Journal of Research Methodology, Volume 14, Number 1 (January - April 2001)

วารสารวิธีวิทยาการวิจัย ปีที่ 14 ฉบับที่ 1 (มกราคม - เมษายน 2544)

Modelling in Epidemiology in the Past Two Decades

Virasakdi Chongsuvivatwong

ABSTRACT

This article aims to link basic concepts of epidemiology with theories and philosophy of research. Epidemiology emphasizes on quantitative empirical data. It emerged from investigation of health problems and diseases outbreak in population. In late 1800s, the revolution of science in microbiology field distracted scientists from utilizing epidemiological methods. Epidemiology discipline came up again when chronic diseases became major health problems with muti-factorial causes in the twentieth century.

The effects of putative factors on health have to be ruled out from biases, random errors and confounding. Methods to control confounding includes randomized controlled trial, stratification, matching and mathematical modeling.

Mathematical models commonly used in epidemiology include Gaussian regression, logistic regression and Cox regression. In the near future, further development will be in the area of logitudinal data analysis, exact methods and Bayesian analysis.

Journal of Research Methodology, Volume 14, Number 1 (January - April 2001)

วารสารวิธีวิทยาการวิจัย ปีที่ 14 ฉบับที่ 1 (มกราคม - เมษายน 2544)

การจัดรูปแบบจำลองทางระบาดวิทยาในรอบสองทศวรรษที่ผ่านมา

วีระศักดิ์ จงสู่วิวัฒน์วงศ์

บทคัดย่อ

บทความนี้ต้องการให้ผู้อ่านเกิดความเข้าใจพื้นฐานเรื่องการวิจัยทางระบาดวิทยา โดยเน้นการประสานทฤษฎีและปรัชญาการวิจัยกับรูปธรรมและวิวัฒนาการของวิธีวิทยาการวิจัยทางระบาดวิทยา

ระบาดวิทยาใช้การวิจัยเชิงปริมาณและเชิงประจักษ์เป็นหลัก เริ่มต้นก่อกำเนิดขึ้นหลังปรัชญา inductivism โดยอาศัยการเก็บข้อมูลด้านสุขภาพจากประชากรขนาดใหญ่ โดยเฉพาะการศึกษาโรคระบาดเพื่อหาปัจจัยเสี่ยงด้านสิ่งแวดล้อมและพฤติกรรม การพัฒนาทางจุลชีววิทยาในปลายศตวรรษที่ 19 ได้หันเหความสนใจของนักวิทยาศาสตร์ออกไปจากวิธีการทางระบาดวิทยาแต่วิธีวิทยาการวิจัยนี้ได้รับการฟื้นฟูอีกครั้งหนึ่งในต้นศตวรรษที่ 20 เมื่อปัญหาทางสาธารณสุขเกิดจากโรคเรื้อรังที่มีเหตุปัจจัยหลายอย่างร่วมกัน

ในการอธิบายความสัมพันธ์ระหว่างปัจจัยต่าง ๆ กับสุขภาพนั้น อาจจะมีข้อผิดพลาด (error) จากความบังเอิญ, จากความลำเอียง และ จาก confounding วิธีที่ใช้ในการควบคุม confounding มีหลายประการ เช่น การออกแบบ randomized controlled trial, การจัดชั้น (stratification), การจับคู่ (matching) และการสร้างรูปแบบจำลองทางคณิตศาสตร์

รูปแบบจำลองหรือสมการถดถอยที่ใช้มากทางระบาดวิทยาได้แก่ Gaussian regression, logistic regression และ Cox regression ในปัจจุบันวงการระบาดวิทยากำลังพัฒนารูปแบบจำลองสำหรับการติดตามบุคคลระยะยาว, การคำนวณแบบ exact และแบบ Bayesian ซึ่งคงจะใช้มากขึ้นในศตวรรษนี้

วัตถุประสงค์

บทความนี้เป็นส่วนหนึ่งของเอกสารประกอบการประชุมวิชาการเรื่องวิธีวิทยาการวิจัย ซึ่งเป็นการประชุมสหวิทยาการ ผู้เขียนคาดว่าผู้เข้าประชุมและผู้อ่านบทความจะมาจากหลายสาขาวิชาโดยที่ทุกคนล้วนเป็นนักวิธีวิทยาการวิจัยในสาขาวิชาของตน ประโยชน์ที่จะได้จากการฟังบรรยายหรือการอ่านบทความนี้จึงอยู่ที่การนำแนวคิดกว้าง ๆ จากระบาดวิทยาไปประยุกต์ในสาขาวิชาต่าง ๆ มากกว่าการเจาะลงไปในรายละเอียดแต่ละเรื่องของวิธีวิทยาการวิจัยทางระบาดวิทยา

เพื่อให้บรรลุถึงจุดมุ่งหมายทั่วไปดังกล่าว เนื้อหาของบทความนี้ในด้านหนึ่งแล้วจึงเริ่มต้นด้วยการทำความเข้าใจพื้นฐานเกี่ยวกับการวิจัยทางระบาดวิทยา ควบคู่ไปกับการอ้างทฤษฎีและปรัชญาการวิจัยซึ่งนักวิจัยทั่วไปคงจะซาบซึ้งอยู่แล้ว จากนั้นจะยกรูปธรรมวิวัฒนาการทางทฤษฎีและปฏิบัติของทฤษฎีด้านระบาดวิทยา และลงท้ายด้วยการนำเสนอเนื้อหาวิธีการใหม่ ๆ ซึ่งก้าวหน้ากว่าวิธีวิทยาการวิจัยที่สอนอยู่ในหลักสูตรระบาดวิทยาส่วนใหญ่ในปัจจุบัน เนื้อหาที่มีโครงสร้างเช่นนี้น่าจะเป็นโครงร่างและมีบทบาทกระตุ้นให้เกิดการถกเถียงอภิปรายและค้นคว้าเพิ่มเติมของแต่ละกลุ่ม ซึ่งเป็นจุดมุ่งหมายสูงสุดของการนำเสนอครั้งนี้

ขอบเขตของวิชาระบาดวิทยา

ระบาดวิทยา หรือ วิทยาการระบาด เป็นวิชาที่ว่าด้วยการวิจัยด้านสุขภาพเพื่อศึกษาการแจกแจงสภาวะสุขภาพ (ทั้งดีและเลว) ในระดับประชากร, หาปัจจัยที่มีผลต่อสภาวะนั้น ตลอดจนทดลองเพื่อป้องกันหรือแก้ปัญหา

ระบาดวิทยาเป็นวิชาที่เน้นวิธีวิทยาการวิจัยมากที่สุดวิชาหนึ่งในเชิงวิทยาศาสตร์สุขภาพ เนื่องจากประชากรที่สนใจมีขนาดใหญ่ สภาวะด้านสุขภาพมีหลายมิติ ปัจจัยที่มีผลต่อสภาวะสุขภาพมีหลากหลายปัจจัย วิธีวิทยาการวิจัยด้านระบาดวิทยาจึงซับซ้อนและต้องการการพัฒนาอย่างต่อเนื่อง

ระบาดวิทยาใช้การวิจัยเชิงปริมาณและเชิงประจักษ์เป็นหลัก วิธีการทางสถิติจึงเป็นพื้นฐานที่สำคัญ นักวิธีวิทยาการวิจัยทางระบาดวิทยาชั้นนำของโลกในปัจจุบันส่วนหนึ่งพัฒนามาจากนักคณิตศาสตร์ในสาขาสถิติ การพัฒนาวิธีวิทยาวิจัยทางระบาดวิทยาส่วนหนึ่งเป็นการพัฒนาออกแบบและการบริหารจัดการการได้มาซึ่งข้อมูลขนาดใหญ่ อีกส่วนหนึ่งเป็นการพัฒนาวิธีการวิเคราะห์ข้อมูล

