น ๆ ที่ไม่เป็นไปตามข้อตกลง เช่น การคัดเลือกตัวอย่างไม่เป็นไปตามเกณฑ์การคัดเข้าและคัดออก หรือการประเมินผลไม่เป็นไปตามที่กำหนด

ดังนั้น ในการวิเคราะห์ข้อมูลการวิจัยทางคลินิก เมื่อไม่สามารถติดตามผลได้ครบหรือมีผลการวิจัยที่ไม่สมบูรณ์แล้วผู้วิจัยจะต้องมีความระมัดระวังเรื่องการวิเคราะห์ข้อมูลและแปลผลด้วย รูปแบบการวิเคราะห์ข้อมูลสำหรับประเด็นนี้มี 2 รูปแบบ คือ การวิเคราะห์แบบ Per protocol (PP) เป็นการวิเคราะห์ข้อมูลที่ใช้เฉพาะผลการวิจัยจากกลุ่มตัวอย่างที่เป็นไปตามเงื่อนไขการวิจัยเท่านั้น โดยตัดกลุ่มตัวอย่างที่สูญหายจากการติดตามหรือละเมิดต่อแบบแผนการวิจัยออก หากมีกลุ่มตัวอย่างเช่นนี้จำนวนมาก การสรุปผลการวิจัยจะไม่ใช้ผลจากการวิเคราะห์แบบ PP

สำหรับการวิเคราะห์แบบ Intention to treat (ITT) นั้นเป็นรูปแบบที่ใช้กันแพร่หลายมากกว่า PP โดยมีหลักการสำคัญคือ หลังจากกลุ่มตัวอย่างได้รับการจำแนกกลุ่มแล้วทุกรายจะต้องอยู่ในกลุ่มตามที่กำหนดไว้ ไม่ว่ากลุ่มตัวอย่างจะสูญหาย ออกจากการติดตามหรือละเมิดต่อแบบแผนการวิจัย และกลุ่มตัวอย่างที่ไม่สามารถวัดผลได้จะมีการคำนวณค่าแทนข้อมูลที่สูญหาย ซึ่งมีหลายวิธีขึ้นอยู่กับสถานการณ์

ขอให้พิจารณาตัวอย่างการวิเคราะห์แบบ PP

และ ITT ในการวิจัยต่อไปนี้ ซึ่งมีวัตถุประสงค์เพื่อเปรียบเทียบภาวะแทรกซ้อนของวิธีการระงับปวดแบบ epidural เทียบกับการใช้ patient-controlled analgesia (PCA) ในกลุ่มตัวอย่าง 40 คน ผู้วิจัยทำการสุ่มแบ่งเป็น 2 กลุ่ม โดยมีกลุ่มทดลอง (epidural) 20 คน และกลุ่มควบคุม (PCA) 20 คน จากผลการวิเคราะห์แบบ PP พบว่าทั้ง 2 วิธี ให้ผลแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p\text{-value} = 0.047$) (ตารางที่ 2) โดยที่กลุ่ม epidural เกิดภาวะแทรกซ้อนน้อยกว่าเมื่อเทียบกับกลุ่ม PCA แต่ในความเป็นจริงแล้วภาวะแทรกซ้อนที่น้อยกว่าอาจเกิดจากการสลับกลุ่มของกลุ่มตัวอย่างก็เป็นได้ ทั้งนี้เนื่องจากการศึกษานี้มีผู้ป่วย 4 คนในกลุ่ม epidural ไม่ได้รับการระงับปวดด้วยวิธี epidural แต่ได้รับการระงับปวดด้วย PCA แทน ซึ่งคิดเป็นร้อยละ 20 (ในจำนวนนี้เกิดภาวะแทรกซ้อน 3 คน) ฉะนั้นผลการทดสอบที่สนใจอาจจะน้อยกว่าที่ควรจะเป็น (under - estimated) และเกิดความเข้าใจผิดในการสรุปผล รวมถึงอาจนำไปสู่ปัญหาของ selection bias ได้ เพราะไม่ได้พิจารณาผู้ป่วยที่ไม่ได้เป็นไปตามแบบแผนการวิจัย โดยเฉพาะในกรณีที่เป็นการศึกษา ฤทธิ์ข้างเคียงจากยา¹⁰ ส่วนผลการวิเคราะห์แบบ ITT ที่ได้กำหนดให้ผู้ป่วย “ทุกคน” ที่ถูกสุ่มเข้ากลุ่มมารวมเข้าสู่วิเคราะห์นั้น พบว่าผลการทดสอบไม่มีนัยสำคัญทางสถิติ ($p\text{-value} = 0.330$) สำหรับวิธี ITT นี้ หากสัดส่วนของกลุ่มตัวอย่างที่เกิดการสลับกลุ่มนั้นมากเกินไป ร้อยละ 20 จะยิ่งทำให้การแปลผลยากมากขึ้นด้วย¹¹ จะเห็นว่าการวิเคราะห์โดยวิธี PP และ ITT อาจให้ผลแตกต่างกันอย่างชัดเจน ดังนั้นผู้วิจัยควรเพิ่มความระมัดระวังในการเลือกใช้ระหว่างสองวิธีนี้

