สถิติกับงานวิทยาเอ็นโดดอนต์ ตอนที่ 2: ความเชื่อถือได้ (Statistics in Endodontics Part 2: Reliability)

สิทธิโชค โอศิริ

สาขาวิชาวิทยาเอ็นโดดอนต์ ภาควิชาทันตกรรมหัตถการและวิทยาเอ็นโดดอนต์ คณะทันตแพทยศาสตร์ มหาวิทยาลัยมหิดล

บทคัดย่อ

บทความนี้เสนอภาพรวมเชิงลึกเกี่ยวกับความเชื่อถือได้ ความสำคัญ และวิธีการต่าง ๆ ที่ใช้ในการประเมินความเชื่อถือได้ ความเชื่อถือได้มีความสำคัญในการวิจัยและการประเมินเนื่องจากเป็นตัวแทนของความสอดคล้องและความสามารถในการทำซ้ำ การมีความเชื่อถือได้ที่สูงทำให้เรามั่นใจในผลการวิจัย ซึ่งทำให้ความรู้ในหลายด้านพัฒนาขึ้น ในทางกลับกันข้อมูลที่ไม่น่าเชื่อถือสามารถนำไปสู่การวิเคราะห์ที่ผิดพลาด ภายในบทความยังได้กล่าวถึงนิยามและประเภทของความเชื่อถือได้ รวมถึงวิธีการประเมินความเชื่อถือได้ต่าง ๆ เช่น ความเชื่อถือได้ในการทดสอบซ้ำเน้นไปที่ผลลัพธ์ที่สอดคล้องจากการวัดซ้ำๆ ในช่วงเวลา ความเชื่อถือได้ระหว่างผู้ประเมินเน้นไปที่ความสอดคล้องของการประเมินระหว่างผู้ประเมินที่ต่างกันของข้อมูลเดียวกัน ในขณะที่ความเชื่อถือได้ภายในผู้ประเมินเน้นไปที่ความสอดคล้องของผู้ประเมินคนเดียวในช่วงเวลาที่ต่างกัน นอกจากนี้ยังมีรายละเอียดเกี่ยวกับวิธีต่าง ๆ ในการประเมินความเชื่อถือได้ เช่น ร้อยละของความเห็นพ้องระหว่างผู้ประเมิน ค่าแคปปา สัมประสิทธิ์สหสัมพันธ์ภายในชั้น และสถิติอื่น ๆ ที่เกี่ยวข้อง

คำสำคัญ: ความเชื่อถือได้, ความเชื่อถือได้ระหว่างผู้ประเมิน, ความเชื่อถือได้ภายในผู้ประเมิน, สถิติแคปปา, สัมประสิทธิ์สหสัมพันธ์ภายในชั้น

Abstract

This article provides an in-depth exploration of the concept of reliability, its significance, and various methods used for its assessment. Reliability is crucial in research and evaluation as it represents consistency and reproducibility in various contexts. High reliability instills confidence in the research or evaluation outcomes, advancing knowledge in several domains. Conversely, unreliable data can lead to erroneous analyses. Key aspects of reliability discussed include its definition and types. Test-retest reliability focuses on the consistent outcomes of repeated measurements over time. Inter-rater reliability examines the consistency of evaluations between different assessors of the same data, while intra-rater reliability emphasizes the consistency of a single assessor over different times. Various methods to assess reliability, such as percent agreement, Cohen's kappa, and Intra-Class Correlation (ICC), among others, are also detailed.

Keywords: reliability, inter-rater reliability, intra-rater reliability, kappa, intraclass correlation coefficient.

Correspondence: ผศ.ทพ.สิทธิโชค โอศิริ
ภาควิชาทันตกรรมหัตถการและวิทยาเอ็นโดดอนต์ คณะทันตแพทยศาสตร์ มหาวิทยาลัยมหิดล
เลขที่ 6 ถนนโยธี แขวงทุ่งพญาไท เขตราชเทวี กรุงเทพฯ 10410
โทรศัพท์: 0-2200-7825-6
Email address: sittichoke.osi@mahidol.ac.th
Received: 20 September 2023 Revised: 4 December 2023 Accepted: 7 December 2023

บทนำ

ในบริบทการศึกษาวิจัยทางเอ็นโดดอนติกส์นั้น จะต้องมีการพิจารณาในเรื่อง ความเชื่อถือได้ (reliability) เกี่ยวกับความคงที่ (consistency) และการทำซ้ำได้ (reproducibility) ของการวัด การสังเกต หรือการวินิจฉัย จากผู้ประเมินต่างคนกัน (different rater) หรือคนเดียวกัน (same rater) ซึ่งการประเมินที่ถูกต้องและน่าเชื่อถือจะช่วยให้ทันตแพทย์สามารถตัดสินใจอย่างมีเหตุผล และนำผลการศึกษาวิจัยไปใช้และต่อยอดสู่การดูแลผู้ป่วยดีขึ้น นำไปสู่ผลลัพธ์การรักษาที่ดีที่สุด อีกทั้งในบริบทของงานวิจัยทางวิทยาศาสตร์ความเชื่อถือได้ที่สูง (high level of reliability) และ ความเชื่อถือได้จะเพิ่มความเชื่อถือ (credibility) ความสมเหตุสมผล (validity) ของผลการศึกษา ซึ่งนำไปสู่งานวิจัยที่มีหลักฐานเชิงประจักษ์ระดับสูง (high level of evidence)

การวัดในทางวิทยาเอ็นโดดอนต์ มีความสำคัญอย่างยิ่งในการศึกษาวิจัยทางเอ็นโดดอนติกส์ อย่างไรก็ตาม มีหลายปัจจัยที่สามารถทำให้การวัดเหล่านี้มีข้อผิดพลาดหรือความแปรปรวน (variability) รวมถึงความแตกต่างจากผู้ประเมินต่างคนกัน ซึ่งอาจเนื่องจากประสบการณ์ การฝึกอบรม และการตีความข้อมูลที่มีความแตกต่างกัน แม้แต่ผู้ประเมินคนเดียวกันอาจมีความแปรปรวนในการวัดซ้ำ ๆ ทำให้เกิดความไม่สอดคล้องในผลลัพธ์ ด้วยเหตุนี้การทดสอบความเชื่อถือได้ (reliability testing) จึงมีความสำคัญเพื่อสร้างความมั่นใจในความสม่ำเสมอและความสามารถในการทำซ้ำได้ของผลการวัด นำไปสู่การศึกษาวิจัยที่มีหลักฐานเชิงประจักษ์ระดับสูง สุดท้ายผลลัพธ์ที่น่าเชื่อถือจะช่วยให้ทันตแพทย์สามารถตัดสินใจอย่างมีเหตุผล และนำผลการวิจัยไปใช้เพื่อปรับปรุงคุณภาพการดูแลผู้ป่วยได้ดียิ่งขึ้น ซึ่งท้ายที่สุดนำไปสู่การรักษาที่มีคุณภาพและผลลัพธ์ที่ดีขึ้น

บทความนี้รวบรวมรายละเอียดเนื้อหาเกี่ยวกับความเชื่อถือได้ รวมถึงวิธีการที่ใช้ในการประเมิน ความเชื่อถือได้ ปัจจัยที่มีผลต่อการประเมินเหล่านี้ บทความนี้มีเป้าหมายที่จะสร้างความเข้าใจอย่างละเอียดเรื่องความสำคัญของความเชื่อถือได้ ซึ่งมีส่วนช่วยสร้างการพัฒนาการศึกษาวิจัยทางเอ็นโดดอนติกส์ต่อไป

1. ความสำคัญและวิธีการในการประเมินความเชื่อถือ

ความเชื่อถือได้ (reliability) เป็นสิ่งสำคัญในการประเมินและการวิจัย เนื่องจากเป็นสิ่งที่แทนความคงที่ (consistency) และการทำซ้ำได้ (reproducibility) ในบริบทต่าง ๆ ของข้อมูลหรือการวัดที่มีความเชื่อถือได้สูงทำให้เกิดความมั่นใจว่าผลการวิจัยหรือประเมินนั้นเป็นข้อมูลที่น่าเชื่อถือ นอกจากนี้การมีความเชื่อถือได้ที่ดียังเป็นการสร้างพัฒนาองค์ความรู้ในหลาย ๆ ด้าน นอกจากนี้ยังช่วยให้การสรุปของผลการวิจัยมีค่าและน่าเชื่อถือ แต่อย่างไรก็ตามข้อมูลที่ไม่น่าเชื่อถือสามารถทำให้เกิดความผิดพลาดในการวิเคราะห์ได้ กล่าวโดยสรุปคือความเชื่อถือได้ที่ดีช่วยให้การวัดและการวิจัยมีความถูกต้องและน่าเชื่อถือมากยิ่งขึ้น

2. นิยามของความสมเหตุสมผล (Validity) และความเชื่อถือได้ (Reliability) [1, 2]