ประวัติพัฒนาการโดยย่อ

วิชาระบาดวิทยาก่อกำเนิดขึ้นหลังจากปรัชญา Inductivism หรือ การได้มาซึ่งความรู้จากการสังเกตหรือเก็บข้อมูลจากธรรมชาติมากกว่าการถกเถียงโดยอิงทฤษฎี ความรู้ที่ได้จาก induction นั้นต้องมาจากการสังเกตจำนวนมาก ๆ โดยเฉพาะเมื่อธรรมชาตินั้น ๆ มีความไม่แน่นอนมาก และสิ่งที่ถูกสังเกตหรือเก็บข้อมูลต้องเป็นตัวแทนที่ดีของสิ่งที่จะนำข้อสรุปนั้นไปประยุกต์ ระบาดวิทยาซึ่งกำเนิดขึ้นในวงการแพทย์จึงจะต้องอาศัยข้อมูลขนาดใหญ่ในประชากรและให้ความสำคัญกับการเป็นตัวแทนประชากรของข้อมูล

องค์ความรู้ทางการสาธารณสุขพัฒนาทางระบาดวิทยาเริ่มจากการศึกษาสถิติชีพในศตวรรษที่ 17 และเริ่มการทดลองรักษาโรคลักปิดลักเปิด (scurvy) โดยมีกลุ่มควบคุมในกลางศตวรรษที่ 18 ต่อมากลางศตวรรษที่ 19 ระหว่าง 1848-1854 Sir John Snow ซึ่งเป็นวิสัญญีแพทย์ในสำนักพระราชวังอังกฤษออกไปสอบสวนหาสาเหตุของอหิวาตกโรคซึ่งระบาดในกรุงลอนดอนโดยใช้วิธีการแจงนับและสร้างตารางจรง่าย ๆ จากการสำรวจครัวเรือนผู้ดื่มน้ำจากบริษัทขายน้ำสองแห่งเป็นประจำ พบว่าครัวเรือนที่ดื่มน้ำจากบริษัทขายน้ำแห่งหนึ่งมีอัตราตายจากอหิวาตกโรคประมาณ 8-9 เท่าของครัวเรือนที่ดื่มน้ำจากอีกบริษัทหนึ่ง นอกจากนี้เขายังได้ระงับการระบาดของอหิวาตกโรคในถนนสายหนึ่งโดยถอดเอาสูบโยกในถนนสายนั้นออกไป ข้อสรุปของเขาคือ อหิวาตกโรคระบาดโดยการดื่มน้ำที่ปนเปื้อน ต่อมาอีกประมาณ 20 ปีการศึกษาโดยวิธีการแจงนับแบบเดียวกันได้ข้อสรุปว่าไข้ไทฟอยด์ในเมือง North Towton ตอนเหนือของประเทศอังกฤษก็เกิดจากการติดต่อทางการสัมผัสกับสิ่งขับถ่ายของผู้ป่วย การศึกษาทั้งสองรายการนี้ได้ข้อสรุปของวิธีการระบาดของโรคก่อนการค้นพบเชื้อ *Vibrio cholera* หรือเชื้ออหิวาต์ และเชื้อ *Salmonella typhi* หลายสิบปี ผลการศึกษาทั้งสองได้ข้อสรุปว่าการเก็บข้อมูลระดับประชากรสามารถนำไปสู่ข้อสรุปสาเหตุของโรคได้ โดยคำว่าสาเหตุ (cause) ในทางระบาดวิทยาไม่จำเป็นจะต้องจับตัวตนของ agent ได้ แต่อาจจะจับพฤติกรรมหรือสิ่งแวดล้อมที่นำไปสู่การเกิดโรค นักระบาดวิทยาไม่จำเป็นต้องรอให้นักวิจัยในห้องทดลองจับ agent ให้ได้ก็สามารถเริ่มงานของตนได้ แต่เป็นที่แน่นอนว่าความเข้าใจธรรมชาติของ agent ในห้องปฏิบัติการจะช่วยให้งานวิจัยในภาคสนามดำเนินไปถูกทิศทางยิ่งขึ้น

หลังการวิจัยที่สำคัญทั้งสองไม่นาน วิชาจุลชีววิทยาซึ่งใช้การทดลองในหลอดทดลองและสัตว์ทดลองก็เจริญขึ้นอย่างรวดเร็ว งานวิจัยทางจุลชีววิทยามีความก้าวหน้ากว่าการวิจัยแจงนับภาคสนาม วิชาสถิติซึ่งถือกำเนิดส่วนหนึ่งจากสถิติชีพและระบาดวิทยาก็หันไปประยุกต์แก้ปัญหาในห้องทดลองและโจทย์ทางพันธุกรรม ทำให้วิธีการวิจัยทางระบาดวิทยาก็ซบเซาไปหลายทศวรรษ จนกระทั่งในต้นศตวรรษที่ 20 มีการวิจัยภาคสนามในสหรัฐอเมริกาค้นพบว่าการขาดวิตามิน niacin เป็นสาเหตุของ pellagra หรือโรคชนิดหนึ่งซึ่งมีอาการทางผิวหนัง ระบบประสาทและระบบย่อย

อาหาร และอีกการศึกษาระยะยาวหลายทศวรรษต่อมาจึงค้นพบว่าการสูบบุหรี่เป็นสาเหตุของโรคมะเร็ง กลุ่มนักระบาดวิทยาในศตวรรษที่ 20 นี้ได้พลิกฟื้นความสำคัญของวิชาระบาดวิทยาอีกครั้งหนึ่ง จากนั้นนักระบาดวิทยาได้ศึกษาหาสาเหตุของโรคต่าง ๆ มากมาย ที่สำคัญมากคือสาเหตุของโรคหัวใจขาดเลือดและพบว่ามีปัจจัยต่าง ๆ มากมายเข้ามาเกี่ยวข้อง ทั้งด้านโภชนาการ การออกกำลังกาย การสูบบุหรี่ และบุคลิกภาพ ในขณะที่การวิจัยทางชีววิทยายังคงดำเนินไปเพื่อหากลไกหรือสารต่าง ๆ ในร่างกายหรือในบุหรี่ที่ทำให้เกิดโรค ความรู้ทางระบาดวิทยาสามารถนำมาประยุกต์ใช้ป้องกันปัญหาเหล่านั้นได้เลยโดยไม่ต้องรอคำตอบทางชีววิทยา

ในครึ่งหลังของศตวรรษที่ 20 นี้ ความสำเร็จที่สำคัญในทางวิธีวิทยาการวิจัยไม่ใช่อยู่ที่การพิชิตโรคใดโรคหนึ่ง แต่อยู่ที่การพัฒนาความสามารถในการพิสูจน์ปัจจัย (exposure) ที่ส่งผลทำให้เกิดโรคในระยะยาว ซึ่งปรกติจะหาความสัมพันธ์เชิงเหตุเชิงผลได้ยาก (เพราะต้องใช้เวลานานกว่าจะเห็นผลและมีระดับของความสัมพันธ์ระหว่างเหตุกับผลไม่สูงมากอย่างโรคเฉียบพลัน) เช่น การเกิดโรคมะเร็งเกิดหลังจากสูบบุหรี่เป็นประจำหลายสิบปี การเกิดโรคหัวใจขาดเลือดเกิดหลังจากมีวิถีชีวิตที่ไม่ถูกต้องหลายสิบปีเช่นกัน ซึ่งจำเป็นต้องออกแบบงานวิจัยให้ดี ต้องจัดการติดตามบุคคลจำนวนมากในระยะยาว

