Thai Journal of Pharmacy Practice

บทความวิจัย

ความถูกต้องของข้อมูลทะเบียนมะเร็งไทยเมื่อใช้การเชื่อมโยงข้อมูลแบบความน่าจะเป็น

วรรณพร วัฒนะวงษ์[1], แสวง วัชระธนกิจ[2]

[1]นักศึกษาหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาเภสัชศาสตร์ คณะเภสัชศาสตร์ มหาวิทยาลัยอุบลราชธานี
[2]กลุ่มวิชาเภสัชกรรมปฏิบัติ คณะเภสัชศาสตร์ มหาวิทยาลัยอุบลราชธานี

บทคัดย่อ

วัตถุประสงค์: เพื่อสร้างและทดสอบโมเดลการเชื่อมโยงข้อมูลในทะเบียนมะเร็งไทยและข้อมูลจากฐานข้อมูลโรงพยาบาลตติยภูมิแห่งหนึ่งด้วยหลักความน่าจะเป็น และเพื่อทดสอบความถูกต้องของข้อมูลทะเบียนมะเร็งไทยจากการใช้วิธีการดังกล่าว **วิธีการ**: การศึกษาเชื่อมโยงข้อมูลทั้งสองแหล่งโดยใช้หมายเลขอ้างอิงที่เข้ารหัสใหม่ ด้วยการเชื่อมโยงแบบกำหนดตายตัวแบบหนึ่งต่อหนึ่ง และการเชื่อมโยงแบบความน่าจะเป็นจำนวน 17 โมเดล แต่ละโมเดลประกอบด้วยตัวแปรที่เป็นตัวระบุ (identifier) และไม่ใช่ตัวระบุ (non-identifiers) ได้แก่ หมายเลขประจำตัวอ้างอิง หมายเลขโรงพยาบาลอ้างอิง ชื่อ นามสกุล ที่อยู่ที่ติดต่อได้ วันที่เกิด วันที่เสียชีวิต รหัสจังหวัดที่อยู่ตามทะเบียนบ้านออกโดยกรมการปกครอง เพศ และรหัสการวินิจฉัยโรค โดยเลือกมาจำนวน 3-8 ตัวแปรเพื่อสร้างแต่ละโมเดล พร้อมกำหนดค่าคะแนนจุดตัดคะแนนที่เหมาะสม **ผลการวิจัย:** ฐานข้อมูลทั้งสองมีผู้ป่วยจำนวน 7,243 รายที่ตรงกัน คิดเป็นร้อยละ 89.72 จากผู้ป่วยในฐานข้อมูลทะเบียนมะเร็งไทยทั้งหมด 8,073 คน หรือร้อยละ 36.72 ของผู้ป่วยจากฐานข้อมูลทั้งสองรวม 19,725 ราย ทุกโมเดลของการเชื่อมโยงแบบความน่าจะเป็นให้ค่าพยากรณ์ผลบวกร้อยละ 99.83-99.94 ค่าพยากรณ์ผลลบร้อยละ 97.04-99.98 ค่าความไวร้อยละ 94.74-99.97 และค่าความจำเพาะร้อยละ 99.90-99.97 **สรุป:** การเชื่อมโยงข้อมูลมะเร็งจากสองฐานข้อมูลโดยใช้การเชื่อมโยงแบบความน่าจะเป็นทำให้พบผู้ป่วยที่ตรงกันมากขึ้น จึงเหมาะสมกับการใช้ในงานใด ๆ หรืองานวิจัยที่ข้อมูลมีการปกปิดตัวตนและการรักษาความลับมีความสำคัญ อย่างไรก็ตามความถูกต้องของโมเดลที่ใช้การเชื่อมโยงข้อมูลแบบความน่าจะเป็นนั้นขึ้นอยู่กับการเลือกตัวแปรที่ใช้ในโมเดลเป็นหลัก ดังนั้นนักวิจัยควรพิจารณาเลือกตัวแปรที่ใช้ในสมการทำนายอย่างรอบคอบ

คำสำคัญ: ความถูกต้องของข้อมูล ฐานข้อมูลทะเบียนมะเร็งไทย ฐานข้อมูลโรงพยาบาล การเชื่อมโยงข้อมูลแบบความน่าจะเป็น

รับต้นฉบับ: 1 เม.ย. 2566, **ได้รับบทความฉบับปรับปรุง**: 3 พ.ค. 2566, **รับลงตีพิมพ์**: 8 พ.ค. 2566

ผู้ประสานงานบทความ: แสวง วัชระธนกิจ กลุ่มวิชาเภสัชกรรมปฏิบัติ คณะเภสัชศาสตร์ มหาวิทยาลัยอุบลราชธานี อำเภอวารินชำราบ จังหวัดอุบลราชธานี 34190 **E-mail**: sawaeng.w@ubu.ac.th

TJPP วารสารเภสัชกรรมไทย ปีที่ 16
เล่มที่ 3 ก.ค.-ก.ย. 2567

PRINCE OF SONGKLA UNIVERSITY
Thai Journal of Pharmacy Practice

Research article

Validity of Data in Thailand Cancer-Based Registry with the Application of Probabilistic Record Linkage

Wannaporn Wattanawong[1], Sawaeng Watcharathanakij[2]

[1]Student in Doctor of Philosophy Program in Pharmaceutical Sciences, Ubon Ratchathani University

[2]Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Ubonratchathani University

Abstract

Objectives: To develop and validate probabilistic data linkage models linking data in cancer registry and data in one tertiary care hospital, and to validate data in cancer registry by using such method. **Method:** The study linked the data from two sources by using deterministic record linkage with 1-1 matching with newly encoded identification numbers and 17 probabilistic record linkage based models. Each model employed either identifiers or non-identifiers including encoded identification numbers, encoded hospital number, first name, last name, contact address in house registration, date of birth, date of death, zip code of residence in house registration issued by Department of Provincial Administration, sex, and ICD-10. Each model employed 3 to 8 variables, and the optimal cutoff points were determined. **Result:** There were 7,243 matched patients in both databases accounting for 89.72% of 8,073 patients in cancer registry and 36.72% of 19,725 patients in two databases. All models of probabilistic record linkage had positive predictive values between 99.83 to 99.94%, negative predictive values between 97.04 to 99.98%, sensitivity between 94.74-99.97%, and specificity between 99.90-99.97%. **Conclusion:** Linking data on cancer patients between two databases using probabilistic record linkage led to a higher number of matched patients than using deterministic record linkage. Therefore, it could be applied in any work or research where data anonymity and confidentiality are important. However, the validity of models using probabilistic record linkage largely depends upon selected variables in the model. Therefore, researchers should carefully select the variables used in the prediction equation.