Table 2 ตัวอย่างผลการวิเคราะห์แบบ Per protocol (PP) เทียบกับแบบ Intention to treat (ITT) ระหว่างวิธี Epidural และ PCA¹¹

<i>Per protocol analysis</i>			
	Epidural group (n = 16)	PCA group (n = 24)	P-v alue
Complications; n (%)	3 (19%)	13 (54%)	0.047*
<i>Intention - to - treat analysis</i>			
	Epidural group (n = 20)	PCA group (n = 20)	P - value
Complications; n (%)	6 (30%)	10 (50%)	0.330

*p-value < 0.05

6. มีความสัมพันธ์ระหว่างการวิเคราะห์ความสัมพันธ์ (Association) และการวิเคราะห์สาเหตุ (Causation)

ในบางกรณีที่ผู้วิจัยสนใจศึกษารูปแบบความสัมพันธ์ระหว่างสองตัวแปรที่เป็นข้อมูลเชิงตัวเลข (numerical data) ตัวอย่างเช่นความสัมพันธ์ระหว่างอุณหภูมิห้องผ่าตัดกับการเกิดภาวะอุณหภูมิร่างกายต่ำหลังผ่าตัดนั้น การที่จะตัดสินว่าอุณหภูมิห้องผ่าตัดเป็นสาเหตุของการเกิดภาวะอุณหภูมิร่างกายต่ำหลังผ่าตัด สิ่งแรกที่ต้องค้นหาคือ ตัวแปรทั้งสองมีความสัมพันธ์กัน (associate) หรือไม่ และความสัมพันธ์ที่เกิดขึ้นนั้นมีน้ำหนัก (strong) เพียงพอหรือไม่ที่จะบอกว่าอุณหภูมิห้องผ่าตัด เป็นสาเหตุ (cause) ของการเกิดภาวะ อุณหภูมิร่างกายต่ำหลังผ่าตัด หรือเกิดขึ้นจากความบังเอิญเพียงเท่านั้น สถิติเบื้องต้นในการประเมินความสัมพันธ์ว่าไม่ได้เกิดจากความบังเอิญ คือ การพิจารณาจากค่า p แต่ค่าสถิติเพียงอย่างเดียวก็ไม่สามารถบอกถึงความสัมพันธ์เชิงสาเหตุได้เสมอไป การบ่งบอกความเป็นสาเหตุต้องมียอดประกอบหลายประการ ดังเกณฑ์ในการตัดสินว่าความสัมพันธ์ระหว่างตัวแปรใดๆ เป็นสาเหตุและผลของกันและกันหรือไม่ของ Bradford - Hill¹² ซึ่งหากความสัมพันธ์ใด ๆ ที่มีลักษณะเข้าเกณฑ์นี้มาก ก็หมายความว่าน่า