ความสำเร็จที่สำคัญอีกประการหนึ่งในเชิงวิธีวิทยาการวิจัย คือ ความสามารถในการหาสาเหตุหลาย ๆ สาเหตุพร้อมกันในตัวอย่างของโรคหัวใจขาดเลือดเป็นต้น ความสำเร็จในแนวสุดท้ายนี้เกิดจากการนำวิธีการทางสถิติเข้าร่วมวงการระบาดวิทยา การพัฒนาที่สำคัญในช่วงนั้นคือการออกแบบงานวิจัยซึ่งเดิมมีเพียงแบบเดียว คือ **การสำรวจแบบภาคตัดขวาง** (cross-sectional study) เพิ่มเติมอีก 3 แบบหลัก คือ ประการแรก case-control study ซึ่งสืบสวนผู้ป่วยและกลุ่ม control ย้อนหลัง ซึ่งเหมาะสำหรับการสืบสวนโรคที่พบได้น้อย, ประการที่สอง cohort study ซึ่งติดตามบุคคลแต่ละคนระยะยาวเพื่อทดสอบความสัมพันธ์ระหว่างพฤติกรรมหรือสภาวะ ณ จุดเริ่มติดตามกับการเกิดโรค ณ จุดสุดท้ายของการติดตาม และ ประการที่สาม Randomized Controlled Trial ซึ่งเป็นการทดลองรักษาหรือป้องกันโรคโดยจัดให้ผู้ถูกทดลองอยู่ในกลุ่มใดกลุ่มหนึ่งโดยการสุ่ม (random allocation) เพื่อลด bias และ confounder

ในช่วงปลายศตวรรษที่ 20 วิชาระบาดวิทยาพัฒนาอย่างรวดเร็วที่สุดโดยประสานการพัฒนาการวิจัยเชิงสถิติ วิทยาการคอมพิวเตอร์ และเทคโนโลยีชีวภาพ ระบาดวิทยาในปัจจุบันจึงเป็นศาสตร์ที่มีความเข้มแข็งมากที่สุดศาสตร์หนึ่งและเป็นเครื่องมือที่ใช้ในการวิจัยทางคลินิกและทางสาธารณสุขที่ขาดไม่ได้ การตัดสินใจทางการแพทย์ซึ่งเดิมเป็น "การประกอบโรคศิลป" ซึ่งอาศัยความเชื่อหรือญาณ (intuition) พัฒนาไปเป็น "Evidence-Based Medicine" หรือการแพทย์ที่อาศัยหลักฐานจากการวิจัยที่ผ่านมาเป็นเกณฑ์ตัดสินความถูกต้องทางวิชาการ นั่นคือวิธี

วิทยาการวิจัยด้านระบาดวิทยามีบทบาทด้านนี้สูงมาก ข้อสรุปที่น่าเชื่อถือในทางการแพทย์ปัจจุบันส่วนใหญ่จะต้องได้มาจากการศึกษาทดลองที่ดีและมีขนาดใหญ่ และในปลายศตวรรษที่ 20 ได้เกิดวิธีวิทยาการวิจัยที่เรียกว่า Systematic Review หรือการทบทวนผลการทดลองทางคลินิกเรื่องใดเรื่องหนึ่งอย่างเป็นระบบ ทำให้สามารถรวบรวมวิพากษ์การทดลองที่เหมือน ๆ กันในโลกได้อย่างกว้างขวาง คัดเลือกเหลือเฉพาะการวิจัยที่เหมาะสมแล้วใช้วิธีสถิติที่เรียกว่า Meta-Analysis หาข้อสรุปผลการทดลองรวมจากการทดลองทั้งหมดในอดีตจนถึงปัจจุบันให้แพทย์นำไปตัดสินใจ

การจำแนกแยกแยะข้อผิดพลาด

ณ กึ่งศตวรรษที่ 20 นักระบาดวิทยาสรุปว่า ในการหาสาเหตุของปัญหานั้น ความผิดพลาด (error) แบ่งเป็นสามข้อ คือ *ข้อแรก*อาจจะเกิดจากความลำเอียง (bias), *ข้อที่สอง* อาจจะเกิดจากการปนกันยุ่งเหยิง (confound) ไม่รู้ว่าปัจจัยไหนเป็นสาเหตุจริงหรือปัจจัยไหนเป็นตัวแปรที่พลอยฟ้าพลอยฝนทำให้คนเข้าใจผิดว่าอาจจะเป็นสาเหตุ, และ*ข้อสาม*อาจจะเกิดจากความบังเอิญ (random error)

Bias และ random error ในทางระบาดวิทยานั้นเข้าใจได้ไม่ยาก เพราะเหมือนกับการวิจัยทางสังคมศาสตร์ส่วนใหญ่ นั่นคือ bias คือ systematic error อาจจะเกิดจากการเลือกผู้ถูกศึกษาอย่างไม่เหมาะสม หรือเกิดจากวิธีวัด หรือวิธีวิเคราะห์ผิดพลาด ส่วน random error เกิดจากความผิดพลาดเนื่องจาก chance หรือ ยถากรรม ซึ่งเป็น random error ซึ่งแบ่งเป็น Type I และ Type II error, โดยที่ Type I error ซึ่งอ่านได้จากตำราสถิติทั่วไป

Confounding ปัญหาสำคัญในวิธีวิทยาการวิจัย

สาเหตุของข้อผิดพลาดที่เรียกว่า confounding นั้นเป็นสิ่งที่นักระบาดวิทยากังวลมากที่สุด เช่น การประเมินเปรียบเทียบผลการรักษาระหว่างวิธีการใช้ยากับวิธีการผ่าตัด หมอผ่าตัดมักเลือกคนที่แข็งแรงไปผ่าและปล่อยให้คนที่อาการหนักรักษาด้วยยา ข้อสรุปที่ว่าการรักษาทางการผ่าตัดได้ผลดีกว่าการรักษาด้วยยาที่มาจากข้อมูลประเภทนี้มีปัญหาของ confounding มาก กล่าวคือ สภาพร่างกายก่อนผ่าตัดเป็น confounder สภาพร่างกายที่แย่ทำให้ภาวะแทรกซ้อนและอัตราตายจากการรักษาสูง การรักษาที่ให้กับคนที่สภาพร่างกายดีมักจะมีผลการรักษาที่ดีจึงถูกเหมาเอาอย่างผิด ๆ ว่าเป็นการรักษาที่ดี ส่วนการรักษาที่ให้กับผู้ป่วยที่ร่างกายอ่อนแอมักลงท้ายด้วยผลการรักษาที่ไม่สู้จะดีจึงถูกเหมาเอาอย่างผิด ๆ ว่าเป็นการรักษาที่ไม่ดี