ความสมเหตุสมผล (Validity) หมายถึง ความถูกต้องของวัดหรือประเมินสิ่งที่เราต้องการวัด ตรงกับค่าจริง เป็นการตรวจสอบว่าเครื่องมือที่ใช้ในการวัดจริง ๆ สามารถวัดสิ่งที่เราต้องการได้โดยถูกต้องและมีความเกี่ยวข้องกับสิ่งที่ต้องการวัดนั้น ๆ หรือไม่ เช่น การวัดความยาวรากฟันด้วยเครื่องมืออิเล็กทรอนิกส์ (Electronic Apex Locator) ความสมเหตุสมผลของการวัดนี้คือความสามารถในการตรวจสอบความยาวของรากฟันอย่างแม่นยำ ใกล้เคียงกับค่าจริง

ความเชื่อถือได้ (Reliability) หมายถึง ความเสถียรและความสม่ำเสมอของการวัดหรือการประเมิน เมื่อใช้เครื่องมือวัดเดียวกันในสถานการณ์หรือเงื่อนไขที่คล้ายคลึงกัน ผลการวัดจะต้องใกล้เคียงและสามารถทำซ้ำได้ ความเชื่อถือได้จึงเป็นเครื่องมือในการทดสอบว่าความแตกต่างที่ปรากฏในผลการวัดเป็นจริง หรือเป็นเพียงผลของความผิดพลาดในการวัด เราสามารถวัดความเชื่อถือได้ได้หลายรูปแบบ เช่น ความเชื่อถือได้ของการทดสอบซ้ำ (test-retest reliability) ความเชื่อถือได้ระหว่างผู้ประเมิน (interrater reliability) หรือความเชื่อถือได้ภายในผู้ประเมิน (intra-rater reliability)

โดยสรุป “ความสมเหตุสมผล” จัดเป็นการตรวจสอบความถูกต้องและความเกี่ยวข้องของการวัดต่อเป้าหมายวัด ในขณะที่ “ความเชื่อถือได้” คือความเสถียรและสามารถทำ

ซ้ำได้ของผลการวัด ทั้งสองแนวคิดเป็นสิ่งที่สำคัญในการวิจัยและการประเมิน เพื่อรับรองความถูกต้อง ความเชื่อถือ และคุณภาพของผลลัพธ์

3. ชนิดของความเชื่อถือได้ [1, 2]

3.1 ความเชื่อถือได้ของการทดสอบซ้ำ (Test-Retest Reliability):

หมายถึง ความสามารถของเครื่องมือการวัดหรือแบบประเมินที่จะสามารถให้ผลลัพธ์ที่คล้ายคลึงกันเมื่อใช้ซ้ำในช่วงเวลาที่ต่างกัน โดยไม่มีการเปลี่ยนแปลงของสิ่งที่วัด วิธีการคือจะทำการวัดหรือประเมินเปรียบเทียบ 2 ครั้งหรือมากกว่าในช่วงเวลาที่กำหนด เช่น การทดสอบความยาวของรากฟันด้วยเครื่องมืออิเล็กทรอนิกส์หลายครั้งเพื่อตรวจสอบความสม่ำเสมอของผลลัพธ์ โดยความเชื่อถือได้ของการทดสอบซ้ำจะสูงเมื่อผลการวัดในแต่ละครั้งให้ค่าที่ใกล้เคียงกัน

3.2 ความเชื่อถือได้ระหว่างผู้ประเมิน (Inter-Rater Reliability)

หมายถึง ความสอดคล้องของการประเมินระหว่างผู้ประเมินที่ต่างกันตั้งแต่สองคนขึ้นไป ในการประเมินข้อมูลหรือสิ่งเดียวกัน วิธีการคือให้ผู้ประเมินที่แตกต่างกันมาวัดสิ่งเดียวกัน แล้วเปรียบเทียบผลการวัดของแต่ละผู้ประเมินว่ามีความเหมือน หรือแตกต่างกันมากน้อยเพียงใด เช่น ในการประเมินภาพถ่ายรังสีของพยาธิสภาพรอบรากฟันโดยให้ทันตแพทย์หลายคนที่มีประสบการณ์และความเชี่ยวชาญต่างกันประเมินภาพรังสีเหล่านั้น ความเชื่อถือได้ระหว่างผู้ประเมินจะสูงหากมีความสอดคล้องในการประเมินผลของภาพรังสีนั้น ๆ

3.3 ความเชื่อถือได้ภายในผู้ประเมิน (Intra-Rater Reliability)

หมายถึง ความสามารถของผู้ประเมินคนเดียวกันในการให้ผลการวัดที่สม่ำเสมอเมื่อวัดซ้ำ ๆ ในช่วงเวลาที่ต่างกัน วิธีการคือผู้ประเมินคนเดียวกันประเมินข้อมูลซ้ำ ๆ ในช่วงเวลาที่แตกต่างกัน โดยประเมินซ้ำในช่วงระยะเวลาที่ห่างมากเพียงพอที่ไม่สามารถจดจำผลของครั้งที่แล้วได้ นอกจากนี้อาจทำได้โดยสลับลำดับของสิ่งที่ประเมิน

4. วิธีการประเมินความเชื่อถือได้

การวัดความเชื่อถือได้ของเครื่องมือการวัดเป็นสิ่งสำคัญในการวิจัยหรือการประเมินผลเพื่อที่จะมั่นใจว่าผลที่ได้มาจากการวัดนั้นเชื่อถือได้และสามารถใช้ในการตัดสินใจได้อย่างมีประสิทธิภาพ ซึ่งวิธีการประเมินความเชื่อถือได้ทางสถิติจึงเป็นเครื่องมือที่ช่วยให้ผู้วิจัยมั่นใจในคุณภาพของเครื่องมือวัดหรือวิธีการวัดนั้น ๆ เมื่อต้องการเลือกการวิเคราะห์ความเชื่อถือได้ควรพิจารณาคำถามหลักดังนี้ [3]:

1. วัตถุประสงค์ของการวิเคราะห์
2. ประเภทของตัวแปรที่ถูกวิเคราะห์
3. จำนวนผู้ประเมิน

สถิติที่นำมาใช้ในการประเมินความเชื่อถือได้ในการวิจัยสามารถแบ่งได้เป็นหลายแนวทาง (ตารางที่ 1) [3, 4]

ตารางที่ 1 สถิติที่นำมาใช้ในการประเมินความเชื่อถือได้ในการศึกษาวิจัย โดยขึ้นกับชนิดของตัวแปร (ตามระดับการวัด) และจำนวนผู้ประเมิน

	ระดับการวัดของตัวแปร (Level of measurement)					
	นามบัญญัติ		จัดลำดับ		อันตรภาค และสัดส่วน	
	ผู้ประเมิน 2 คน	ผู้ประเมิน มากกว่า 2 คน	ผู้ประเมิน 2 คน	ผู้ประเมิน มากกว่า 2 คน	ผู้ประเมิน 2 คน	ผู้ประเมิน มากกว่า 2 คน
สถิติที่ควรเลือกใช้	1. Cohen's kappa 2. ICC 3. Weighted kappa	1. Fleiss' kappa 2. ICC	1. Weighted kappa 2. ICC	1. Kendall coefficient of concordance 2. ICC	1. Bland-Altman plots 2. ICC	1. ICC

4.1 ดัชนีหลักที่ใช้ในการประเมิน (Key indices)

1. ร้อยละของความสอดคล้องระหว่างผู้ประเมิน (Percent agreement)

เป็นดัชนีพื้นฐานที่ให้ข้อมูลเกี่ยวกับระดับความเชื่อถือได้หรือความสอดคล้อง คำนวณได้ง่ายที่สุด โดยไม่มีการจำกัดในจำนวนผู้ประเมิน แต่อาจมีโอกาสปฏิเสธความสอดคล้องที่มีอยู่ได้ (ignoring agreement by chance) [3] ค่าของ percent agreement ที่ถือว่ายอมรับได้ในการวิจัยคืออย่างน้อย 70% [5] มีวิธีการคำนวณคือ จำนวนผลการประเมินที่สอดคล้องกันหารด้วยจำนวนการประเมินทั้งหมด ตัวอย่างการคำนวณเช่น เมื่อมีผู้ประเมินภาพรังสี 2 คนที่ประเมินการพบรอยโรครอบปลายรากฟันในภาพรังสี และได้ผลดังตารางที่ 2

ตารางที่ 2 ผลการประเมินรอยโรครอบปลายรากฟันในภาพรังสีของผู้ประเมิน 2 คน

	ผู้ประเมินคนที่ 1: มีรอยโรค	ผู้ประเมินคนที่ 1: ไม่มีรอยโรค
ผู้ประเมินคนที่ 2: มีรอยโรค	50	10
ผู้ประเมินคนที่ 2: ไม่มีรอยโรค	5	35

จากตารางข้างต้น:

จำนวนครั้งที่ผู้ประเมินทั้งสองคนประเมินสอดคล้องกันนี่คือจำนวนครั้งที่ทั้งคู่ประเมินว่ามีรอยโรค (50 ครั้ง) และไม่มีรอยโรค (35 ครั้ง) รวม 85 ครั้ง

หาจำนวนครั้งทั้งหมดที่ทำการประเมิน เท่ากับ 50+10+5+35 = 100

คำนวณร้อยละของความสอดคล้อง เท่ากับ (85/100) x 100 = 85

ตัวอย่างงานวิจัย เช่น งานวิจัยซึ่งมีวัตถุประสงค์เพื่อวัดความสอดคล้องในการตีความภาพรังสีของโรคอักเสบรอบปลายรากฟันระหว่างผู้เชี่ยวชาญสองคน [6]