จะเป็นเหตุเป็นผลกันมากด้วย เกณฑ์ในการพิจารณาดังกล่าวมี 9 ประการ ดังต่อไปนี้

- 1) Strength การศึกษาปัจจัยเสี่ยงเมื่อมีความสัมพันธ์ต่อการเกิดโรคไม่ได้แสดงว่าปัจจัยดังกล่าวเป็นสาเหตุของการเกิดโรคเสมอไป แต่ระดับความสัมพันธ์ที่ค่อนข้างสูง ความน่าจะเป็นสาเหตุและผลก็น่าจะมากขึ้นด้วย
- 2) Consistency ผลการศึกษาอื่น ๆ ที่ทำการทดลองแบบเดียวกันได้ผลคล้ายคลึงหรือสอดคล้องกัน
- 3) Temporality หรือ Temporal Relationship ปัจจัยสาเหตุต้องเกิดก่อนตัวแปรที่เป็นผลเสมอ ตัวอย่างเช่น การสูบบุหรี่ทำให้เกิดโรคมะเร็งปอด ไม่ใช่เป็นมะเร็งปอดเกิดก่อนการสูบบุหรี่
- 4) Specificity ความสัมพันธ์ระหว่างปัจจัยนั้นมีความเฉพาะเจาะจง เช่น การสูบบุหรี่เพียงอย่างเดียวทำให้เกิดโรคมะเร็งปอด
- 5) Dose-response relationship เมื่อระดับปัจจัยเสี่ยงมากขึ้นทำให้ปัจจัยผลลัพธ์มากขึ้นด้วย
- 6) Plausibility มีคำอธิบายหรือทฤษฎีที่สมเหตุสมผลในการอธิบายกลไกการเกิดโรค เช่น เมื่อสูบบุหรี่แล้วสารก่อมะเร็งในบุหรี่ไปทำให้เนื้อเยื่อปอดแบ่งตัวผิดปกติและทำให้เป็นมะเร็ง
- 7) Coherence การแปลผลความเป็นเหตุและ

ผลจะต้องไม่ขัดแย้งกับธรรมชาติของการเกิดโรคหรืออาจกล่าวได้ว่าการศึกษาทางระบาดวิทยาสอดคล้องกับการศึกษาทางห้องปฏิบัติการ

8) Experiment หากมีการศึกษาเชิงทดลอง ผู้วิจัยสามารถควบคุมปัจจัยสิ่งแวดล้อมได้เกือบทั้งหมด ดังนั้นจะช่วยเพิ่มความน่าจะเป็นของการเป็นสาเหตุได้มากขึ้น

การศึกษาวิจัยทางด้านการแพทย์และสาธารณสุข มีเป้าหมายเพื่อทำความเข้าใจ ค้นหาปัญหาสุขภาพ และสาเหตุของปัญหาสุขภาพ รวมทั้งแนวทางในการแก้ไขปัญหาเหล่านั้น เพื่อให้บรรลุวัตถุประสงค์ดังกล่าว องค์ความรู้ด้านวิทยาการระบาดได้มีการเสนอแนวคิดในการศึกษาโดยมีรูปแบบการศึกษา ดังต่อไปนี้¹³

1) การศึกษาเชิงสังเกต (Observational study)

ก. เพื่อการวิเคราะห์สถานการณ์และตั้งสมมติฐานเกี่ยวกับสาเหตุ คือ การศึกษาเชิงพรรณนา (Descriptive study) ได้แก่ Ecological study (Correlation study) และ Cross - sectional study ซึ่งไม่มีกลุ่มเปรียบเทียบ

ข. เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องกับการเกิดโรคโดยมุ่งพิสูจน์ว่าปัจจัยนั้น ๆ อาจเป็นสาเหตุของปัญหาสุขภาพได้หรือไม่ คือ การศึกษาเชิงวิเคราะห์ (Analytical study) ได้แก่ Cross - sectional study, Case-control study และ Cohort study

2) การศึกษาเชิงทดลอง (Experimental study) เพื่อศึกษาประสิทธิภาพ ประสิทธิผลของโครงการหรือสิ่งทดลองต่าง ๆ ได้แก่ Clinical trial, Field trial และ Community trial

แนวคิดในการศึกษาเกี่ยวกับโรคตามรูปแบบการศึกษาวิทยาการระบาดที่จะตอบคำถามเชิงสาเหตุได้คือนั้น ได้แก่ การศึกษาเชิงวิเคราะห์และเชิงทดลอง และที่แย่ที่สุดคือการศึกษาแบบ Cross - sectional

แต่การศึกษาเชิงทดลองในแง่ความเป็นสาเหตุหรือปัจจัยเสี่ยงต่อการเกิดโรค จะทำการศึกษาไม่ได้เพราะเป็นการผิดจริยธรรม แต่อาจทำได้ด้วยการทดลองในการพยายามลดปัจจัยเสี่ยงลงและเทียบกับอีกกลุ่มหนึ่งว่ามีความแตกต่างในการเกิดโรคหรือไม่