การแก้ปัญหา confounding ในระดับของการออกแบบงานวิจัย

วิธี Randomized Controlled Trial

การแก้ confounding ที่สำคัญอยู่ที่การออกแบบการวิจัย ถ้าเป็นการทดลองต้องใช้ randomized allocation คืออย่าให้ผู้ป่วยหรือผู้รักษาเลือกวิธีการรักษาเอง ต้องใช้วิธีสุ่มหรือจับสลากว่าใครจะได้การรักษาแบบใด โดยวิธีนี้ผู้ที่ได้รับการรักษาทั้งสองวิธีจะมีการแจกแจงของคุณสมบัติต่าง ๆ เช่น อายุ เพศ รวมทั้งสภาพความพร้อมของร่างกายไม่แตกต่างกัน วิธีการรักษาไม่ถูก confound หรือพลอยฟ้าพลอยฝนไปกับปัจจัยใด ๆ ทั้งสิ้น การออกแบบการวิจัยโดยวิธีนี้จึงเป็นวิธีที่ดีที่สุดในการแก้ปัญหา (ถ้าจะให้ดีขึ้นไปกว่านี้ต้อง double-blind คือทำให้ทั้งผู้รับผิดชอบการรักษาและผู้ป่วยไม่ทราบว่าการรักษาที่ให้อยู่นั้นใช้วิธีใดโดยการใช้ยาปลอมหรือ placebo แต่วิธีนี้ใช้ได้กับการรักษาด้วยยาเท่านั้น)

การปรับปรุงวิธีการวิจัยแบบสังเกต

สำหรับการวิจัยที่ไม่ใช่การทดลอง เช่น cross-sectional study, case-control study และ cohort study ที่กล่าวมาแล้วในตอนต้น exposure เกิดขึ้นตามธรรมชาติ ผู้วิจัยไม่สามารถ allocate ให้กับผู้ถูกทดลองได้ด้วยปัญหาทางจริยธรรมและปัญหาการจัดการ จะต้องมีวิธีการแก้ confounding โดยวิธีต่าง ๆ ดังนี้

วิธีแรกได้แก่การคัดเลือกศึกษาเฉพาะบางกลุ่ม (ใช้ selection criteria) เช่น ถ้าอายุเป็นปัจจัยเสี่ยงสำคัญของการหายจากโรคหรือของการเกิดโรค ก็อาจจะจำกัดเลือกกลุ่มศึกษาให้มีอายุอยู่ในกลุ่มเดียวเท่านั้น

วิธีที่สองคือการจัดเป็นชั้น (strata) แล้วเปรียบเทียบกลุ่มทั้งสองให้อยู่ในชั้น (stratum) เดียวกันเช่นศึกษาว่าการกินยาคุมกำเนิดเพิ่มความเสี่ยงต่อมะเร็งมดลูกหรือไม่ การศึกษาเช่นนี้จะมีอายุเป็น confounder เมื่อบุคคลอายุมากขึ้นความเสี่ยงต่อโรคมะเร็งมากขึ้นและหญิงที่มีอายุมากย่อมมีความน่าจะเป็นที่เคยได้กินยาคุมต่างกับหญิงที่อายุน้อย การเปรียบเทียบประวัติการกินยาคุมกำเนิดระหว่างผู้ป่วยที่เป็นโรคมะเร็งมดลูกกับผู้หญิงปรกติ ต้องแยกเปรียบเทียบภายในกลุ่มอายุเดียวกันเพื่อขจัดปัญหา confounding effect ของอายุ ตัวแปรอายุกลายถูกนำมาจัดเป็นชั้น (stratification factor) ผลของอายุก็จะไม่มาปะปน (confound) กับตัวแปรที่ต้องการทดสอบ

วิธีที่สาม เพื่อจะให้แน่ใจยิ่งขึ้นว่าตัวแปรอื่น ๆ ที่ไม่ทราบหลายตัวจะถูกกันออกไปไม่ให้มารบกวนการแปลผล การออกแบบอีกระดับหนึ่งคือการจับคู่ (matching) ให้คู่ที่กำลังเปรียบเทียบมีความคล้ายคลึงกันมากขึ้น เช่น อายุเท่ากัน เพศเดียวกัน อยู่ในฐานะทางสังคมเดียวกัน หรืออาจ

จะเปรียบเทียบภายในครอบครัวเพื่อให้ปัจจัยทางพันธุกรรมและสิ่งแวดล้อมถูกกันออกไป อย่างไรก็ตาม matching อาจจะทำให้เกิด over-matching กล่าวคือตัวแปรที่ต้องทดสอบสมมติฐานก็จะถูก matched out ออกไปด้วยโดยไม่ได้ตั้งใจ เช่น ถ้าพฤติกรรมบางอย่างเป็นสิ่งที่สงสัยว่าจะเป็นสาเหตุของโรค ถ้าเลือกจับคู่ภายในครอบครัวเดียวกัน คู่ที่จับก็มักจะมีพฤติกรรมคล้ายกันมากกว่าการจับคู่อย่าง random เพราะมาจากครอบครัวเดียวกัน

การออกแบบงานวิจัยช่วยลด confounding ได้ในระดับหนึ่ง อย่างไรก็ตาม ในการวิจัยส่วนใหญ่มักจะมีตัวแปรค่อนข้างมาก ไม่อาจจะนำมาเป็น stratification factor และ matching factor ได้หมด ในที่สุดก็ต้องลงท้ายด้วยการแก้ไข confounding ในขั้นตอนของการวิเคราะห์โดยใช้ multi-variate model ซึ่งจะได้กล่าวต่อไป

พัฒนาการทางเทคโนโลยีในการวิเคราะห์ข้อมูลในช่วงสองทศวรรษสุดท้าย

การพัฒนาที่สำคัญที่สุดในช่วงสองทศวรรษสุดท้าย คือการพัฒนาทางเทคโนโลยีการวิเคราะห์ข้อมูล ในที่นี้จะกล่าวถึงด้านที่สำคัญสองด้านก่อน คือ เรื่องระบบฐานข้อมูลและ high resolution graphics แล้วจึงต่อด้วยการวิเคราะห์ multi-variate ซึ่งต่อเนื่องกับข้างบน

ระบบฐานข้อมูลคอมพิวเตอร์

วิธีการแจงนับ (tally) ซึ่งใช้มาหลายร้อยปีเป็นวิธีหลักในการวิจัยเช่นเดียวกันกับการวิจัยระดับประชากรสาขาอื่น ๆ มาจนถึงปัจจุบัน การแจงนับในภายหลังเปลี่ยนจากการใช้สมุดจดบันทึกไปเป็นการเจาะบัตรข้อมูลคอมพิวเตอร์และต่อมาพัฒนาไปเป็นป้อนด้วย keyboard บันทึกลงในแผ่นดิสก์คอมพิวเตอร์ จนถึงปัจจุบันซึ่งสามารถใช้เทคโนโลยี Optical Character Reading (OCR) อ่านข้อมูลจากกระดาษโดยตรงและแปลงเป็นตัวเลขโดยไม่ต้องใช้คนป้อน เทคโนโลยีเหล่านี้ทำให้สามารถเก็บฐานข้อมูลที่มีระเบียบสูง เรียกใช้ (retrieve) และแจงนับข้อมูลมหาศาลในเวลาอันสั้น นอกจากนี้กระบวนการด้านการจัดการฐานข้อมูลได้พัฒนาไปเป็น relational database ซึ่งมีข้อมูลหลายระดับเชื่อมโยงถึงกัน เช่น ด้านผู้รับบริการข้อมูลจำเพาะของท้องถิ่นเชื่อมกับข้อมูลของแต่ละครัวเรือนในท้องถิ่นนั้นและข้อมูลของครัวเรือนเชื่อมกับข้อมูลของสมาชิกภายในครัวเรือน ในขณะเดียวกันทางด้านการให้บริการก็มีข้อมูลจำเพาะระดับโรงพยาบาลซึ่งเชื่อมกับข้อมูลจำเพาะของแพทย์แต่ละคน และในที่สุดก็เชื่อมกันระหว่างแพทย์แต่ละคนที่ดูแลผู้ป่วยแต่ละคนในการตรวจรักษาแต่ละครั้ง โดยที่แพทย์คนเดียวกันอาจจะรักษาผู้ป่วยหลายคน และผู้ป่วยแต่ละคนได้รับการดูแลจากแพทย์ต่าง ๆ ในเวลาต่าง ๆ จะเป็นได้ว่าระบบข้อมูล relational database สลับซับซ้อนกว่าการแจงนับธรรมดาที่มีเพียง file เดียวซึ่งมี record แทนตัวบุคคลในแนวราบ และ fields แทน

variables ในแนวนอนเท่านั้น การเก็บข้อมูลต่าง ๆ ไว้ในระบบคอมพิวเตอร์ที่สามารถเชื่อมโยงได้ถึงกันนี้เป็นรากฐานของการศึกษาระบาดวิทยาในประเทศที่เจริญแล้วในรอบ 20-30 ปีที่ผ่านมา