2. สถิติแคปปา (Kappa Statistics) หรือสถิติแคปปาแบบไม่ถ่วงน้ำหนัก (unweight kappa):

สถิติแคปปาแบบไม่ถ่วงน้ำหนัก (unweight kappa) มักจะถูกเรียกว่า สถิติแคปปาของโคเฮน (Cohen's kappa) เป็นวิธีที่นิยมใช้ในการศึกษาที่ใช้ผู้ประเมิน 2 คนเพื่อประเมินข้อมูลตัวแปรระดับนามบัญญัติ (nominal variable) ไม่แนะนำให้คำนวณแคปปาของโคเฮนเพื่อประเมินความตกลงกันตัวแปรระดับจัดลำดับ (ordinal variable) ถึงแม้ว่าทางทฤษฎีจะเป็นไปได้ เนื่องจากมีสถิติที่เหมาะสมมากกว่าซึ่งจะได้กล่าวถึงต่อไป [7]

สำหรับสถิติแคปปา ค่า คือ สัดส่วนของความเห็นพ้องที่สังเกตได้ (observed agreement) ที่ลบด้วยความเห็นพ้องที่คาดหวัง (expected agreement) แล้วหารด้วยค่า 1 ลบด้วยความเห็นพ้องที่คาดหวัง สามารถคำนวณค่าสัมประสิทธิ์แคปปา ได้จากสูตร

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

โดย

p_0 คือ สัดส่วนของความเห็นพ้องที่สังเกตได้

p_e คือ ความเห็นพ้องที่คาดหวัง

ตัวอย่างการคำนวณค่าสถิติแคปปาในทางเอ็นโดดอนติกส์

เมื่อมีผู้ประเมินภาพรังสี 2 คนที่ประเมินการพบรอยโรครอบปลายรากฟันในภาพรังสี และได้ผลดังตารางที่ 2 ดังตารางที่ 2

ความเห็นพ้องที่สังเกตได้ = [(จำนวนการประเมินทั้งหมดของผู้ประเมินคนที่ 1 และ 2 ที่ได้ผลว่ามีรอยโรค)+[(จำนวนการประเมินทั้งหมดของผู้ประเมินคนที่ 1 และ 2 ที่ได้ผลว่าเป็นไม่มีรอยโรค)]/จำนวนการประเมินทั้งหมด = (50+35)/100 = 0.85 หรือ 85%

ความเห็นพ้องที่คาดหวัง

ความเห็นพ้องที่คาดหวังคือความเห็นพ้องที่อาจเกิดขึ้นโดยบังเอิญ สามารถคำนวณได้ดังนี้

ความเห็นพ้องกรณีพบรอยโรค: คำนวณโดยการคูณความน่าจะเป็นของผู้ประเมินทั้งสองที่จะประเมินว่าพบรอยโรค ซึ่งคำนวณได้จาก (จำนวนการประเมินที่พบรอยโรคโดยผู้ประเมินคนที่ 1/จำนวนการประเมินทั้งหมด) * (จำนวนการประเมินที่พบรอยโรคโดยผู้ประเมินคนที่ 2/จำนวนการประเมินทั้งหมด) ในกรณีนี้ ความเห็นพ้องกรณีพบรอยโรค = (55/100) * (60/100) = 0.33 หรือ 33%

ความเห็นพ้องกรณีไม่พบรอยโรค: คำนวณโดยการคูณความน่าจะเป็นของผู้ประเมินทั้งสองที่จะประเมินว่าไม่พบรอยโรค ในกรณีนี้ความเห็นพ้องกรณีไม่พบรอยโรค = (45/100) * (40/100) = 0.18 หรือ 18%

ความเห็นพ้องที่คาดหวังรวม = 0.33 + 0.18 = 0.51 หรือ 51%

การคำนวณค่าแคปปาตามกรณีตัวอย่าง

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.85 - 0.51}{1 - 0.51} = 0.69$$

ค่า κ สามารถแปรผันได้ตั้งแต่ -1 ถึง 1

ค่า Kappa = 1: หมายถึงความเห็นพ้องสมบูรณ์ คือ ผู้ประเมินทุกคนมีความเห็นพ้องกันอย่างสมบูรณ์ โดยไม่มีความแตกต่างในการประเมินเลย

ค่า Kappa = 0: หมายถึงความเห็นพ้องที่เกิดขึ้นเท่ากับที่คาดหวัง คือ ความเห็นพ้องที่สังเกตได้ไม่ได้มีมากกว่าที่คาดหวังจากการบังเอิญ และไม่ได้บ่งบอกถึงความเชื่อถือได้ที่แท้จริงของการประเมินนั้น ๆ เช่น การประเมินรอยโรครอบปลายรากฟันในภาพรังสีของผู้ประเมิน 2 คน ผู้ประเมินคนที่ 1 ระบุว่ามีรอบโรครอบปลายรากรากฟันใน 50% ของผู้ป่วย ผู้ประเมินคนที่ 2 ระบุว่ามีรอยโรครอบปลายรากฟันใน 50% ของผู้ป่วยเช่นกัน แต่เป็นผู้ป่วยที่ถูกประเมินว่าพบรอยโรคเป็นคนละคนกับที่ผู้ประเมินคนที่ 1 ระบุ ในกรณีนี้ถึงแม้จะดูเหมือนว่าผู้ประเมินทั้งสองมีความเห็นพ้องกัน 50% แต่ความเห็นพ้องนี้เกิดขึ้นเท่ากับที่คาดหวังจากการบังเอิญเท่านั้น เนื่องจากผู้ประเมินทั้งสองประเมินผลการพบรอยโรครอบปลายรากฟันจากภาพรังสีของผู้ป่วยที่แตกต่างกันโดยสิ้นเชิง ในกรณีนี้ ค่า Kappa จะเป็น 0

ค่า Kappa < 0 (Negative): หมายถึงความเห็นพ้องที่น้อยกว่าที่คาดหวัง กรณีนี้อาจเกิดขึ้นเมื่อมีความขัดแย้งสูงระหว่างผู้ประเมิน มีการประเมินที่สวนทางกันอย่างเห็นได้ชัด เช่น การประเมินรอยโรครอบปลายรากฟันในภาพรังสีของผู้ประเมิน 2 คน และผลลัพธ์ที่ได้จากทั้งสองคนมีความขัดแย้งสูง โดยผู้ประเมินคนแรกมองว่ามีรอยโรครอบปลายรากฟันส่วนใหญ่ ในขณะที่ผู้ประเมินคนที่สองมองว่าไม่มีรอยโรครอบปลายรากฟันส่วนใหญ่ ในกรณีนี้ ความเห็นพ้องที่สังเกตได้จะน้อยกว่าความเห็นพ้องที่คาดหวังจากการบังเอิญ ซึ่งทำให้ค่า Kappa อาจมีค่าเป็นลบ

การได้ค่า Kappa เป็นค่าศูนย์หรือลบบ่งบอกถึงความจำเป็นที่จะต้องทบทวนเกณฑ์การประเมินหรือการฝึกอบรมผู้ประเมินเพื่อลดความคลาดเคลื่อนและเพิ่มความเห็นพ้องในการประเมิน

สำหรับค่า Kappa ที่คำนวณได้จะมีค่าตั้งแต่ -1 ถึง +1 การตีความค่าสถิติแคปปามีหลายวิธี โดยวิธีที่นิยมคือ Landis and Koch (1977) จะระบุค่าดังตารางที่ 3 [8]

ตัวอย่างงานวิจัย การประเมินคุณภาพของการรักษารากฟันโดยทันตแพทย์เฉพาะทางสาขาวิทยาเอ็นโดดอนต์ 2 ราย ในเรื่องความยาวของวัสดุอุดคลองรากฟัน ความแน่นของวัสดุอุดคลองรากฟัน ความผายของคลองรากฟัน และการมีข้อผิดพลาดที่เกิดขึ้น วิเคราะห์สถิติแคปปาได้ 1.00 0.63 1.00 และ 0.86 ตามลำดับ [9]

ตารางที่ 3 แสดงการแปลผลค่าแคปปา ของ Landis and Koch (1977)

ค่าสถิติ Kappa	ระดับความสอดคล้อง
น้อยกว่า 0.00	ไม่มีความสอดคล้อง (Poor, worse than chance)
0.00–0.20	ความสอดคล้องน้อย (Slight)
0.21–0.40	ความสอดคล้องพอใช้ (Fair)
0.41–0.60	ความสอดคล้องปานกลาง (Moderate)
0.61–0.80	ความสอดคล้องดี (Good)
0.81–1.00	ความสอดคล้องดีมาก (Very good)

3. สถิติแคปปาถ่วงน้ำหนัก (Weighted Kappa, κ_w) [10]