Keywords: validity of data, cancer registry database, hospital database, probabilistic record linkage

บทนำ

การเชื่อมโยงข้อมูลเป็นส่วนหนึ่งของกระบวนการจัดเตรียมข้อมูลเพื่อการวิเคราะห์ หลังจากที่ข้อมูลได้ผ่านการปรับให้เป็นมาตรฐานเดียวกัน (data standardization) แล้ว การเชื่อมโยงข้อมูลช่วยให้ข้อมูลย่อย ๆ ถูกโยงกันเป็นเซตข้อมูลขนาดใหญ่ที่มีความละเอียดและมีหลายมิติมากขึ้น ทั้งยังช่วยขจัดระเบียนที่ซ้ำซ้อนกันในเซตข้อมูล ตัวอย่างงานที่ต้องมีการเชื่อมโยงข้อมูล ได้แก่ การวิจัยด้านการแพทย์หรือระบาดวิทยาโดยเชื่อมโยงข้อมูลสำมะโนประชากร ข้อมูลผู้ป่วย และข้อมูลการรักษาเฉพาะโรค เพื่อศึกษาปัจจัยเสี่ยงหรือผลสัมฤทธิ์ในการรักษาโรคในกลุ่มประชากรที่สนใจ (1) การเชื่อมโยงข้อมูลสามารถนำไปประยุกต์ใช้กับข้อมูลของโรคต่าง ๆ ได้ โดยเฉพาะโรคมะเร็งที่เป็นปัญหาสำคัญอันดับต้นของระบบสาธารณสุขของประเทศต่าง ๆ ทั่วโลก เพราะทำให้เกิดความสูญเสียทั้งชีวิตและส่งผลกระทบต่อเศรษฐกิจ

องค์การอนามัยโลกประมาณการณ์ว่าในช่วงปี ค.ศ.2007-2030 จะมีผู้ป่วยที่เสียชีวิตด้วยโรคมะเร็งทั่วโลกเพิ่มขึ้นจาก 7.9 ล้านคน เป็น 11.5 ล้านคน และมีผู้ป่วยรายใหม่ทั่วโลกเพิ่มขึ้นจาก 11.3 ล้านคน เป็น 15.5 ล้านคน (2) สำหรับประเทศไทยในปี พ.ศ.2556-2558 พบผู้ป่วยโรคมะเร็งรายใหม่เพศชาย 61,416 ราย เพศหญิง 65,139 ราย อัตราอุบัติการณ์การเสียชีวิตตามอายุ (mean annual age-standardized incidence rate) ต่อประชากร 100,000 รายเท่ากับ 143.8 ในเพศชาย และ 134.2 ในเพศหญิง (3) คณะทำงานเครือข่ายผู้เชี่ยวชาญด้านโรคมะเร็งของเขตบริการสุขภาพ ภายใต้นโยบายและการกำกับดูแลของกระทรวงสาธารณสุข service plan สาขาโรคมะเร็ง กำหนดตัวชี้วัดระดับเป้าประสงค์ในเรื่อง ลดอัตราตาย ลดอัตราป่วย และลดระยะเวลาการรอคอย ซึ่งต้องอาศัยการบริหารจัดการโดยใช้ข้อมูลสารสนเทศโรคมะเร็ง (cancer informatics) ซึ่งเป็นตัวชี้วัดระดับยุทธศาสตร์ โดยในส่วนของยุทธศาสตร์ที่ 6 สารสนเทศโรคมะเร็งมี 3 ตัวชี้วัด ได้แก่ 1) ร้อยละของโรงพยาบาลในสังกัดกระทรวงสาธารณสุขที่มีการรายงานข้อมูลทะเบียนมะเร็งระดับโรงพยาบาล 2) ร้อยละของโรงพยาบาลในสังกัดกระทรวงสาธารณสุขที่มีการส่งข้อมูลทะเบียนมะเร็งไปยังเว็บไซต์ และ 3) เขตบริการสุขภาพมีการทำทะเบียนมะเร็งระดับประชากร (population based) เกณฑ์ผ่านการประเมินตัวชี้วัดปี 2564-2565 คือ ≥ร้อยละ 90 (สะสม) ในตัวชี้วัดที่ 1 และ 2 (4) ดังนั้นจึงต้องมีการจัดทำทะเบียนมะเร็งระดับโรงพยาบาลเพื่อนำไปรวบรวมในระดับประเทศสำหรับตอบตัวชี้วัดและเป้าหมายดังกล่าวข้างต้น

โดยทั่วไปการจัดทำทะเบียนมะเร็งในประเทศไทยใช้เกณฑ์ในการจัดทำเดียวกันซึ่งแบ่งข้อมูลเป็น 5 กลุ่ม คือ ข้อมูลบุคคล ข้อมูลการป่วย ข้อมูลการรักษา ข้อมูลสถานภาพ และแหล่งข้อมูล ข้อมูลบุคคลเก็บตามที่ปรากฏในเวชระเบียนหรือฐานข้อมูลผู้ป่วยของสถานพยาบาล ทั้งนี้เป็นไปตามเกณฑ์ของ The International Association of Cancer Registries (IACR) (5) เพื่อให้การรวบรวมข้อมูลโรคมะเร็งที่เป็นปัจจุบันถูกต้องและน่าเชื่อถือ อย่างไรก็ตามการนำข้อมูลผู้ป่วยโรคมะเร็งจากเวชระเบียนกระดาษ หรือเวชระเบียนอิเล็กทรอนิกส์เข้าสู่ทะเบียนมะเร็งของประเทศมีหลากหลายรูปแบบขึ้นกับบริบทของแต่ละสถานพยาบาล เช่น บางโรงพยาบาลนำข้อมูลจากโรงพยาบาลบันทึกในรูปของสเปรดชีตหรือโปรแกรมจัดทำทะเบียนมะเร็งที่ทางโรงพยาบาลพัฒนาขึ้นมาเอง เพื่อส่งให้สถาบันมะเร็งแห่งชาติรวบรวมจัดทำรายงานประจำปี และมีการลงทะเบียนโดยบุคลากรที่มีความหลากหลาย ได้แก่ แพทย์ พยาบาล นักเวชสถิติ นักวิชาการคอมพิวเตอร์ หรือเจ้าหน้าที่บันทึกข้อมูลที่ผ่านการอบรมการจัดทำทะเบียนมะเร็งเบื้องต้นเป็นอย่างน้อย