High Resolution Graphics

ระบบ High-Resolution Graphics ซึ่งเพิ่งนำเข้ามาสู่วงการคอมพิวเตอร์เมื่อไม่เกิน 10 ปีนี้ ทำให้นักวิเคราะห์สามารถเขียนกราฟที่ลงรายละเอียดต่างๆ ได้มาก เมื่อประมาณ 15 ปีที่แล้ว กราฟที่เขียนจากคอมพิวเตอร์ค่อนข้างหยาบและส่วนใหญ่ตอบสนองการนำเสนอง่าย ๆ เช่น pie chart, histogram ปัจจุบันโปรแกรมคอมพิวเตอร์สามารถสร้าง dot plots, box plots ซึ่งแสดง out-lier ได้ชัดเจน ตลอดจนช่วยในด้าน scatterplots, multi-variate plots, contour plots, perspective plots เพื่อให้เห็นอิทธิพลของตัวแปรต่าง ๆ หลายตัวพร้อมกันในกราฟเดียวกัน

Multi-Variate Analysis

ดังกล่าวข้างต้นแล้วว่า การแก้ปัญหา confounding ด้วย multi-variate analysis เป็นวิธีการที่สำคัญที่สุด ในการนี้ต้องใช้เทคโนโลยีด้านคอมพิวเตอร์

คำว่า model หมายถึงการจัดรูปแบบทางคณิตศาสตร์ ในบทความนี้ทั้งหมดเป็น linear model ซึ่งก็คือสมการที่พยากรณ์ตัวแปรตามตัวเดียวจากตัวแปรอิสระหลาย ๆ ตัว ในที่นี้จะนำเสนอ model ที่สำคัญที่สุดในทางระบาดวิทยาเพียงบาง model

การจัดรูปแบบ (Modeling) กับการแก้ปัญหา Confounding

Gaussian Regression

การคิดค้นทางคณิตศาสตร์ของ Carl Federick Gauss ที่พบว่า random error ส่วนใหญ่ในธรรมชาติมีการแจกแจงแบบ normal distribution (ซึ่งเรียกอีกอย่างหนึ่งว่า Gaussian distribution) และการหาค่าความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามด้วยการแก้สมการเชิงพีชคณิตให้ได้ค่า error ต่ำสุด วิธี Least Square Methods เป็นการค้นพบกฎธรรมชาติที่ยิ่งใหญ่อย่างหนึ่ง ถ้ามีตัวแปรอิสระเพียงตัวเดียวก็จะได้เส้นตรงจากสมการถดถอยซึ่งเรียกว่า Linear Regression คำว่า linear หมายถึงว่าเป็นเส้นตรง ส่วนคำว่า regression หมายถึงว่าเกิดจากการพยายามถอยจุดต่าง ๆ ที่กระจัดกระจายให้กลับไปอยู่ในเส้นนี้ ต่อมาเมื่อมีการพัฒนาสมการนี้เป็นรูปแบบจำลอง (model) ชนิดต่าง ๆ ก็พลอยทำให้รูปแบบจำลองที่ตามมาเหล่านั้นมีลักษณะบางประการคล้ายกับ linear regression จึงเรียกรูปแบบจำลองเหล่านั้นว่า Linear models หรือรูปแบบจำลองเชิงเส้น

รูปแบบจำลองที่พัฒนาไปจาก linear regression เป็นอันดับแรกคือ Multiple Regression ซึ่งมีตัวแปรอิสระหลายตัวโดยที่ตัวแปรตามยังเป็นค่าต่อเนื่องอยู่ ทั้ง Linear Regression และ Multiple Regression ต่างอาศัยทฤษฎีการแจกแจง random errors ของ Gauss จึงเรียกรวมได้ว่า Gaussian Regression

ใน Multiple Regression ภายใต้สมการเดียวตัวแปรแต่ละตัวจะส่งผลอิสระจากกัน การนำตัวแปรที่เป็น confounder มาไว้ใน model เดียวกันกับตัวแปรหลักจึงเป็นการทำให้อิทธิพลอิสระของตัวแปรหลักปรากฏขึ้น นั่นก็คือตัวแปรหลักไม่ถูก confound อีกต่อไป หลักเกณฑ์ใช้กับ regression อย่างอื่น ๆ ด้วย ซึ่งเรียกรวม ๆ ว่า Multi-Variate Analysis

Multi-variate analysis สำหรับแพทย์และนักสาธารณสุขก่อนปี 1980 จำกัดอยู่ภายใต้ Gaussian distribution เท่านั้น วิธีนี้ใช้อธิบายการทดลองในห้องปฏิบัติการวัดค่าต่าง ๆ ในทางชีววิทยาได้ แต่ไม่สามารถประยุกต์กับปัญหาตัวแปรทวินาม (คือตัวแปรที่มีค่าเพียงสองค่า) เช่น การเจ็บป่วย (ป่วย/ไม่ป่วย) และความตาย(ตาย/ไม่ตาย) ในช่วงก่อน 1980 นักระบาดวิทยาจึงไม่ค่อยได้ใช้ multi-variate analysis

Logisitic Regression

ย้อนกลับไปประมาณทศวรรษที่ 1970 นักสถิติได้แสดงให้เห็นเชิงทฤษฎีว่า แบบจำลองความสัมพันธ์แบบ Gaussian ซึ่งมีตัวแปรตามเป็นค่าต่อเนื่องซึ่งใช้กันมากในทางวิทยาศาสตร์ธรรมชาติสาขาต่าง ๆ นั้นสามารถดัดแปลงใช้กับปรากฏการณ์ทางระบาดวิทยาซึ่งตัวแปรตามเป็น binary data (0=ปรกติ, 1=ป่วย) ได้

จากข้อมูลที่มีตัวแปรตามเป็น binary นี้ Model ที่ได้จะมีพยากรณ์ logit ซึ่งเป็นดัชนีวัดความเสี่ยงในทางทฤษฎี โดยที่ logit ซึ่งมีค่าได้ตั้งแต่ $-\infty$ ถึง ∞ เช่นเดียวกันกับตัวแปรตามที่มีค่าต่อเนื่อง

แบบจำลองการแก้สมการเพื่อประเมินค่าความเสี่ยงแบบ logit นี้ เรียกว่า logit regression หรือ logistic regression การแก้สมการ logistic regression จะใช้วิธีพีชคณิตไม่ได้ ต้องใช้วิธีการทำซ้ำ (iteration) ดังกล่าวข้างต้นเพื่อให้ได้ค่าพยากรณ์ที่สอดคล้องกับข้อมูลเชิงประจักษ์มากที่สุด (maximum likelihood)