สถิติ Kappa ถูกใช้เป็นการวัดความเห็นพ้องหรือความสอดคล้องระหว่างผู้ประเมิน 2 คน โดยทั่วไปการคำนวณ Kappa จะเน้นที่ความเห็นพ้องที่เป็นตัวแปรระดับนามบัญญัติ แต่ในบางครั้งข้อมูลที่เรามีอาจจะเป็นตัวแปรระดับจัดอันดับ (ordinal variable) ซึ่งการให้คะแนนไม่เท่ากันเนื่องจากมีความหมายตามลำดับ ดังนั้น เราจึงจำเป็นต้องใช้สถิติแคปปาถ่วงน้ำหนัก ซึ่งเป็นการปรับปรุงของค่า Kappa ให้น้ำหนักกับความไม่เห็นพ้องต่างกัน ตามความรุนแรงหรือความสำคัญของความแตกต่างนั้น ๆ โดยความไม่เห็นพ้องระหว่างการประเมินที่มีความแตกต่างกันมากจะได้รับน้ำหนักที่สูงกว่าความไม่เห็นพ้องที่แตกต่างกันน้อย ตัวอย่างเช่น การประเมินระดับความรุนแรงของอาการปวด มีการให้คะแนนตั้งแต่ 1 ถึง 5 ถ้าผู้ประเมินคนหนึ่งให้คะแนนเป็น 2 และอีกคนหนึ่งให้คะแนนเป็น 3 ความไม่เห็นพ้องนี้จะได้น้ำหนักน้อยกว่ากรณีที่คนหนึ่งให้คะแนน 1 และอีกคนให้คะแนน 5 เพราะการประเมินในกรณีหลังแตกต่างกันมากกว่า สำหรับการตีความค่าสถิติแคปปาถ่วงน้ำหนัก สามารถใช้วิธีเดียวกับสถิติแคปปาแบบไม่ถ่วงน้ำหนัก (ตารางที่ 3)

ตัวอย่าง เช่น การวิจัยที่มุ่งศึกษาถึงการมีรอยคอดรากฟัน (root isthmus) ในส่วนต่าง ๆ ของรากฟันที่ โดยใช้ระดับการประเมิน 5 ระดับ ได้แก่ 1 ไม่มีอย่างแน่นอน 2 น่าจะไม่มี 3 ไม่ชัดเจน 4) น่าจะมี และ 5 มีอย่างแน่นอน มีการใช้สถิติแคปปาถ่วงน้ำหนักเพื่อประเมินเชื่อถือได้ระหว่างประเมิน [11]

4. สถิติฟลีสแคปปา (Fleiss Kappa, κ_f) [12]

สถิติฟลีสแคปปา เป็นสถิติที่ถูกพัฒนาขึ้นโดย Joseph Fleiss เป็นการขยายความสามารถในการคำนวณจาก Cohen's kappa สถิติฟลีสแคปปามีความสามารถที่ในการวัดความเชื่อถือได้ในข้อมูลตัวแปรระดับนามบัญญัติ กรณีที่มีผู้ประเมินมากกว่าสองคน สำหรับการตีความค่าสถิติฟลีสแคปปาถ่วงน้ำหนักสามารถใช้วิธีเดียวกับสถิติแคปปาแบบไม่ถ่วงน้ำหนัก (ตารางที่ 3)

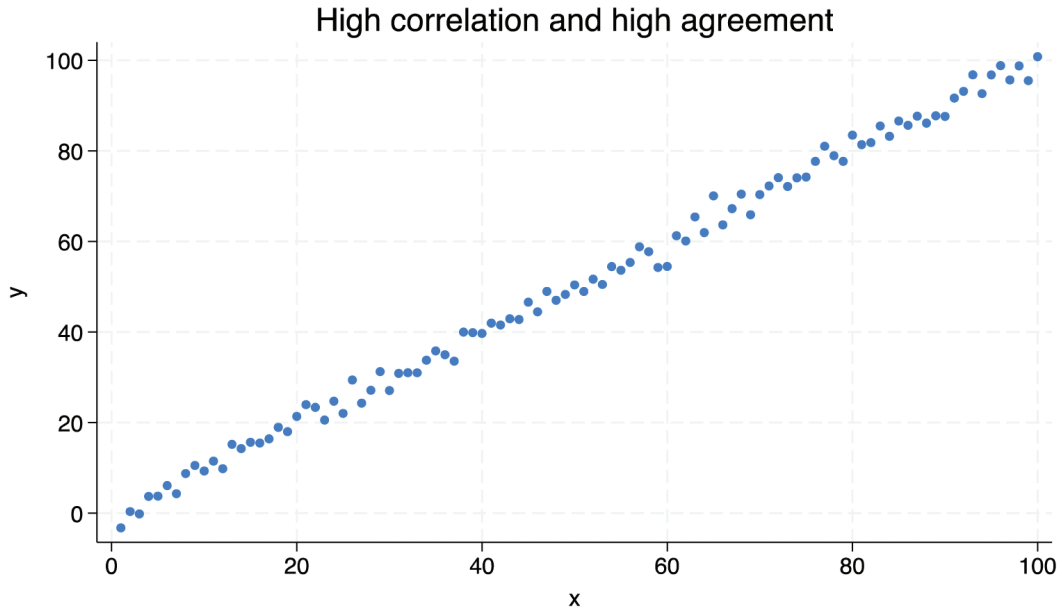
ในปัจจุบันการคำนวณ Fleiss' Kappa มักทำผ่านโปรแกรมทางสถิติ เนื่องจากการคำนวณสามารถซับซ้อนเมื่อมีผู้ประเมินหลายคนและข้อมูลจำนวนมาก โปรแกรมเหล่านี้มักมีการคำนวณที่ทำได้รวดเร็วและแม่นยำ ช่วยให้นักวิจัยสามารถวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ

5. สัมประสิทธิ์คอนคอร์ดานส์ของเคนดาล์ล (Kendall coefficient of concordance, W) [13]

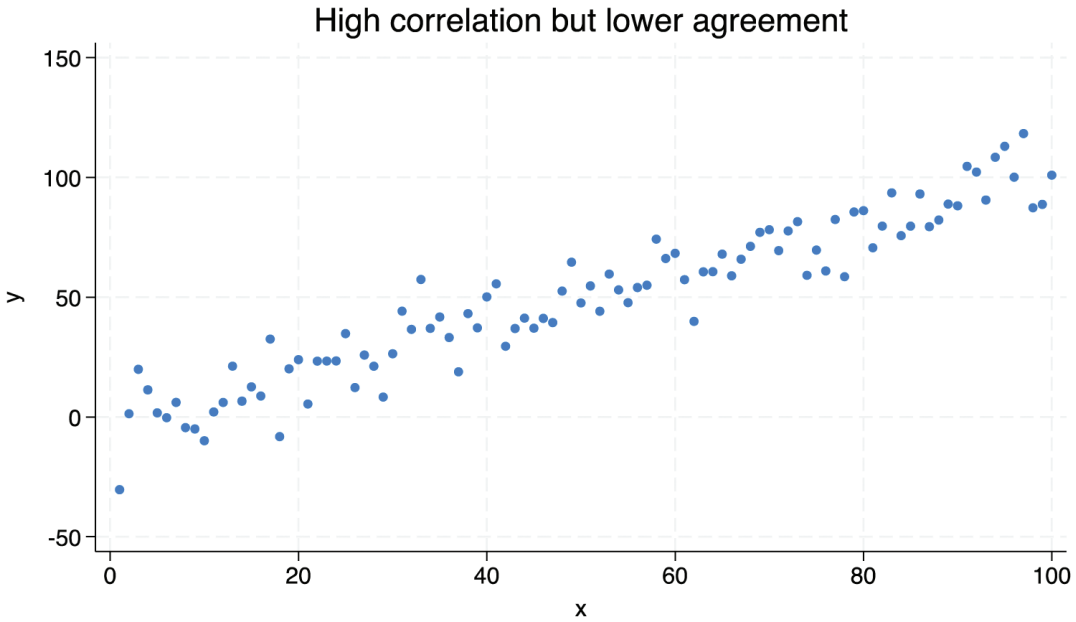
สัมประสิทธิ์คอนคอร์ดานส์ของเคนดาล์ล หรือ Kendall coefficient of concordance หรือ W ของ Kendall เป็นสถิติที่ใช้วัดความเห็นด้วย (concordance) หรือความสอดคล้องกันของข้อมูลจัดอันดับที่ได้จากผู้ประเมินหลาย ๆ คน สถิตินี้เป็นเครื่องมือที่มีประสิทธิภาพสำหรับการวัดความเห็นด้วยในการจัดอันดับข้อมูล ไม่ว่าจะเป็นการประเมินความพึงพอใจ หรือการวิจัยทางวิทยาศาสตร์ที่ต้องการความเห็นด้วยของผู้เชี่ยวชาญหลายคน ค่า W แตกต่างจากค่าแคปปาโดยที่ค่าแคปปาสามารถมีค่าเป็นลบได้ แต่ค่า W จะมีค่าระหว่าง 0 ถึง 1 เท่านั้น โดยที่ 0 หมายถึงไม่มีความเห็นด้วยเลย และ 1 หมายถึงมีความเห็นด้วยอย่างสมบูรณ์ สำหรับการตีความค่าของสัมประสิทธิ์คอนคอร์ดานส์ของเคนดาลล์สามารถใช้วิธีเดียวกับสถิติแคปปาแบบไม่ถ่วงน้ำหนัก (ตารางที่ 3)