สำหรับการพัฒนาระบบสารสนเทศและระบบฐานข้อมูลทะเบียนมะเร็งในประเทศไทยนั้น เริ่มการพัฒนาฐานข้อมูลโดยใช้โปรแกรมในการรวบรวมข้อมูลคือ Thai Cancer Based Program (TCB) ซึ่งในปัจจุบันกระทรวงสาธารณสุขได้กำหนดให้โรงพยาบาลในสังกัดกระทรวงสาธารณสุขจัดทำรายงานข้อมูลมะเร็งระดับโรงพยาบาลและส่งข้อมูลทะเบียนมะเร็งไปยังเว็บไซต์ผ่านทางโปรแกรม TCB แต่เนื่องจากความหลากหลายของระบบสารสนเทศและผู้ปฏิบัติงาน จึงอาจเกิดความคลาดเคลื่อนของข้อมูลที่รวบรวมได้ ซึ่งโดยทั่วไปการตรวจสอบความถูกต้องของข้อมูลในฐานข้อมูลนั้นมักใช้วิธีการมาตรฐาน คือ การเชื่อมโยงด้วยหมายเลขบัตรประจำตัวประชาชน 13 หลัก ซึ่งเป็นการเชื่อมโยงข้อมูลแบบกำหนดตายตัว (deterministic data linkage) แต่ถ้าไม่มีข้อมูลที่บ่งชี้ผู้ป่วยโดยตรง เช่น ไม่ทราบหมายเลขบัตรประจำตัวประชาชน อาจใช้วิธีการเชื่อมโยงข้อมูลแบบความน่าจะเป็น (probabilistic data linkage) (1) ที่มีข้อดีกว่า คือ สามารถบอกได้ว่าบุคคลนั้นเป็นบุคคลเดียวกัน (หรือเป็นระเบียน

เดียวกัน) จากตัวแปรอื่น และทำให้เกิดการใช้ประโยชน์จากข้อมูลได้อย่างเต็มที่

งานวิจัยที่ศึกษาเรื่องการเชื่อมโยงข้อมูลจากฐานข้อมูลด้านสุขภาพมีจำนวนมากในประเทศสหรัฐอเมริกา อังกฤษ และออสเตรเลีย (6-8) ทั้งนี้การเชื่อมโยงข้อมูลจากแหล่งข้อมูลเดียวกันหรือจากแหล่งข้อมูลข้อมูลต่างกัน มีเป้าหมายเพื่อพิจารณาว่ามีระเบียนใดเป็นข้อมูลของบุคคลเดียวกัน ในทางปฏิบัติเพื่อรักษาความเป็นส่วนตัวของบุคคลที่เป็นเจ้าของข้อมูล องค์กรจำเป็นต้องปกปิดตัวบ่งชี้ที่ระบุถึงตัวบุคคลได้ เช่น ชื่อ-นามสกุล หรือเลขประจำตัวประชาชน การขาดตัวบ่งชี้เหล่านี้ไปจึงเป็นอุปสรรคสำคัญต่อการเชื่อมโยงข้อมูลแบบที่ใช้กันโดยทั่วไป คือ การเชื่อมโยงข้อมูลแบบกำหนดตายตัวที่ใช้กุญแจหลักหรือตัวบ่งชี้ที่ไม่มีค่าซ้ำ (unique identifier) เช่น เลขประจำตัวประชาชน โดยกำหนดเงื่อนไขว่า ตัวบ่งชี้ในระเบียนจากแหล่งหรือฐานข้อมูลหลักและระเบียนใด ๆ จากแหล่งหรือฐานข้อมูลที่สองต้องมีค่าตรงกัน (exact match) จึงจะถือว่าทั้งสองระเบียนเป็นข้อมูลของบุคคลคนเดียวกัน แต่ในทางปฏิบัติข้อมูลจากแหล่งต่าง ๆ อาจมีความคลาดเคลื่อน จึงทำให้ผลการเชื่อมโยงข้อมูลนั้นพบว่าไม่ได้เป็นระเบียนเดียวกัน เช่น หมายเลขประจำตัว 9781264 กับ 9782264 ที่มีตัวเลขแตกต่างกันเพียงตำแหน่งเดียว ดังนั้นการเชื่อมโยงข้อมูลแบบเชิงความน่าจะเป็น (1) ที่อาศัยหลักการความน่าจะเป็นว่า ข้อมูลจากทั้งสองแหล่งจะเป็นระเบียนเดียวกันหรือไม่ โดยใช้ตัวแปรตัวเดียวหรือหลาย ๆ ตัวร่วมกัน อาทิเช่น ตัวแปรชื่อ-นามสกุล “ดวงเดือน ดวงตะวัน” และตัวแปรเพศ “หญิง” ของระเบียนจากฐานข้อมูลหลัก และ “ดวนเดือน ดวงตะวัน” และ “หญิง” จากฐานข้อมูลรอง มีความน่าจะเป็นในระดับใดที่จะเป็นระเบียนของบุคคลเดียวกันเป็นต้น ข้อมูลแบบเชิงความน่าจะเป็นอาศัยการจับคู่ตัวบ่งชี้แบบประมาณ (approximate match) โดยมีหลักการวัดความต่างกันของระเบียนข้อมูลดังนี้ สมมติให้ตัวบ่งชี้หรือตัวระบุ (identifiers) ที่ใช้เชื่อมโยงข้อมูลระเบียน a จากฐานข้อมูลหลัก และระเบียน b จากฐานข้อมูลที่สองมีอยู่ k ตัว ความต่างกันของตัวบ่งชี้ทั้ง k ตัว จะถูกประมวลผลด้วยฟังก์ชันการตัดสินใจเพื่อหาข้อสรุปว่า ระเบียน a และระเบียน b น่าจะเป็นระเบียนของบุคคลคนเดียวกันหรือไม่ ทั้งนี้ตัวระบุแต่ละตัวนั้นมีโอกาสหรือความสามารถแยกระเบียนว่าตรงกันหรือไม่ตรงกันแตกต่างกันออกไป ขึ้นกับจำนวนค่าของตัวระบุนั้น ๆ เช่น เพศ มี 2 ค่า คือ ชายหรือหญิง ดังนั้นโอกาสที่ข้อมูลจะตรงกันหรือเป็นระเบียนเดียวกัน (matched) คือ ร้อยละ 50 (1/2) หรือ เดือนเกิด มี 12 เดือน ดังนั้นโอกาสที่ข้อมูลจะตรงกันหรือเป็นระเบียนเดียวกัน (matched) คือ ร้อยละ 8.30 (1/12) โอกาสที่ข้อมูลจะตรงกันโดยบังเอิญจะน้อยกว่าเพศ ดังนั้นการรวมตัวระบุหลายตัวจะเพิ่มจำนวน unique values ทำให้ลดโอกาสที่จะตรงกันโดยบังเอิญ ตัวระบุที่มักใช้ในการเชื่อมโยงข้อมูลผู้ป่วย เช่น หมายเลขประกันสังคม หมายเลขประจำตัวผู้ป่วย ชื่อตัว ชื่อสกุล วันเกิด วันตาย วันที่ได้รับการวินิจฉัยโรค วันที่รักษา เพศ ที่อยู่ เป็นต้น (9)