คำว่า Likelihood หมายถึงความน่าจะเป็นที่จะพบข้อมูลอย่างที่เราพบ ข้อมูลแต่ละ record หรือของแต่ละคนมีตัวแปรอิสระต่าง ๆ หลายตัว ถ้าแทนค่าสัมประสิทธิ์เข้าไปก็จะพยากรณ์โดยคำนวณว่า likelihood ที่บุคคลคนนั้นจะมีสภาพเช่นนั้นออกมาเป็นตัวเลขสำหรับแต่ละ record เมื่อพิจารณาข้อมูลทั้งหมดที่มีอยู่ likelihood รวมมีค่าเท่ากับผลคูณความน่าจะเป็นของ record

ทั้งหมด ค่าสัมประสิทธิ์ที่ดีที่สุดคือค่าที่เมื่อแทนค่าลงไปครบทุก record แล้วจะทำให้ likelihood รวมมีค่าสูงสุด

กระบวนการหาสัมประสิทธิ์นี้ต้องทดลองให้โปรแกรมคอมพิวเตอร์แทนค่าสัมประสิทธิ์ทุกตัวที่ไม่ทราบค่าลงในทุก record แล้วคำนวณค่า likelihood ออกมา จากนั้นค่อย ๆ ขยับแทนค่าของสัมประสิทธิ์ใหม่เพื่อหาทางให้ได้ค่า likelihood ดีขึ้น แล้วทำซ้ำ (iterate) จนกระทั่ง likelihood ได้ค่าสูงสุดแล้ว ค่าของสัมประสิทธิ์ ณ จุดที่ได้ค่า maximum likelihood นั้นถือว่าจะเป็นสัมประสิทธิ์ที่เหมาะสมที่สุด ณ จุดนั้นจะได้คำตอบของสัมประสิทธิ์ต่าง ๆ ที่ต้องการทั้งหมด วิธีการเหล่านี้จะทำไม่ได้เลยถ้าไม่ใช้คอมพิวเตอร์เนื่องจากต้องอาศัยการคำนวณมากมาย

ตั้งแต่กลางทศวรรษ 1980 เป็นต้นมา PC - Personal Computer หรือคอมพิวเตอร์ส่วนบุคคลได้รับการพัฒนาอย่างรวดเร็ว โปรแกรมที่มี logistic regression ซึ่งเดิมอยู่บนคอมพิวเตอร์ระดับ mainframe เท่านั้น ถูกดัดแปลงลงมาใช้ใน PC และมีโปรแกรมทางสถิติเกิดขึ้นใหม่จำนวนมาก Logistic regression จึงกลายเป็นการวิเคราะห์มาตรฐานสำหรับข้อมูลทางระบาดวิทยา

นอกจากใช้ใน logistic regression แล้ว การหา maximum likelihood สำหรับ model อื่นๆ ก็ใช้หลักการคล้ายกัน วิธีการนี้จึงเป็นคุณูปการใหญ่หลวงสำหรับการแก้สมการทางระบาดวิทยาซึ่งมีตัวแปรอิสระหลายตัวพร้อมกันในรอบ 20 ปีที่ผ่านมา

ในปัจจุบันนอกจากใช้ logistic regression ในการพยากรณ์ outcome ที่เป็นทวินามเช่น ป่วยหรือไม่ป่วยแล้ว ยังมี polytomous (หรือ multinomial) logistic regression ซึ่งพยากรณ์ outcome ที่เป็นพหุนาม (polytomous) เช่น ผลการรักษาที่หายขาด, พิการ, เป็นซ้ำซาก หรือตาย หรือพยากรณ์ผลการตั้งครรภ์ว่าจะแท้ง, คลอดก่อนกำหนด, คลอดปรกติ หรือ ตายคลอด ฯลฯ นอกจากนี้ยังมี ordinal logistic regression ที่พยากรณ์ outcome ที่เป็นลำดับหรือ order เช่น ไม่ป่วย, ป่วยน้อยไม่ต้องนอนโรงพยาบาล, ป่วยหนักต้องนอนโรงพยาบาล, หรือป่วยอย่างรุนแรงจนตาย ฯลฯ

Survival Analysis

นอกเหนือจากผลลัพธ์สุดท้ายคือการป่วยหรือตายแล้ว ช่วงเวลาของการคงอยู่ (survival) โดยปราศจากสภาวะที่ไม่พึงปรารถนาก็เป็นสิ่งสำคัญ มนุษย์เราล้วนแต่ต้องป่วยและต้องตายทั้งสิ้น ในบางครั้งการป้องกันก็ดี การรักษาก็ดี เป็นการยืดเวลาแห่งการคงอยู่โดยปราศจากสภาวะที่ไม่พึงประสงค์ทั้งสิ้น ถ้าพยากรณ์หาปัจจัยเร่งบุคคลให้ไปสู่สภาวะที่ไม่พึงประสงค์ได้ ก็อาจจะนำไปสู่ความรู้สำหรับการป้องกันปัญหาที่มีประสิทธิผลดี

Logistic regression ไม่เหมาะสำหรับการวิเคราะห์ข้อมูลที่มีเวลามาเกี่ยวข้อง เพราะ logistic regression ไม่สนใจเรื่องเวลามากนัก ถ้าจะวิเคราะห์ต้องกำหนดขีดเส้นเวลาหลาย ๆ ค่าว่า outcome ณ เวลานั้น ๆ เป็นอย่างไร เนื่องจากเวลาเป็น continuum ถ้าจะทำเช่นนั้นต้องสร้าง model นับไม่ถ้วน

วิธีการวิเคราะห์ที่ถูกต้องคือ survival analysis ซึ่งมีวิธีต่าง ๆ แตกย่อยไปหลายวิธี Survival analysis นั้นเริ่มต้นมาตั้งแต่ประมาณ 300 ปีที่แล้วโดย Edmund Halley ผู้ค้นพบดาวหางฮัลเล่ย์ โดยวิธีที่เรียกว่าตารางชีพ หรือ Life-Table Analysis วิธีนี้คำนวณอัตราการอยู่รอดของแต่ละช่วงอายุแล้วคูณสะสมเป็นอัตราการคงอยู่สะสม (cumulative survival) ตั้งแต่แรกเกิด และคำนวณอายุขัย (life expectancy) เฉลี่ยของประชากร เดิมนักระบาดวิทยาใช้ cumulative survival ในการเปรียบเทียบระหว่างกลุ่ม exposure ต่าง ๆ โดยเปรียบเทียบให้การเริ่ม expose เหมือนการเกิด และการเกิดโรคเหมือนการตายใน life-table ต่อมา Kaplan และ Meier เสนอว่าเพื่อให้ละเอียดขึ้น แทนที่จะคำนวณ cumulative survival เป็นกลุ่มอายุตามวิธีการทางประชากรศาสตร์ก็คำนวณ cumulative survival จนถึงวันเริ่มป่วยของแต่ละคนเขียนเป็น survival curve และเปรียบเทียบการคงอยู่จาก curve นี้ นอกจากนั้นยังใช้วิธี Log-Rank Test เปรียบเทียบจำนวนตายหรือป่วยที่สะสมทั้งหมดกับจำนวนที่น่าจะเป็นถ้ากลุ่มย่อยต่าง ๆ มีความเสี่ยงเท่ากันจริง (observed vs. expected number of failure or death)

ในการแก้ไขปัญหาของ confounder ก็ต้องจัดให้ confounder เป็น stratification factor แต่ถ้าต้องการควบคุม confounder หลายตัวพร้อมกันสำหรับ survival analysis ต้องใช้ Cox regression