6. วิธีของ Bland-Altman (Bland-Altman method) [14, 15]

การคำนวณค่า Pearson's product correlation coefficient (r) ถือเป็นทางเลือกที่ดูเหมือนจะมีความเหมาะสมเมื่อต้องการประเมินระดับความเชื่อถือได้ระหว่างผู้ประเมินสองคนสำหรับตัวแปรระดับอันตรภาค (interval variable) หรืออัตราส่วน (ratio variable) อย่างไรก็ตาม ค่าสหสัมพันธ์ (correlation) ที่สูง บ่งบอกถึงความสัมพันธ์เชิงเส้นที่สูงระหว่างการวัดสองครั้ง โดยค่าใกล้ 1 หรือ -1 หมายถึงความสัมพันธ์ที่มาก กรณีที่ข้อมูลมีความเชื่อถือได้หรือความสอดคล้องสมบูรณ์จะปรากฏเมื่อจุดทั้งหมดอยู่บนเส้นเท่ากัน (หรือ y=x) ซึ่งจะมีค่า r ประมาณ 0.5 (ภาพที่ 1) ดังนั้น ค่าหากค่าสหสัมพันธ์สูง แต่จุดข้อมูลกระจายตัวไม่อยู่บนเส้นเท่ากัน (หรือ y=x) อาจไม่ได้บ่งบอกถึงระดับความสอดคล้องที่ดี (ภาพที่ 2)



ภาพที่ 1: แผนภาพการกระจายข้อมูลที่ค่าสหสัมพันธ์สูงและระดับความสอดคล้องสูง



ภาพที่ 2: แผนภาพการกระจายข้อมูลที่ค่าสหสัมพันธ์สูงแต่ระดับความสอดคล้องต่ำ

ดังนั้นเพื่อแก้ปัญหาเหล่านี้และประเมินความสอดคล้องอย่างถูกต้อง การสร้างแผนภูมิของ Bland-Altman จะมุ่งเน้นไปที่ความแตกต่างระหว่างการวัดสองวิธี แทนที่จะเป็นการเปรียบเทียบโดยตรงเช่นกับการความสัมพันธ์เชิงเส้น ในแกนนอนของกราฟจะแสดงค่าเฉลี่ยของการวัดสองวิธี และแกนตั้งจะแสดงความแตกต่างระหว่างการวัดสองวิธี

ขั้นตอนในการสร้างการสร้างแผนภูมิ:

1. คำนวณค่าเฉลี่ยของการวัดสองวิธี:

สำหรับแต่ละคู่ของการวัด (x, y) คำนวณ mean = (x+y)/2

2. คำนวณความแตกต่างระหว่างการวัด: สำหรับแต่ละคู่ของการวัด (x, y)

คำนวณ difference = x - y

3. การสร้างแผนภูมิ: ใช้ค่าเฉลี่ยเป็นแกน x และความแตกต่างเป็นแกน y

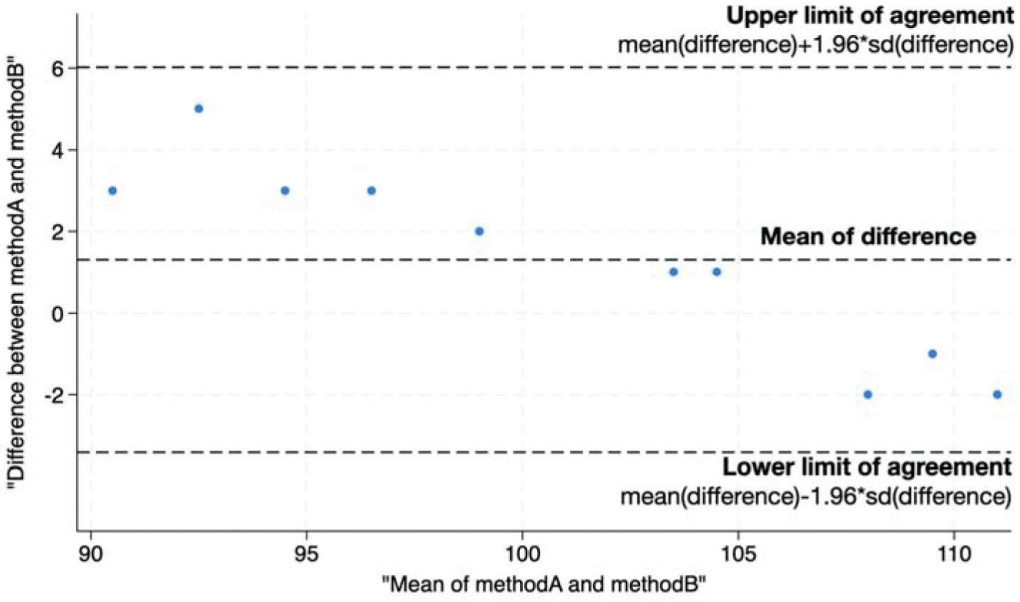
คำนวณและเพิ่มเส้นแนวนอน: คำนวณค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของความแตกต่าง แล้ววาดเส้นแนวนอนที่แสดงค่าเฉลี่ยและ 1.96 เท่าของค่าเบี่ยงเบนมาตรฐาน

ความแตกต่างเฉลี่ยในแนวแกนนอนจะแสดงความแตกต่างเฉลี่ยระหว่างการวัดสองวิธี หากความแตกต่างนี้เป็น 0 หมายความว่ามีความสอดคล้องที่ดีระหว่างการวัดสองวิธี ส่วนเส้น 95% limits of agreement) แสดงขอบเขตบนและล่าง (upper and lower limit) ซึ่งเป็นระยะที่คาดว่าจะครอบคลุมความแตกต่าง 95% ระหว่างการวัดสองวิธี หากข้อมูลที่ต้องการอยู่นอกเส้นเหล่านี้อาจบ่งบอกถึงปัญหาที่เกิดขึ้นในการวัด

โดยสรุป การอ่านและตีความแผนภูมิ Bland-Altman นั้นควรคำนึงถึงปัจจัยหลายประการ ดังนี้

1. ความแตกต่างระหว่างการวัดสองวิธี: แกนตั้งของแผนภูมิแสดงความแตกต่างระหว่างการวัดสองวิธี ซึ่งผู้อ่านควรสังเกตว่าความแตกต่างเหล่านี้กระจายอย่างไร หากความแตกต่างส่วนใหญ่อยู่ใกล้ศูนย์ นั่นหมายถึงความสอดคล้องที่ดีระหว่างสองวิธีการวัด

2. ค่าเฉลี่ยของการวัดสองวิธี: แกนนอนแสดงค่าเฉลี่ยของการวัดทั้งสองวิธี ควรสังเกตว่าจุดข้อมูลเหล่านี้กระจายอย่างไรตามแกนนอน



ภาพที่ 3: ตัวอย่างการสร้างแผนภูมิของ Bland-Altman

3. ความชันและการกระจายของจุดต่าง ๆ: หากจุดข้อมูลกระจายตัวอย่างสม่ำเสมอ นั่นหมายถึงความสม่ำเสมอในการวัด หากข้อมูลมีการกระจายที่ไม่สม่ำเสมอหรือแผนภาพมีความชัน อาจบ่งชี้ถึงความแปรปรวนในการวัด

4. เส้นขอบเขตบนและล่าง: แสดงความแตกต่างสูงสุดที่ยอมรับได้ระหว่างการวัดสองวิธี หากจุดข้อมูลส่วนใหญ่อยู่ภายในเส้นนี้ จะบ่งบอกถึงความสอดคล้องที่ดี

7. สัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation, ICC)

สัมประสิทธิ์สหสัมพันธ์ภายในชั้น เป็นเครื่องมือสถิติที่ใช้ใช้ในการประเมินความเชื่อถือได้หรือความสอดคล้องของการวัดในระหว่างผู้ประเมินหลายคน หรือการวัดในครั้งต่าง ๆ ด้วยผู้ประเมินเดียวกัน โดยเฉพาะเมื่อข้อมูลมีระดับการวัดที่เป็นอันตรภาคหรือสัดส่วน เช่น การหาความเชื่อถือได้ในการวัดขนาดของรอยโรคระหว่างผู้ประเมิน 2 คน [2] ตัวอย่างเช่นงานวิจัยเพื่อประเมินความเชื่อถือได้ของการวัดความหนาของรากฟัน 3 วิธี [16]

สัมประสิทธิ์สหสัมพันธ์ภายในชั้นเป็นตัวบ่งชี้ความสอดคล้องในข้อมูลประเภทต่อเนื่อง ไม่ว่าจะเป็นการประเมินความเชื่อถือได้ของการทดสอบซ้ำ ความเชื่อถือได้ระหว่างผู้ประเมิน หรือความเชื่อถือได้ภายในผู้ประเมิน โดยมีค่าตั้งแต่ 0 จนถึง 1 หากค่า ICC ที่ใกล้เคียง 1 แสดงว่ามีความสอดคล้องสูง แต่ถ้าค่า ICC ใกล้ 0 หมายความว่าข้อมูลทั้งสองกลุ่มไม่สอดคล้องกัน เช่น การประเมินขนาดของรอยโรคบนรากฟัน ในกรณีที่ ICC ใกล้เคียงค่าศูนย์ อาจสื่อถึงว่าผู้ประเมินแต่ละคนมีการประเมินขนาดของรอยโรคบนรากฟันที่แตกต่างกันอย่างมาก หรือว่าผู้ประเมินเดียวกันในการวัดซ้ำมีผลลัพธ์ที่แตกต่างกันมากในแต่ละครั้ง