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างและทดสอบโมเดลการเชื่อมโยงข้อมูลระหว่างทะเบียนมะเร็งไทยและฐานข้อมูลของโรงพยาบาลตติยภูมิแห่งหนึ่งด้วยหลักความน่าจะเป็น และเพื่อทดสอบความถูกต้องของข้อมูลทะเบียนมะเร็งไทยด้วยการเชื่อมโยงข้อมูลแบบความน่าจะเป็น

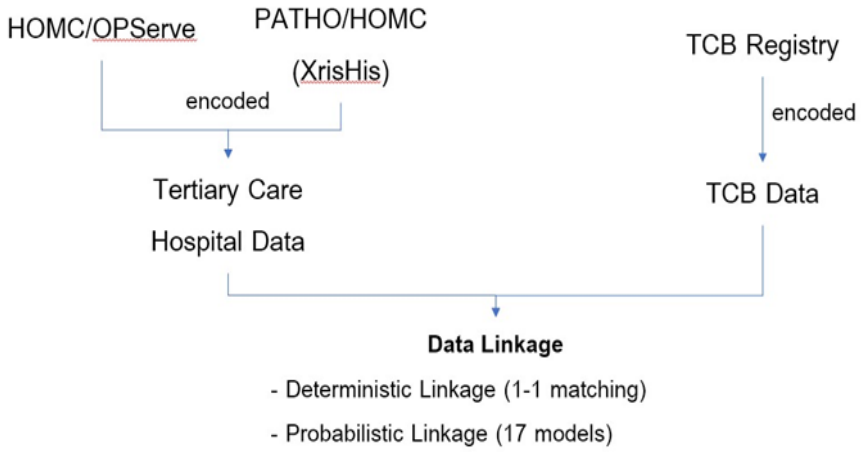
วิธีการวิจัย

สถานที่วิจัย

การวิจัยนี้ผ่านการพิจารณาจริยธรรมการวิจัยในมนุษย์ของโรงพยาบาลที่ทำการวิจัย เลขที่ 011/2563 สถานที่วิจัยเป็นโรงพยาบาลระดับตติยภูมิแห่งหนึ่งในภาคตะวันออกเฉียงเหนือของประเทศไทย มีจำนวนเตียงให้บริการผู้ป่วยจำนวน 1,158 เตียง โรงพยาบาลเป็นศูนย์ความเชี่ยวชาญด้านโรคมะเร็ง โดยมีผู้ป่วยโรคมะเร็งรายใหม่ที่เข้ารับการรักษามากกว่า 1,000 คนต่อปี

การรวบรวมข้อมูล

การรวบรวมข้อมูลจากโรงพยาบาลทำโดยนักวิชาการคอมพิวเตอร์ใช้คำสั่ง structured query language (SQL) รวบรวมข้อมูลผู้ป่วยโรคมะเร็งสัญชาติไทยทุกรายที่มีผลการวินิจฉัยยืนยันว่าเป็นโรคมะเร็งด้วยรหัส ICD-10 (C00-C97) (5) ที่เข้ารับการรักษาในโรงพยาบาลระหว่างวันที่ 1 มกราคม 2558 ถึง 31 ธันวาคม 2561 ฐานข้อมูลของโรงพยาบาลใช้ข้อมูลจาก 2 โปรแกรม ได้แก่ HOMC/OPServe ที่ใช้ดึงข้อมูลตัวแปรข้อมูลส่วนบุคคล ข้อมูลการรักษา ข้อมูลโรค และข้อมูลการติดตาม และ Patho/HOMC (XrisHis) ที่ให้ข้อมูล ได้แก่ ผลทางพยาธิวิทยา behavior และ grade จากนั้นทำการเข้ารหัสหมายเลขประจำตัวใหม่อย่างเป็นระบบเพื่อให้รหัสประจำตัวไม่ซ้ำกัน และใช้การเข้ารหัสเดียวกันนี้กับข้อมูลที่ดึงมาจากฐานข้อมูลทะเบียนมะเร็ง



รูปที่ 1. กระบวนการการเชื่อมโยงข้อมูลของผู้ป่วยโรคมะเร็งในฐานข้อมูลทะเบียนมะเร็งเทียบกับข้อมูลต้นฉบับจากฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์ของโรงพยาบาล