Cox Regression และ Proportional Hazard Model

ผลงานเชิงทฤษฎีของ David Cox ซึ่งตีพิมพ์ในปี 1972 เมื่อพัฒนาไปสู่ software สถิติต่าง ๆ แล้วปรากฏว่าเป็นที่นิยมของวงการแพทย์มาก จนถึงปัจจุบันงานตีพิมพ์ชิ้นนี้เป็นผลงานที่มีการอ้างอิง (citation) ในวงการวิทยาศาสตร์มากที่สุดในโลก

Cox Regression เริ่มต้นจากทฤษฎีที่มีข้อสมมติว่า ไม่ว่าความเสี่ยงต่อการเกิดโรคหรือความตายของบุคคลจะเปลี่ยนแปลงไปตามกาลเวลาอย่างไร การวิเคราะห์จะถือว่าอิทธิพลของตัวแปรอิสระที่มีต่อความเสี่ยงจะคงที่ไม่เปลี่ยนตามกาลเวลา ที่เรียกว่า Proportional Hazard Model เช่น ผู้ป่วยโรคเบาหวานจำนวนหนึ่งจะเกิดอาการแทรกซ้อน ความเสี่ยงนี้อาจจะเพิ่มขึ้นตามกาลเวลาหรืออย่างไรไม่แน่ชัด ผู้หญิงหรือผู้ชายอาจจะเสี่ยงต่างกันหรือไม่ก็ไม่แน่ชัด ถ้าต้องการทดสอบสมมติฐานว่าการออกกำลังกายช่วยได้หรือไม่โดยใช้ Cox regression ผลประโยชน์

จากการออกกำลังจะคงที่สม่ำเสมอไม่ว่าตอนเริ่มเป็นเบาหวานหรือเป็นมานานแล้ว อัตราที่ลดลงได้(ซึ่งไม่เปลี่ยนตามเวลา)นี้เรียกว่า hazard ratio ซึ่งในเบื้องต้นสมมติว่าคงที่ตลอดเวลา

ข้อสมมติ proportional hazard ไม่เป็นจริงในหลายกรณีเช่นการรักษาโรคไตวายเรื้อรังโดยวิธีผ่าตัดปลูกไตเทียบกับการรักษาแบบล้างไต เมื่อเริ่มผ่าตัดนั้น ผู้ที่ถูกผ่าตัดมีความเสี่ยงสูงกว่าเนื่องจากอาจจะได้รับผลแทรกซ้อนจากการผ่าตัด แต่เมื่อระยะเวลาผ่านไป ถ้าไม่มีปัญหาอะไรผู้ที่ได้รับการผ่าตัดจะเสี่ยงน้อยกว่าผู้ที่ต้องล้างไตเป็นประจำ นั่นก็คือ hazard ratio เปลี่ยนไปตามกาลเวลาซึ่งข้อสมมติ proportional hazard ใช้ไม่ได้ในกรณีนี้ แต่อย่างไรก็ตามทฤษฎีของ Cox regression ได้รับการพัฒนาต่ออย่างรวดเร็วจนสามารถแก้ปัญหาในตัวอย่างดังกล่าวได้ทั้งหมดข้อสมมติ proportional hazard สามารถแก้ไขดัดแปลงได้ ปัจจุบันจึงเป็นทฤษฎีที่ใช้มากที่สุดในการติดตามบุคคลระยะยาวเพื่อหาปัจจัยเสี่ยง

Longitudinal Data Analysis

การพัฒนาทางทฤษฎีที่สำคัญที่จะกล่าวถึงประการสุดท้ายคือ Longitudinal Data Analysis หรือ การวิเคราะห์ข้อมูลซึ่งติดตามวัด outcome variable หรือตัวแปรตามซ้ำ ๆ ในบุคคลคนเดียวกันระยะยาวบางครั้งตัวแปรอิสระก็เปลี่ยนแปลงไปด้วย เช่น ต้องการทดสอบสมมติฐานว่าเด็กขาดวิตามินทำให้ป่วยเป็นหวัดได้ง่ายขึ้นหรือไม่ ต้องติดตามสังเกตว่าเป็นหวัดหรือไม่ (ซึ่งเป็นตัวแปรตาม)เป็นระยะ ๆ ในขณะเดียวกันก็ต้องวัดระดับการขาดวิตามิน (ซึ่งเป็นตัวแปรอิสระ) เป็นระยะ ๆ เช่นกัน นอกจากนี้ยังต้องพิจารณาว่าอายุเด็กซึ่งเป็น confounder ก็เปลี่ยนไปเรื่อย

ถ้าถือว่า unit of analysis คือการตรวจเด็กแต่ละครั้งโดยไม่คำนึงถึงว่าเด็กคนเดียวกันถูกตรวจหลาย ๆ ครั้งก็ถือว่าผิด เพราะเด็กแต่ละคนอาจจะตอบสนองต่อการขาดวิตามินเหมือนการตอบสนองของตนที่ผ่านมามากกว่าเหมือนการตอบสนองของเด็กคนอื่น นอกจากนี้ถ้าในช่วงที่ผ่านมาเป็นหวัด อาจจะทำให้ภูมิต้านทานเปลี่ยนไปอาจจะเป็นหวัดได้ง่ายขึ้นในช่วงต่อไปหรืออาจจะมีภูมิต้านทานทำให้เป็นหวัดน้อยลง ผลกระทบของการขาดวิตามินอาจจะไม่เปลี่ยนความเสี่ยงในสัปดาห์นั้นทันที แต่เปลี่ยนความเสี่ยงในสัปดาห์ถัดไป สรุปแล้ว longitudinal data analysis มีคำถามที่น่าสนใจในเชิงวิธีวิทยาการวิจัยมากมาย ไม่เฉพาะแต่ผลกระทบจากสิ่งที่สงสัยเท่านั้น แต่รวมถึงความสัมพันธ์ระหว่าง outcome ในช่วงเวลาต่าง ๆ ด้วย

Model ของ Longitudinal Data Analysis มีความหลากหลายน่าสนใจ กล่าวคือแบ่งเป็น 3 models ใหญ่ ๆ คือ ประการแรก Marginal Model หรือ Population-Average Model, ประการที่สอง Random-Effects Model และประการสุดท้าย คือ Transition Model

Marginal Model หรือ Population-Average Model เป็นการคิดค่าเฉลี่ยประมาณการทั้งประชากรโดยผู้วิจัยต่อระบุความสัมพันธ์ของ outcome ภายในตัวบุคคลเดิมตามกาลเวลา เช่น กรณีแรกอาจจะกำหนดว่าสหสัมพันธ์ (correlation) ระหว่างครั้งใด ๆ ก็ตามเหมือนกันทั้งนั้นไม่เกี่ยวกับกาลเวลา (exchangeable) ถ้าคนไหนเป็นหวัดบ่อยก็จะเป็นบ่อย ๆ ถ้าคนไหนไม่ค่อยเป็นหวัดก็จะไม่ค่อยเป็น หรือกรณีที่สองอาจจะกำหนดว่า สัมพันธ์ระหว่างสัปดาห์ที่ถัดกันจะสูงแต่ถ้าห่างกันออกไประดับสหสัมพันธ์จะน้อยลงตามลำดับ (auto-regressive) วิธีการทางสถิติที่สำคัญสำหรับ Marginal Model คือ Generalized Estimating Equations หรือ GEE ซึ่งแก้สมการหลายสมการพร้อม ๆ กันโดยหลักการทางแคลคูลัสที่เรียกว่า Score Statistics หลังการพัฒนาเชิงทฤษฎีไม่นานนัก ปัจจุบันมี software ที่ใช้ในการวิเคราะห์แบบนี้ในระดับหนึ่ง