7.1 ขั้นตอนการเลือกใช้แบบจำลองในการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้นสำหรับการศึกษาความเชื่อถือได้ระหว่างผู้ประเมิน

เมื่อต้องการวิเคราะห์ความเชื่อถือได้ในการประเมินผ่านสัมประสิทธิ์สหสัมพันธ์ภายในชั้น การเลือกแบบจำลอง (model) ที่เหมาะสมเป็นสิ่งสำคัญ [17] เพื่อให้ได้ผลวิเคราะห์ที่แม่นยำและสะท้อนความเป็นจริง มีขั้นตอนดังนี้ [2]

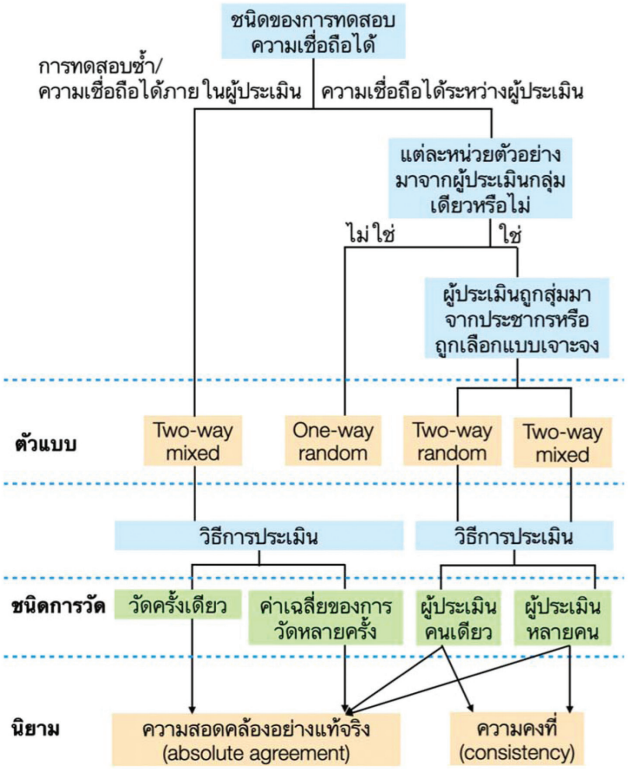
1. ต้องการรู้ว่าในแต่ละหน่วยตัวอย่าง มีผู้ประเมินที่เป็นกลุ่มเดียวกันทำการประเมินหรือไม่

2. กลุ่มของผู้ประเมินที่เราเลือกมาเป็นกลุ่มที่ถูกเลือกมาโดยการสุ่มจากประชากรหรือเป็นกลุ่มที่มีความเฉพาะเจาะจง

3. เราต้องการวิเคราะห์ความเชื่อถือได้ของผู้ประเมินคนเดียวหรือหลายคน

4. ความสนใจนิยามความคงที่ (consistency) หรือความสอดคล้องอย่างแท้จริง (absolute agreement)

จากข้อ 1 และ 2 นั้นจะช่วยเราเลือกแบบจำลอง (model) ที่เหมาะสม ส่วนข้อ 3 จะช่วยเลือกว่าควรวิเคราะห์ผู้ประเมินประเภทใด (type) และข้อ 4 จะช่วยแนะนำในการเลือกประเภทของการประเมินที่เราสนใจตามนิยาม (definition) ดังจะที่จะกล่าวถึงต่อไปนี้



ภาพที่ 4: แผนภาพแสดงกระบวนการเลือกแบบจำลองสัมประสิทธิ์สหสัมพันธ์ภายในชั้นตามการออกแบบการทดลองของการศึกษาความเชื่อถือได้จาก Koo TK และ Li M Y (2016)

1. เลือกตัวแบบ (Model)

a. One-Way Random-Effects Model

ข้อมูลแต่ละตัวอย่างจะถูกประเมินโดยผู้ประเมินหลายคนหรือกลุ่มผู้ประเมินที่แตกต่างกัน โดยผู้ประเมินหรือกลุ่มผู้ประเมินจะถูกสุ่มมาจากประชากรขนาดใหญ่ ซึ่งตัวแบบประเภทนี้ จะไม่นิยมใช้ในงานวิจัยทางการแพทย์ เนื่องจากในการศึกษาในงานวิจัยทางการแพทย์ควรจะใช้ผู้ประเมินเพียงกลุ่มเดียว ซึ่งจะช่วยให้มั่นใจได้ว่ามีมาตรฐานเดียวกันในการประเมิน ผู้ประเมินทุกคนได้รับการฝึกฝนหรือมีความเข้าใจในเกณฑ์การประเมินที่เหมือนกัน

b. Two-Way Random-Effects Model

แต่ละตัวอย่างจะถูกประเมินโดยกลุ่มผู้ประเมินเดียวกัน โดยกลุ่มผู้ประเมินจะถูกสุ่มมาจากประชากรขนาดใหญ่ ตัวแบบประเภทนี้เป็นที่นิยมใช้กัน และสามารถนำไปใช้กับผู้ประเมินทั่วไปที่มีลักษณะเดียวกัน(generalization) เช่น การประเมินความสำเร็จของการรักษาคลองรากฟัน โดยใช้กลุ่มผู้ประเมินที่สุ่มมาจากหลายสถาบัน หลายคลินิก หรือทันตแพทย์ทั่วไป ความเชื่อถือได้ของผลการประเมินนี้จะสะท้อนถึงผู้ประเมินอื่น ๆ ทั่วไปได้

c. Two-Way Mixed-Effects Model

ควรใช้ตัวแบบประเภท Two-Way Mixed-Effects เมื่อผู้ประเมินที่ถูกเลือกเป็นผู้ประเมินที่สนใจเฉพาะ (rater of interest) ผลที่ได้จะแทนความเชื่อถือได้ของผู้ประเมินเฉพาะที่เข้าร่วมการทดลองหรือวิจัยเท่านั้น ซึ่งผลไม่สามารถนำไปใช้กับผู้ประเมินอื่น ๆ แม้ว่าจะมีลักษณะเดียวกันกับผู้ประเมินที่ถูกเลือก เช่น การประเมินความสำเร็จของการรักษาคลองรากฟัน โดยทันตแพทย์เฉพาะทางเอ็นโดดอนติกส์ ผลลัพธ์ที่ได้จะแสดงความเชื่อถือได้ของการประเมินโดยกลุ่มนี้เท่านั้น และไม่สามารถนำไปใช้กับทันตแพทย์ทั่วไปหรือผู้ประเมินที่มีความเชี่ยวชาญแตกต่างได้

2. เลือกประเภท (Type)

การเลือกประเภทการวัดสหสัมพันธ์นั้นขึ้นอยู่กับวิธีการวัดในเกณฑ์วิธี (Protocol) การวิจัยที่จะนำไปใช้งานจริง เพื่อให้เข้ากับการใช้งานที่เกิดขึ้นจริง ตัวอย่างเช่น ถ้าเราวางแผนที่จะใช้ค่าเฉลี่ยของ ผู้ประเมิน 3 คนเป็นฐานในการประเมิน ควรเลือกประเภท “ค่าเฉลี่ยของ k ผู้ประเมิน (ผู้ประเมินมากกว่า 1 คน)” ในทางกลับกัน ถ้าเราวางแผนที่จะใช้การวัดจากผู้ประเมินคนเดียวเป็นฐานการวัดจริง ควรเลือกประเภท “ผู้ประเมินคนเดียว” แม้ในการศึกษาความเชื่อถือได้จะมีผู้ประเมิน 2 คนหรือมากกว่าก็ตาม

3. เลือกนิยาม (Definition)

สำหรับทั้ง 2 รูปแบบ คือ 2-way random-effects และ 2-way mixed-effects มีการนิยาม ICC อยู่ 2 ประเภท คือ “ความสอดคล้องอย่างแท้จริง” และ “ความคงที่” การเลือกนิยาม ICC ขึ้นอยู่กับว่าเราคิดว่าความสอดคล้องอย่างแท้จริง หรือความคงที่ของผู้ประเมินเป็นสิ่งที่สำคัญกว่ากัน