การรวบรวมข้อมูลจากฐานข้อมูลทะเบียนมะเร็งไทยได้จากการส่งออกข้อมูล 33 เขตข้อมูลในรูปแบบสเปรดชีตจากโปรแกรม TCB ได้แก่ หมายเลขอ้างอิง ชื่อ-นามสกุล วัน/เดือน/ปีเกิด เพศ สถานภาพสมรส สัญชาติ เชื้อชาติ ศาสนา ที่อยู่ตามทะเบียนบ้าน ที่อยู่ที่ติดต่อได้หรือที่อยู่ปัจจุบัน ชื่อโรงพยาบาล/รหัสโรงพยาบาล เลขที่อ้างอิงผู้ป่วยในโรงพยาบาล วันที่วินิจฉัยว่าเป็นมะเร็ง อายุ ณ วันที่วินิจฉัย วิธีวินิจฉัย หมายเลขชิ้นเนื้อพยาธิวิทยา วันที่ตัดชิ้นเนื้อหรือวันที่ส่งชิ้นเนื้อ วันที่อ่านชิ้นเนื้อ ตำแหน่งอวัยวะที่เป็น การกลับเป็นซ้ำ ผลทางพยาธิวิทยา behavior grade T N M Stage Extend Metastasis สภาพที่เป็นอยู่ล่าสุด วันที่ติดต่อล่าสุด วันที่เสียชีวิต สาเหตุการเสียชีวิต refer in (ส่งมาจากไหน) refer out (ส่งต่อไปยัง) และวิธีการรักษา พร้อมเข้ารหัสหมายเลขประจำตัวเช่นเดียวกับวิธีการที่กล่าวมาแล้วข้างต้น

การเชื่อมโยงข้อมูล

ผู้วิจัยใช้การเชื่อมโยงแบบกำหนดตายตัวแบบ 1 ต่อ 1 โดยใช้หมายเลขประจำตัวอ้างอิง (encoded identification numbers) และเชื่อมโยงข้อมูลแบบความน่าจะเป็นซึ่งทำการเข้ารหัสใหม่อย่างเป็นระบบ เพื่อไม่ให้สามารถทราบได้ว่าหมายเลขประจำตัวประชาชน 13 หลักที่แท้จริงคือหมายเลขใด เพื่อใช้ค้นหาตัวบุคคลที่ตรงกันในโมเดลที่ใช้ทดสอบ การเชื่อมโยงแบบความน่าจะเป็นสามารถทำได้ 2 แบบ คือ 1) การเชื่อมโยงโดยใช้ข้อมูลจากทั้งสองแหล่ง โดยไม่ต้องใช้ผลการเชื่อมโยงในส่วนที่ไม่ตรงกันจากการเชื่อมโยงแบบกำหนดตายตัว หรือ 2) การเชื่อมโยงเฉพาะข้อมูลส่วนที่ไม่ตรงกัน หลังจากได้ผลจากการเชื่อมโยงแบบกำหนดตายตัวเป็นที่เรียบร้อยแล้ว ดังแสดงในรูปที่ 1

ผู้วิจัยสร้างโมเดลหรือสมการทำนายด้วยการเชื่อมโยงแบบความน่าจะเป็น โดยมีตัวแปรที่เลือกใช้ทั้งหมดจำนวน 9 ตัวแปร ได้แก่ หมายเลขประจำตัวอ้างอิงหรือหมายเลขโรงพยาบาลอ้างอิง (encoded hospital number) ชื่อ (first name) นามสกุล (last name) ที่อยู่ที่ติดต่อได้ (address) วันที่เกิด (date of birth) วันที่เสียชีวิต (date of death) รหัสจังหวัดที่อยู่ตามทะเบียนบ้านออกโดยกรมการปกครอง (province code) เพศ (sex) และรหัสการวินิจฉัยโรค (ICD-10) ในแต่ละสมการใช้ตัวแปรครั้งละ 3-8 ตัวแปร ทำให้ได้โมเดลการเชื่อมโยงแบบความน่าจะเป็นจำนวน 17 โมเดลที่มีการให้ค่าน้ำหนักของตัวแปรแต่ละตัวแตกต่างกัน (รูปที่ 1) การกำหนดค่าน้ำหนักได้จากการทบทวนวรรณกรรมจากงานของ Kranker (10)

หากข้อมูลจากทั้งสองแหล่งตรงกันจะได้ค่าน้ำหนักคะแนนของตัวแปรเป็นบวก แต่หากข้อมูลจากทั้งสองแหล่งไม่ตรงกันจะได้ค่าคะแนนติดลบ ยกตัวอย่างเช่น โมเดลที่ 1 ดังแสดงในตารางที่ 1 ตัวแปรหมายเลขอ้างอิง (id1) ให้คะแนนน้ำหนักเมื่อข้อมูลตรงกันเท่ากับ 15 และเมื่อข้อมูลไม่ตรงกันเท่ากับ -5 ในโมเดลที่ 1 กำหนด cutoff ที่คะแนนเท่ากับ 21

ผู้วิจัยยังนำข้อมูลส่วนที่ไม่ตรงกันหรือไม่เป็นระเบียนเดียวกัน (unmatched) จากการเชื่อมโยงแบบกำหนดตายตัว (ส่วน A และ C ในรูปที่ 2) มาเชื่อมโยงแบบ

ตารางที่ 1. ค่าคะแนนของแต่ละตัวแปรในโมเดลการเชื่อมโยงแบบความน่าจะเป็น: ตัวอย่างโมเดลที่ 1

ตัวแปรที่	ตัวแปร	ค่า	น้ำหนัก	
			ข้อมูลตรงกัน	ข้อมูลไม่ตรงกัน
ตัวแปรที่ 1	หมายเลขอ้างอิง	id1	15	-5
ตัวแปรที่ 2	ชื่อ	first name	5	-3
ตัวแปรที่ 3	ชื่อสกุล	last name	7	-3
ตัวแปรที่ 4	เลขที่บ้านที่อยู่อาศัยตามทะเบียนบ้าน	address1	5	-3
ตัวแปรที่ 5	วันเกิด	date of birth	8	-2
ตัวแปรที่ 6	รหัสจังหวัดที่อยู่ตามทะเบียนบ้าน	province code	5	-3
ตัวแปรที่ 7	เพศ	sex	5	-3
ตัวแปรที่ 8	รหัสกลุ่มโรค	ICD10	5	-3

ความน่าจะเป็นด้วยโมเดลที่ทำการทดสอบ เพื่อให้ทราบว่าจะพบผู้ป่วยที่เป็นบุคคลเดียวกันเพิ่มขึ้นหรือไม่ (รูปที่ 2)