Random-Effects Model สำหรับระบาดวิทยา เป็น model ที่ตั้งข้อสมมติว่าบุคคลแต่ละคนมีความเสี่ยง (logit) ภายในไม่เท่ากัน บางคนเสี่ยงมากบางคนเสี่ยงน้อย ปัจจัยภายนอกอื่น ๆ มีผลเพิ่มหรือลดความเสี่ยงพื้นฐานนี้เท่านั้น ความเสี่ยงที่แตกต่างกันไปนี้ส่วนใหญ่ถือว่าเป็น Gaussian random distribution ซึ่งมีค่าเบี่ยงเบนมาตรฐาน ถ้าค่าเบี่ยงเบนมาตรฐานกว้างความเสี่ยงของแต่ละคนจะแตกต่างกันมาก เมื่อมี random effects ของแต่ละคนแล้วก็ไม่จำเป็นต้องคำนึงถึงความสัมพันธ์ภายในบุคคลอีกต่อไป Software สำหรับการวิเคราะห์ Random-Effects Model ในปัจจุบันหาได้ไม่ยากนัก

Transition Model (หรือ Marcov model) เป็น model ที่เน้นการเปลี่ยนผ่านจากเวลาหนึ่งไปสู่เวลาถัดไปซึ่งเหมาะสมสำหรับการหาคำตอบว่าการเกิดโรคในช่วงหนึ่งทำให้ภูมิต้านทานในช่วงต่อไปเพิ่มขึ้นหรือลดลง การวิเคราะห์ transition model เป็นการวิเคราะห์ที่ง่ายที่สุดเพราะไม่ต้องการ software ใดพิเศษนอกเหนือจาก logistic regression ซึ่งผู้วิเคราะห์จะต้องใส่สภาวะการป่วยในช่วงที่ผ่านมาเป็นตัวแปรอิสระตัวหนึ่งสำหรับพยากรณ์การป่วยในปัจจุบัน

การพัฒนาวิธีวิทยาวิจัยในช่วงต่อสหัสวรรษที่กำลังจะได้เห็น

การพัฒนาทางด้านวิทยาการคอมพิวเตอร์น่าจะเป็นหลักสำหรับการพัฒนาวิธีวิทยาการวิจัยในช่วงทศวรรษนี้ ปัญหาการวิเคราะห์ที่น่าสนใจแต่ไม่สามารถทำได้ในอดีตเพราะคอมพิวเตอร์ไม่มีความสามารถมากพอได้แก่การวิเคราะห์แบบ exact probability methods สำหรับ multi-variate analysis ซึ่งจะให้ค่าประมาณการที่ถูกต้องมากกว่าวิธีการที่ใช้อยู่ในปัจจุบันถ้าข้อมูลมีขนาดเล็ก นอกจากนี้ยังมีแนวคิดแบบ Bayesianism แนวคิดนี้หาคำตอบจากข้อมูลอย่างมีเงื่อนไข กล่าวคือผู้ต้องการคำตอบต้องระบุว่าเงื่อนไขความเชื่อโดยทั่วไปของเขาก่อนเห็นข้อมูลเป็นอย่างไร จากนั้นจึงคำนวณเพื่อหาว่าภายใต้เงื่อนไขของข้อมูลที่มีอยู่ความเชื่อของเขาน่าจะถูกหรือผิดเพียงไร แนวคิดแบบนี้ได้เริ่มใช้ในสถิติแขนงอื่นและน่าจะแพร่หลายเข้ามาสู่สาขาระบาดวิทยามากขึ้นในไม่ช้า

สรุป

บทความนี้ชี้ให้เห็นวิธีวิทยาการวิจัยสำหรับวิชาระบาดวิทยาซึ่งผ่านการพัฒนามาประมาณสามศตวรรษ เริ่มต้นจากการค้นหาสาเหตุของโรคระบาดก่อนความเจริญก้าวหน้าทางเทคโนโลยีจุลชีววิทยา จากนั้นเงียบหายไปเป็นเวลานานจนเข้าสู่ยุคฟื้นฟูหลังสงครามโลกครั้งที่สอง เนื่องจากการนำวิชาสถิติเข้ามาประยุกต์อย่างเข้มข้น ในรอบสองทศวรรษที่ผ่านมา การบริหารจัดการทางระบาดวิทยาก็ได้ก้าวหน้ามากขึ้นทำให้สามารถติดตามบุคคลจำนวนมากในระยะยาวเพื่อเก็บข้อมูลหาความสัมพันธ์ระหว่างปัจจัยที่เป็นเหตุกับผลที่เกิดขึ้นในอนาคตได้ และที่สำคัญก็คือความก้าวหน้าทางวิทยาการคอมพิวเตอร์ทำให้การเก็บรักษาฐานข้อมูลตลอดจนการวิเคราะห์ข้อมูลจำนวนมากเป็นไปได้ดีขึ้นอย่างมาก การวิเคราะห์ logistic regression และ Cox regression นับเป็นความก้าวหน้าในทางวิธีวิทยาการวิจัยที่สำคัญในสองทศวรรษที่ผ่านมา ส่วนการวิเคราะห์ longitudinal data analysis กำลังเริ่มเป็นที่แพร่หลายมากขึ้น ในอนาคตอันใกล้นี้คาดว่าวิธีการวิเคราะห์ข้อมูลและการได้ข้อสรุปจากข้อมูลจะยิ่งซับซ้อนกว่าสภาพที่เป็นอยู่

เอกสารอ้างอิง

Breslow, N. E., & Day, N. E. (1980). ***Statistical Methods in Cancer Research. Volume I- The Analysis of Case-Control Studies. IARC Scientific Publication No 32.*** International Agency for Research on Cancer. Lyon, France.

Breslow, N. E., & Day, N. E. (1987). ***Statistical Methods in Cancer Research. Volume II- The Design and Analysis of Cohort Studies. IARC Scientific Publications No. 82.*** International Agency for Research on Cancer. Lyon, France.

Chalmers, I. & Altman, D. A. (1995). ***Systematic Reviews.*** BMJ Publishing Group. Birstol, UK.

Clayton, D. & Hills, M. (1998). ***Statistical Models in Epidemiology.*** Oxford Science Publications. UK.

Cox, D. R. (1972). Regression Model and Life Table. ***Journal of the Royal Statistical Society Series B. 34,*** 187-220.

Diggle, P .J. , Liang, K. Y. & Zegger, S. L. (1994). ***The Analysis of Longitudinal Data (Oxford Statistical Science, No 13),*** London, UK.

Hosmer, D. W. & Lemeshow, S. (1989). ***Applied Logistic Regression.*** John Wiley & Son. New York. USA.

Hosmer, D. W. & Lemeshow, S. (1999). ***Applied Survival Analysis: Regression Modeling of Time to Event Data.*** John Wiley & Sons. New York, USA.

Kleinbuam, D. G. (1994). ***Logistic Regression : A Self-Learning Text*** (Springer Series in Statistics. Statistics in the Health Sciences.). USA.

Kleinbuam, D. G. (1999). ***Survival Analysis : A Self-Learning Text*** (Springer Series in Statistics.Statistics in the Health Sciences). USA.

Lilienfeld, D. E. & Stolley, P. D. (1994). ***Foundations of Epidemiology- Third Edition.*** Oxford University Press, London, UK.

McNeil, D. (1996). ***Epidemiological Research Methods.*** John Wiley & Son. West Sussex, UK.

Rothman, K. J. & Greenland S. (1998). ***Modern Epidemiology.*** Chapter 2 - Causation and Causal Inference. Lippincott- Raven Publisher. Philadelphia, PA USA.