ความสอดคล้องอย่างแท้จริง เกี่ยวกับว่าผู้ประเมินที่ต่างกันจะกำหนดคะแนนเดียวกันให้กับหัวข้อเดียวกันหรือไม่ เช่น ถ้าผู้ประเมิน A ให้คะแนน 4 ในหัวข้อหนึ่ง ผู้ประเมิน B ก็ควรจะให้คะแนน 4 เช่นกันเพื่อให้เกิดความสอดคล้องอย่างแท้จริง ในขณะที่ความคงที่ เกี่ยวข้องกับว่าคะแนนของผู้ประเมินในกลุ่มหัวข้อที่เดียวกันนั้นมีความสัมพันธ์ในแบบบวก (additive) หรือไม่ เช่น ถ้าผู้ประเมิน A ให้คะแนน 4 และผู้ประเมิน B มักจะให้คะแนนที่สูงกว่า A โดยเฉลี่ย 2 คะแนน คะแนนของ B ควรเป็น 6 ในที่นี้ความสัมพันธ์ระหว่างคะแนนของ A และ B สามารถแสดงได้โดยสูตร y (คะแนนของ B) = x (คะแนนของ A) + c (ความคลาดเคลื่อนที่เป็นระบบ) ในขณะที่ความสอดคล้องอย่างแท้จริง เกี่ยวกับขนาดที่ค่า y เท่ากับ x เพียงอย่างเดียว

7.2 ขั้นตอนการเลือกใช้แบบจำลองในการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้นสำหรับ การศึกษาความเชื่อถือได้ภายในผู้ประเมิน (Intra-rater Reliability) และความเชื่อถือได้ของการทดสอบซ้ำ (Test-retest Reliability)

ขั้นตอนการเลือกใช้แบบจำลองในการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้นสำหรับ การศึกษาความเชื่อถือได้ภายในผู้ประเมินและความเชื่อถือได้ของการทดสอบซ้ำ จะเลือกใช้ตัวแบบ Two-Way Mixed-Effects เท่านั้น เนื่องจากเป็นการประเมินจากบุคคลหนึ่ง ๆ ไม่จำเป็นที่จะนำผลการประเมินของผู้ประเมินคนเดียวไปใช้กับผู้ประเมินทั่วไปอื่น ๆ เพราะไม่สามารถใช้คะแนนของผู้ประเมินคนเดียวเป็น

ตัวแทนสำหรับผู้ประเมินหลายคนได้ โดยปัจจัยหลักที่ต้องพิจารณาคือเป็นการวัดครั้งเดียว (single measurement) หรือเฉลี่ยจากการวัดหลายครั้ง (multiple measurements) สำหรับแบบจำลองทั้งสองแบบนี้จะสามารถเลือกใช้รูปแบบนิยามของสัมประสิทธิ์สหสัมพันธ์ภายในชั้นได้เพียงรูปแบบนิยามเดียวเท่านั้น คือ ความสอดคล้องอย่างแท้จริง

การแปลผลการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้น มีค่าได้อยู่ระหว่าง 0 ถึง 1 โดยสามารถแปลความหมายได้ดังนี้ [2]

กล่าวโดยสรุปคือขั้นตอนการเลือกรูปแบบสำหรับการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้นสำหรับการศึกษาความเชื่อถือได้ประกอบด้วย การรู้ประเภทของการศึกษาความเชื่อถือได้ที่กำลังดำเนินการ และการตัดสินใจเกี่ยวกับ แบบจำลอง ชนิด และนิยามของการประเมินนั้น ๆ

5. ข้อพิจารณาทางปฏิบัติสำหรับการทำการทดสอบความเชื่อถือได้ในการวิจัยทางเอนโดดอนติกส์

5.1 การกำหนดขนาดตัวอย่างสำหรับการทดสอบความเชื่อถือได้

การกำหนดขนาดตัวอย่างสำหรับการทดสอบความเชื่อถือได้ ก่อนทำการศึกษาหลักมักขึ้นอยู่กับหลายปัจจัย และไม่มีขนาดตายตัวที่ใช้กันทั่วไป อย่างไรก็ตามขนาดตัวอย่างสำหรับการทดสอบความเชื่อถือได้ คือ จำนวน 10-30 ตัวอย่าง [18] แต่ถ้าการวิจัยต้องการความแม่นยำสูงหรือมีปัจจัยมากมายที่ต้องการวิเคราะห์ ขนาดตัวอย่างอาจต้องใหญ่มากขึ้น

สิ่งสำคัญที่สุดคือต้องทำการวิเคราะห์พลัง (power analysis) เพื่อกำหนดขนาดตัวอย่างที่เฉพาะเหมาะสมสำหรับการศึกษานั้น ๆ และการปรึกษากับนักสถิติเป็นสิ่งที่ควรทำอย่างยิ่ง นอกจากนี้การใช้ซอฟต์แวร์สถิติ เช่น โปรแกรมเช่น R SAS STATA หรือ SPSS ที่มีฟังก์ชันสำหรับการกำหนดขนาดตัวอย่าง เป็นวิธีที่ดีที่สุดในการคำนวณขนาดตัวอย่างสำหรับสถิติเหล่านี้

5.2 การเก็บข้อมูลสำหรับทดสอบความเชื่อถือได้

การเก็บข้อมูลสำหรับทดสอบความเชื่อถือได้ มีหลักการและแนวทางที่ควรปฏิบัติตามเพื่อให้ได้ผลลัพธ์ที่มีความน่าเชื่อถือ ดังนี้:

1. การเลือกกลุ่มตัวอย่าง

กลุ่มตัวอย่างควรมีความหลากหลายเพื่อให้สามารถเป็นตัวแทนที่ดีของประชากร นอกจากนี้ตัวอย่างควรมีการกระจายของคะแนนหรือมีตัวแปรที่ครอบคลุมช่วงต่างๆ ของข้อมูลที่คาดว่าจะพบในการศึกษาหลัก เช่น ถ้าผลการประเมินเป็นตัวแปรระดับจัดลำดับ (เช่น คะแนน) ต้องให้แน่ใจว่ามีการกระจายคะแนนแต่ละระดับครอบคลุม ไม่มีเพียงแค่คะแนนใดคะแนนหนึ่ง ตัวอย่างที่ครอบคลุมทุกระดับ ถ้าการประเมินผลการรักษามี 3 ระดับ (healed healing disease) ตัวอย่างที่จะควรนำมาใช้ในการประเมินก็ควรมีผลการรักษาหายทั้ง 3 รูปแบบ

2. การวัดหลายครั้งและการหาค่าเฉลี่ย สำหรับข้อมูลเชิงปริมาณ,

สำหรับข้อมูลเชิงปริมาณ การวัดอย่างน้อย 3 ครั้งมักถูกนำมาใช้เพื่อลดผลกระทบของความผิดพลาดแบบสุ่มหรือความผิดพลาดในการวัด หลังจากที่ได้รับข้อมูลจากการวัดหลายครั้งแล้ว การคำนวณค่าเฉลี่ยของการวัดเหล่านี้เป็นวิธีที่ดีในการเป็นตัวแทนของค่าในการวัดข้อมูลนั้น ๆ [19]

5.3 บทบาทของผู้เชี่ยวชาญหรือมาตรฐานสูงสุด (gold standard) และการเทียบมาตรฐาน (calibration)

การมีบทบาทของผู้เชี่ยวชาญหรือมาตรฐานสูงสุด และการเทียบมาตรฐาน มีความสำคัญอย่างมากในการเพิ่มความเชื่อถือได้ในการทำวิจัย โดยเฉพาะก่อนที่จะทำการทดสอบความเชื่อถือได้ ดังนี้ ผู้เชี่ยวชาญมีบทบาทในการกำหนดเกณฑ์หรือการตั้งค่ามาตรฐานสำหรับการทดสอบหรือการวัด ซึ่งผู้ช่วยชาญสามารถให้คำแนะนำและช่วยเหลือในการออกแบบวิธีการวัดที่มีความน่าเชื่อถือ นอกจากนี้มาตรฐานสูงสุดที่กำหนดไว้ยังใช้เป็นเกณฑ์อ้างอิงในการเปรียบเทียบและประเมินความเชื่อถือได้ของการวัดในการวิจัย

นอกจากนี้ การฝึกอบรม (Training) ให้ผู้แก่ทำการวัดหรือผู้รวบรวมข้อมูลเป็นสิ่งสำคัญเช่นกัน เพื่อให้พวกเขามีความเข้าใจและความสามารถในการทำการวัดหรือรวบรวมข้อมูลอย่างถูกต้องและสม่ำเสมอ จากนั้นการเทียบมาตรฐานของผู้วัดกับผู้เชี่ยวชาญหรือมาตรฐานสูงสุดเป็นอีกกระบวนการที่จำเป็นเพื่อให้แน่ใจว่าผู้วัดมีความแม่นยำรวมถึงเพิ่มความเชื่อถือได้ของการวัด