สมการทำนายจะคำนวณค่าความน่าจะเป็นที่ระเบียนหรือข้อมูลในแถวหนึ่ง ๆ จากสองฐานข้อมูลเป็นระเบียนที่ตรงกันจริงหรือเป็นบุคคลเดียวกันจริง (M-probability) และความน่าจะเป็นที่ระเบียนใด ๆ จะไม่เป็นข้อมูลที่ตรงกันจริง (U-probability) (11) จากนั้นนำผลการเชื่อมโยงข้อมูลที่ได้ทั้งหมดไปเปรียบเทียบกับผลการเชื่อมโยงข้อมูลด้วยวิธีมาตรฐาน (gold standard) ที่ได้จากการเชื่อมโยงแบบกำหนดตายตัวแบบ 1 ต่อ 1 ผู้วิจัยตรวจสอบข้อมูลซ้ำอีกครั้งทุกระเบียนสำหรับผลการเชื่อมโยงของทุกโมเดลที่ทดสอบ

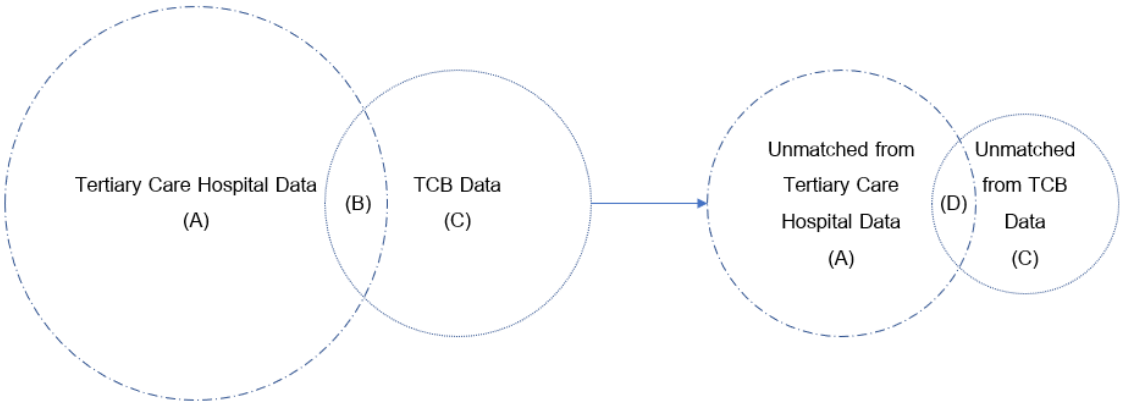
การวิเคราะห์ข้อมูล

การศึกษารายงานความถูกต้องของโมเดล (validity) ในการเชื่อมโยงข้อมูลแบบความน่าจะเป็นด้วยค่าร้อยละความสอดคล้อง (% matched) ค่าพยากรณ์ผลบวก (positive predictive values; PPVs) ค่าพยากรณ์ผลลบ (negative predictive value; NPV) ค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) ซึ่งคำนวณดังแสดงในตารางที่ 2

ผลการวิจัย

ความสอดคล้องที่พบจากการเชื่อมโยง

การศึกษาพบว่า ฐานข้อมูลทั้งสองมีผู้ป่วยจำนวน 7,243 รายที่ตรงกันหรือคิดเป็นร้อยละ 89.72 จากผู้ป่วยในฐานข้อมูลทะเบียนมะเร็งไทยทั้งหมด 8,073 คน หรือร้อยละ 36.72 ของผู้ป่วยจากฐานข้อมูลทั้งสองรวมทั้งหมด 19,725 ราย ผลการเชื่อมโยงข้อมูลจากแหล่งข้อมูลทั้งสองด้วยการเชื่อมโยงแบบความน่าจะเป็น โดยใช้ข้อมูลทั้งหมดเชื่อมโยงพร้อมกัน (ทั้งส่วนที่ตรงกันและไม่ตรงกันจากการเชื่อมโยง



รูปที่ 2. กระบวนการการเชื่อมโยงแบบความน่าจะเป็นด้วยโมเดลความน่าจะเป็นโดยใช้ข้อมูลส่วนที่ตรวจสอบไม่พบจากการเชื่อมโยงแบบตายตัว (A และ C)

ตารางที่ 2. ความถูกต้องของผลการเชื่อมโยงข้อมูลแบบความน่าจะเป็นเทียบกับการเชื่อมโยงแบบกำหนดตายตัว

		ผลของการเชื่อมโยงข้อมูลแบบกำหนดตายตัว		
		สอดคล้อง	ไม่สอดคล้อง	รวม
ผลของการเชื่อมโยงข้อมูลแบบความน่าจะเป็น	สอดคล้อง	A	B	A+B
	ไม่สอดคล้อง	C	D	C+D
	รวม	A+C	B+D	A+B+C+D

ค่าร้อยละความสอดคล้อง = (A+D)/(A+B+C+D); ความไว = A/(A+C); ความจำเพาะ = D/(B+D); PPV = A/(A+B); NPV = D/(C+D)

แบบกำหนดตายตัว) โดยเลือกใช้ตัวแปรในโมเดลที่แตกต่างกันจำนวน 17 โมเดล ให้ค่าร้อยละของความสอดคล้องดังแสดงในตารางที่ 3 โดยพบว่ามีจำนวน 12 โมเดลที่ให้ค่าความสอดคล้องมากกว่าหรือเท่ากับการเชื่อมโยงแบบกำหนดตายตัว โดยมีค่าความสอดคล้องสูงสุดที่ 7,247 คน และเมื่อทำการตรวจสอบข้อมูลซ้ำอีกครั้งในทุกระเบียนพบระเบียนหรือผู้ป่วยที่เป็นบุคคลคนเดียวกันเพิ่มขึ้นอีกจำนวน 3 คน