สรุป

แนวทางแก้ไขความผิดพลาดทางการใช้สถิติที่ดีที่สุดก็คือการป้องกันไว้ตั้งแต่ขั้นต้นและกำหนดในโครงร่างการวิจัยอย่างชัดเจน ผู้วิจัยควรหมั่นเพิ่มพูนความรู้เรื่องสถิติหรือให้ความสำคัญกับการมีนักสถิติมาช่วยเหลือในการวิเคราะห์ข้อมูลระหว่างการวิจัย วารสารวิชาการทางการแพทย์ต่าง ๆ ควรนำแนวทางการรายงานผลทางสถิติต่าง ๆ มาใช้ประโยชน์ อาทิเช่น คำแนะนำในการใช้สถิติสำหรับงานวิจัยที่ตีพิมพ์ในวารสาร Anesthesia & Analgesia¹⁴ หรือ ‘BMJ Checklists for statisticians’¹⁵ ซึ่งเป็นแบบตรวจสอบคุณภาพงานวิจัยในเรื่องสถิติด้วยตนเอง ซึ่งจัดทำโดยกองบรรณาธิการวารสาร British Medical Journal หรือหากวารสารที่จะตีพิมพ์อยู่ในกลุ่มของคณะกรรมการบรรณาธิการวารสารทางการแพทย์นานาชาติ (International Committee of Medical Journal Editors; ICMJE) ก็อาจศึกษาคำแนะนำเพื่อการส่งต้นฉบับลงตีพิมพ์ในวารสารทางการแพทย์ (ICMJE Recommendations)¹⁶ ได้ นอกจากนี้ยังมีอยู่ในเอกสารคำแนะนำสากลในการพัฒนาเนื้อหาของรายงานการวิจัย (International guideline for content development) อื่นๆ เช่น AMA, WAME, AMWA, CONSORT, PhRMA, OIG หรือ COPE ซึ่งหากผู้อ่านวารสารมีการพัฒนาความรู้ด้านสถิติเป็นอย่างดีแล้ว ก็จะสามารถนำความหมายและค่าสถิติที่มีอยู่ในผลการวิจัยไปใช้ในการพัฒนาแนวทางการรักษาผู้ป่วยได้ดียิ่งขึ้น

เอกสารอ้างอิง

1. Robinson DH, Toledo AH. Historical development of modern anesthesia. *J Invest Surg.* 2012 Jun; 25(3):141-9.
2. Goodman NW, Hughes AO. Statistical awareness of research workers in British anaesthesia. *Br J Anaesth.* 1992;68(3):321-4.
3. Mainland D. Chance and the blood count. 1934. *CMAJ.* 1993;148:225-7.
4. Claridge JA, Fabian TC. History and development of evidence-based medicine. *World J Surg.* 2005;29(5):547-53.
5. Yim KH, Nahm FS, Han KA, Park SY. Analysis of statistical methods and errors in the articles published in the Korean Journal of Pain. *Korean J Pain.* 2010;23(1):35-41.
6. Murray C. How to Accuse the Other Guy of Lying with Statistics. *Statistical Science.* 2005; 20(3):239-41.
7. Moses LE, Emerson JD, Hosseini H. Statistics in practice: Analyzing data from ordered categories. *N Engl J Med.* 1984;311:442-8.
8. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol.* 2009;62(5):464-75.
9. Feinstein AR. X and iprP: an improved summary for scientific communication. *J Chronic Dis.* 1987;40:283-8.
10. Sedgwick P. Analysis by per protocol. *BMJ.* 2011:1-2.
11. Aron G, Boutron I, Giraudeau B, Ravaud P. Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. *Arthritis Rheum.* 2005;52(6):1858-65.
12. Hill AB. The Environment and disease: association or causation? *Proc R Soc Med.* 1965;58:295-300.
13. Rothman KJ, Lash TL, Greenland S. Types of Epidemiologic study. In: Rothman KJ, Lash TL, Greenland S, editors. *Modern Epidemiology.* Baltimore: Lippincott Williams & Wilkins; 2012. p. 87-99.
14. Anesthesia & Analgesia Guide for Authors. [Internet]. 2014 [cited 2014 Feb 5]; [about 50 p.]. Available from: <http://www.aeditor.org/GuideforAuthors.pdf>
15. BMJ Checklists for statisticians. *BMJ.* 1996; 312:43-4.
16. Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals. [Internet]. 2013 [cited 2014 Feb 5]; [about 17 p.]. Available from: <http://www.icmje.org/icmje-recommendations.pdf>