5.4 การบันทึกข้อมูลและการเตรียมข้อมูลสำหรับการวิเคราะห์โดยใช้โปรแกรมทางสถิติ

หลักการบันทึกข้อมูลและการเตรียมข้อมูลสำหรับการวิเคราะห์โดยใช้โปรแกรมทางสถิติเป็นกระบวนการที่สำคัญมากในการวิจัย โดยข้อมูลควรถูกบันทึกอย่างเป็นระเบียบและตามแบบที่กำหนดไว้ล่วงหน้า โดยปกติข้อมูลจะถูกบันทึกในตารางที่มีแถวและคอลัมน์ชัดเจน ซึ่งแต่ละคอลัมน์แสดงถึงการวัดผู้ประเมินคนเดียวกันหรือต่างผู้ประเมิน แต่ละแถวแสดงถึงผลการวัดในแต่ละครั้ง ข้อมูลควรถูกบันทึกในรูปแบบที่สามารถนำไปใช้กับโปรแกรมทางสถิติได้ง่าย เช่น Excel CSV ตรวจสอบความถูกต้องของข้อมูล ไม่มีข้อมูลที่ขาดหาย แปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ เช่น การเปลี่ยนข้อมูลจากรูปแบบข้อความเป็นตัวเลข จากนั้นใช้เทคนิคทางสถิติเพื่อวิเคราะห์ข้อมูลที่ โดยการใช้โปรแกรมทางสถิติ เช่น SPSS (Statistical Package for the Social Sciences) R SAS (Statistical Analysis System) STATA เป็นต้น

5.5 ช่วงเวลาที่สามารถดำเนินการทดสอบความเชื่อถือได้

ในการตรวจสอบว่าเครื่องมือวัดหรือข้อมูลที่ใช้ในการวิจัยนั้นมีความสม่ำเสมอและน่าเชื่อถือหรือไม่ มีหลายช่วงเวลาที่สามารถดำเนินการทดสอบความเชื่อถือได้ โดยแต่ละช่วงเวลานั้นมีวัตถุประสงค์และเหตุผลที่แตกต่างกัน

1. ก่อนเริ่มงานวิจัย (Pre-study Assessment): เป็นการตรวจสอบว่าการวัดที่จะใช้ในการวิจัยมีความเหมาะสมและน่าเชื่อถือหรือไม่ ซึ่งช่วยให้มั่นใจได้ว่าข้อมูลที่จะเก็บรวบรวมนั้นจะมีความถูกต้องและเชื่อถือได้

2. ระหว่างการวิจัย (Ongoing Assessment): เป็นการตรวจสอบความเชื่อถือได้ของการวัดตลอดระยะเวลาการวิจัย เพื่อตรวจสอบว่าเครื่องมือนั้นยังคงมีความเชื่อถือได้อยู่หรือไม่ ซึ่งช่วยลดความผิดพลาดและความไม่แน่นอนในการวัด

3. หลังจบการวิจัย (Post-study Assessment): ช่วยให้สามารถประเมินความเชื่อถือได้ของเครื่องมือวัดหลังจากที่ข้อมูลทั้งหมดได้รับการเก็บรวบรวมแล้ว ซึ่งช่วยให้เห็นภาพรวมของความสม่ำเสมอของเครื่องมือตลอดระยะเวลาการวิจัย

4. การวิเคราะห์ช่วงกลาง (Interim Analysis): การทำการวิเคราะห์ช่วงกลางในระหว่างการวิจัยเป็นการตรวจสอบและประเมินความเชื่อถือได้ของข้อมูลที่เก็บรวบรวมมาแล้ว ซึ่งช่วยให้สามารถทำการปรับเปลี่ยนหรือแก้ไขข้อบกพร่องในกระบวนการวิจัยได้ทันท่วงที

อย่างไรการทำการทดสอบความเชื่อถือได้ก่อนเริ่มงานวิจัย มักถือว่าสำคัญมากที่สุด เนื่องจากการทดสอบล่วงหน้าช่วยให้สามารถระบุและแก้ไขปัญหาในการวัดหรือขั้นตอนการวิจัยก่อนที่จะเริ่มเก็บข้อมูลจริง ซึ่งเป็นขั้นตอนที่สำคัญในการรักษาความถูกต้องและความเชื่อถือได้ของการวิจัยทั้งหมด การทดสอบล่วงหน้าช่วยให้วิจัยเริ่มต้นด้วยการวัดที่มีคุณภาพสูงและความน่าเชื่อถือ ทำให้สามารถเชื่อมั่นได้ว่าข้อมูลที่เก็บรวบรวมจะเป็นตัวแทนที่ดีของสิ่งที่ต้องการวัดจริง

5.6 แนวทางการรายงาน (Reporting guidelines)

ในการเขียนรายงานการวิจัย แนวทางการรายงานเป็นเครื่องมือที่สำคัญในการสร้างความเชื่อมโยงและความชัดเจนในการนำเสนอผลลัพธ์การวิจัย โดยเฉพาะในการศึกษาวิจัยทางวิทยาเอ็นโดดอนต์ เช่น PROBE 2023 guidelines for reporting observational studies in endodontics [20] หรือ PRIRATE 2020 guidelines for reporting randomized trials in Endodontics [21] แนวทางเหล่านี้มักจะเน้นย้ำถึงความสำคัญของการรายงานมาตรฐานในการวัดผลและความเชื่อถือได้ของการวัด ซึ่งเป็นปัจจัยสำคัญที่ส่งผลต่อคุณภาพและความเชื่อถือได้ของผลการวิจัย การระบุวิธีการวัด การทดสอบความเชื่อถือได้ของการวัด และการอธิบายวิธีการวิเคราะห์ข้อมูลอย่างละเอียดช่วยให้ผู้อ่านสามารถเข้าใจและตรวจสอบผลการวิจัยได้อย่างเชื่อถือได้

บทสรุป

บทความนี้มุ่งเน้นที่การทบทวนเรื่องความเชื่อถือได้และการประเมินความเชื่อถือได้ในเชิงวิจัยและการประเมินความเชื่อถือได้ถือโดยเป็นปัจจัยสำคัญในการกำหนดว่าข้อมูลหรือผลการวิจัยนั้นมีความถูกต้องและสามารถนำมาใช้ประโยชน์ในการตัดสินใจหรือการสร้างองค์ความรู้ในด้านต่าง ๆ ได้หรือไม่ทั้งนี้ ความเชื่อถือได้มีประโยชน์มากในการสร้างความเชื่อถือได้ต่อผลการวิจัยและเป็นฐานที่สำคัญในการพัฒนาด้านการศึกษาวิจัยทางเอ็นโดดอนติกส์ต่อไป

เอกสารอ้างอิง

1. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. **Sports Med**. 1998;26(4):217-38.
2. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. **J Chiropr Med**. 2016;15(2):155-63.
3. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. **Res Social Adm Pharm**. 2013;9(3):330-8.
4. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. **Tutor Quant Methods Psychol**. 2012;8(1):23-34.
5. Stemler S. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. **Pract Assess Res Eval**. 2004;9:1-19.
6. Lam EWN, Law AS, Nguyen RHN, Basile S, Austah O, Gilbert GH, et al. Interexaminer Agreement in the Radiologic Identification of Apical Periodontitis/Rarefying Osteitis in the National Dental Practice-Based Research Network Predict Endodontic Study. **J Endod**. 2021;47(10):1575-82.
7. Cohen J. A Coefficient of Agreement for Nominal Scales. **Educ Psychol Meas**. 1960;20(1):37-46.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. **Biom J**. 1977;33(1):159-74.
9. Al-Manei KK. Radiographic Quality of Single vs. Multiple-Visit Root Canal Treatment Performed by Dental Students: A Case Control Study. **Iran Endod J**. 2018;13(2):149-54.
10. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychol Bull**. 1968;70(4):213-20.
11. Barros de Oliveira ML, Junqueira RB, Kamburoglu K, Eratam N, Cakmak EE, Sonmez G, et al. Assessment of the Metal Artifact Reduction Tool for the Detection of Root Isthmus in Mandibular Molars with Intraradicular Posts in Cone-beam Computed Tomographic Scans. **J Endod**. 2021;47(10):1583-91.
12. Fleiss JL. Measuring nominal scale agreement among many raters. **Psychol Bull**. 1971;76(5):378.
13. Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences: McGraw-Hill; 1988.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. **Lancet**. 1986;1(8476):307-10.
15. Bland JM, Altman DG. Measuring agreement in method comparison studies. **Stat Methods Med Res**. 1999;8(2):135-60.

16. Sutam N, Jantarat J, Ongchavalit L, Sutimuntanakul S, Hargreaves KM. A Comparison of 3 Quantitative Radiographic Measurement Methods for Root Development Measurement in Regenerative Endodontic Procedures. **J Endod**. 2018;44(11):1665-70.
17. McGraw K, Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. **Psychol Methods**. 1996;1:30-46.
18. Bujang MA. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. **Arch Orofac Sci**. 2017;12:1-11.
19. Koppenhaver SL, Parent EC, Teyhen DS, Hebert JJ, Fritz JM. The effect of averaging multiple trials on measurement error during ultrasound imaging of transversus abdominis and lumbar multifidus muscles in individuals with low back pain. **J Orthop Sports Phys Ther**. 2009;39(8):604-11.
20. Nagendrababu V, Duncan HF, Fouad AF, Kirkevang LL, Parashos P, Pigg M, et al. PROBE 2023 guidelines for reporting observational studies in endodontics: Explanation and elaboration. **Int Endod J**. 2023;56(6):652-85.
21. Nagendrababu V, Duncan HF, Bjorndal L, Kvist T, Priya E, Jayaraman J, et al. PRIRATE 2020 guidelines for reporting randomized trials in Endodontics: a consensus-based development. **Int Endod J**. 2020;53(6):764-73.