ตารางที่ 3. ร้อยละของความสอดคล้องของโมเดลที่ใช้การเชื่อมโยงแบบเชิงความน่าจะเป็น

โมเดลที่	ตัวแปรที่ใช้ในโมเดล[1]	ผู้ป่วยในการเชื่อมโยง (คน)	ความสอดคล้องที่พบ (คน)	ร้อยละที่สอดคล้อง
1	a, c, d, e, f, h, i, j	19,725	7,241	36.71
2	b, c, d, e, f, h, i, j	19,725	7,239	36.70
3*	b, c, d, e, f, h	19,725	7,245	36.73
4*	b, c, e, f, h	19,725	7,244	36.72
5*	b, c, e, g, h	19,725	7,246	36.74
6*	b, d, e, f	19,725	7,244	36.72
7	b, d, e, g	19,725	6,868	34.82
8*	b, h, i, j	19,725	7,247	36.74
9*	b, f, j	19,725	7,247	36.74
10*	b, e, j	19,725	7,247	36.74
11*	a, e, f, h, i, j	19,725	7,252	36.77
12	b, e, f, h, i, j	19,725	7,162	36.31
13*	b, e, f, h	19,725	7,244	36.72
14*	b, e, f	19,725	7,247	36.74
15*	a, e, g	19,725	7,247	36.74
16*	b, e, f	19,725	7,247	36.74
17	b, e, g	19,725	6,897	34.97

1: a = หมายเลขประจำตัว, b= หมายเลขโรงพยาบาล, c = ชื่อ, d = นามสกุล, e = ที่อยู่ที่ติดต่อได้, f = วันเกิด, g = วันที่เสียชีวิต, h = จังหวัด, i = เพศ, j = รหัสการวินิจฉัยโรค (ICD10)

*โมเดลที่มีผลความสอดคล้องมากกว่าหรือเท่ากับการเชื่อมโยงข้อมูลแบบกำหนดตายตัวแบบ 1 ต่อ 1 ซึ่งมีความสอดคล้องที่ร้อยละ 36.72 ค่า cut-off ของโมเดลที่ 1-17 คือ 21, 29, 17, 15, 15, 15, 27, 14, 10, 9, 17, 25, 15, 10, 10, 10 และ 20 ตามลำดับ

ตารางที่ 4. ค่าพยากรณ์ผลบวก ค่าพยากรณ์ผลลบ ค่าความไว และค่าความจำเพาะจากโมเดลที่ใช้การเชื่อมโยงข้อมูลแบบความน่าจะเป็น[1]

โมเดล	ความไว	95% CI	ความจำเพาะ	95%CI	PPV	95% CI	NPV	95% CI
1	99.92	99.88- 99.96	99.97	99.94- 99.99	99.94	99.91- 99.98	99.95	99.92- 99.98
2	99.86	99.81- 99.91	99.95	99.92- 99.98	99.92	99.88- 99.96	99.92	99.88- 99.96
3	99.94	99.91- 99.98	99.95	99.92- 99.98	99.92	99.88- 99.96	99.97	99.94- 99.91
4	99.93	99.89- 99.97	99.95	99.92- 99.98	99.92	99.88- 99.96	99.96	99.93- 99.99
5	99.96	99.93- 99.99	99.95	99.92- 99.98	99.92	99.88- 99.96	99.98	99.95-100
6	99.93	99.89- 99.97	99.95	99.92- 99.98	99.92	99.88- 99.96	99.96	99.93- 99.99
7	94.74	94.43- 95.05	99.95	99.92- 99.98	99.91	99.87- 99.95	97.04	96.80- 97.27
8	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
9	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
10	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
11	99.96	99.93-99.99	99.90	99.86-99.95	99.83	99.78-99.89	99.98	99.95-100
12	98.80	98.65-98.95	99.95	99.92-99.98	99.92	99.88-99.96	99.31	99.19-99.42
13	99.93	99.89-99.97	99.95	99.92-99.98	99.92	99.88-99.96	99.96	99.93-99.99
14	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
15	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
16	99.97	99.95-100	99.95	99.92-99.98	99.92	99.88-99.96	99.98	99.97-100
17	95.14	94.84-95.44	99.95	99.92-99.98	99.91	99.87-99.95	97.26	97.03-97.48

1: โมเดลที่มีผลความสอดคล้องมากกว่าหรือเท่ากับการเชื่อมโยงข้อมูลแบบกำหนดตายตัวแบบ 1 ต่อ 1 ซึ่งมีความสอดคล้องที่ร้อยละ 36.72 (95%CI: 36.05-37.39)

ความถูกต้องของการเชื่อมโยง

เมื่อเปรียบเทียบผลการเชื่อมโยงข้อมูลที่ได้ทั้งหมดกับผลการเชื่อมโยงข้อมูลด้วยวิธีการเชื่อมโยงแบบกำหนดตายตัวแบบ 1 ต่อ 1 พบว่า ทุกโมเดลที่เชื่อมโยงแบบความน่าจะเป็นให้ค่า PPVs ร้อยละ 99.83-99.94, NPV ร้อยละ 97.04-99.98, ค่าความไวร้อยละ 94.74-99.97 และค่าความจำเพาะร้อยละ 99.90-99.97 ดังแสดงในตารางที่ 4

เมื่อนำข้อมูลส่วนที่ไม่ตรง (unmatched) หรือไม่เป็นระเบียนเดียวกันจากการเชื่อมโยงแบบกำหนดตายตัวมาเชื่อมโยงแบบความน่าจะเป็นด้วยโมเดลที่ทดสอบ พบผู้ป่วยที่เป็นบุคคลเดียวกันเพิ่มขึ้นจากที่การเชื่อมโยงแบบกำหนดตายตัวไม่พบ และพบว่าทุกโมเดลที่ทดสอบสามารถระบุผู้ป่วยที่ตรงกันเพิ่มได้จำนวน 2-5 ราย ผู้วิจัยตรวจสอบข้อมูลซ้ำอีกครั้งและพบว่าทุกระเบียนในทุกโมเดลที่ทดสอบมีความถูกต้อง ยกเว้นโมเดลที่ 11 กล่าวคือ พบข้อมูลที่ตรงกันเพิ่มขึ้นถึง 5 ราย แต่ตรวจพบตรงกันจริงเพียง 1 ราย ดังนั้นโมเดลที่ 11 มีโอกาสเกิดผลบวกปลอมหรือผลลวง (false positive rate) เท่ากับ 0.001 (1-0.999) ดังแสดงในตารางที่ 5

การอภิปรายและสรุปผล

ผลการศึกษาความถูกต้องของกระบวนการเชื่อมโยงข้อมูลแบบเชิงความน่าจะเป็นระหว่างทะเบียนมะเร็งไทยกับฐานข้อมูลโรงพยาบาลตติยภูมิแห่งหนึ่งด้วยหลักการความน่าจะเป็น พบว่า ทั้ง 17 โมเดลที่ทดสอบให้ค่า PPV ร้อยละ 99.83-99.94, ค่า NPV ร้อยละ 97.04-99.98, ความไวร้อยละ 94.74-99.97 และค่าความจำเพาะร้อยละ 99.90-99.97 จึงสรุปได้ว่า การเชื่อมโยงข้อมูลจาก 2 ฐานข้อมูลด้วยหลักการความน่าจะเป็นทำให้ระบุได้ว่าข้อมูลเป็นของบุคคลเดียวกันในกรณีที่ไม่มีตัวระบุที่เฉพาะเจาะจงหรือไม่สามารถเข้าถึงได้ตัวแปรดังกล่าวได้ เช่น ไม่ทราบหมายเลขบัตรประจำตัวประชาชน 13 หลัก จึงใช้ตัวแปรอื่นร่วมกันในการทำนาย หรือระบุตัวบุคคล นอกจากนี้ ยังพบว่า การเชื่อมโยงข้อมูลด้วยหลักการความน่าจะเป็นทำ

ตารางที่ 5. ร้อยละของความสอดคล้องของโมเดลแบบความน่าจะเป็นในส่วนของข้อมูลจากที่การเชื่อมโยงแบบกำหนดตายตัวไม่พบ

โมเดลที่	จำนวนผู้ป่วยที่เชื่อมโยง	ความสอดคล้องที่พบ (คน)	ร้อยละของความสอดคล้อง
1	12,482	2	0.03
2	12,482	3	0.05
3	12,482	3	0.05
4	12,482	3	0.05
5	12,482	3	0.05
6	12,482	3	0.05
7	12,482	3	0.05
8	12,482	3	0.05
9	12,482	3	0.05
10	12,482	3	0.05
11	12,482	5	0.08
12	12,482	3	0.05
13	12,482	3	0.05
14	12,482	3	0.05
15	12,482	3	0.05
16	12,482	3	0.05
17	12,482	3	0.05

ให้ได้จำนวนบุคคลที่ตรงกันหรือเป็นบุคคลเดียวกันจริง ๆ ที่มากเพิ่มขึ้นร้อยละ 0.05 จากที่การเชื่อมโยงแบบกำหนดตายตัวตรวจไม่พบ งานวิจัยนี้มีจำนวนข้อมูลเพียงหลักหมื่นคน หากข้อมูลเป็นหลักสิบล้าน ก็อาจระบุตัวตนหรือพบเพิ่มขึ้นได้ถึงหลักพัน ซึ่งช่วยลดความผิดพลาดที่เกิดจากการเชื่อมโยงแบบกำหนดตายตัว เพราะการเชื่อมโยงแบบกำหนดตายตัวมักใช้รหัสประจำตัวในการเชื่อมโยง หากมีการกรอกตัวเลขรหัสประจำตัวผิดพลาด (ขาดเกินหรือสลับตำแหน่ง) จะทำให้ความถูกต้องของการเชื่อมโยงข้อมูลลดลง ในขณะที่การเชื่อมโยงแบบความน่าจะเป็นนั้นใช้ตัวแปรอื่น ๆ หลายตัวร่วมกันในการทำนายว่าเป็นบุคคลหรือระเบียนเดียวกันหรือไม่ ย่อมทำให้พบจำนวนบุคคลที่ตรงกันมากขึ้น ทั้งนี้การเชื่อมโยงแบบความน่าจะเป็นนั้นอาจให้ผลถูกต้องลดลงหากตัวแปรที่ใส่ในโมเดลนั้นมีความถูกต้องน้อยดังโมเดลที่ 7 ดังนั้นผู้วิจัยจึงต้องเลือกตัวแปรด้วยความรอบคอบ

งานวิจัยนี้มีข้อจำกัดคือไม่ได้วิเคราะห์ความไว (sensitivity analysis) ของค่าคะแนนจุดตัดของแต่ละโมเดลและของค่าคะแนนของแต่ละตัวแปรที่ใช้ในโมเดล อย่างไรก็ตาม จากการทดลองเบื้องต้นพบว่า หากเปลี่ยนค่าคะแนนจุดตัดต่างกันไป จะทำให้ผู้ป่วย 1 คนเจอคู่ที่ตรงกันมากกว่าเดิม บางครั้งอาจเจอถึง 30 คู่ เนื่องจากเป็นผลการวิเคราะห์ข้อมูลจากโมเดลการเชื่อมโยงแบบความน่าจะเป็นที่แสดงผลด้วยความน่าจะเป็น ผู้วิจัยจึงดำเนินการตรวจสอบข้อมูลซ้ำอีกครั้งทุกระเบียนในทุกโมเดลที่ทดสอบ นอกจากนี้ในการเชื่อมโยงข้อมูลแบบความน่าจะเป็น ผู้วิจัยเลือกใช้คำสั่งการจับคู่ที่ดีที่สุด (best matched) จึงพิจารณาไม่เปลี่ยนแปลงค่าจุดตัดของโมเดล ซึ่งสอดคล้องกับการศึกษาของ Capuani และคณะ (12) ที่ใช้ค่าจุดตัด 7.06 เพียงค่าเดียวเท่านั้น แต่ยังให้ค่าความไวสูงถึงร้อยละ 94.4 (95% CI: 90.0 – 97.0) และค่าความจำเพาะร้อยละ 100.0 (95%CI: 98.0 – 100.0) นอกจากนี้งานวิจัยนี้ไม่ได้เน้นการพัฒนาฐานข้อมูล แต่มุ่งพัฒนาวิธีการที่ใช้สนับสนุนการตรวจสอบความถูกต้องของข้อมูลที่ต้องแลกเปลี่ยนระหว่างหน่วยงานต่าง ๆ เช่น โรงพยาบาลหรือหน่วยงานที่เก็บข้อมูลด้านสุขภาพเพื่อให้สามารถใช้ประมวลผลเกี่ยวกับสถานการณ์ของโรคและการให้บริการทางการแพทย์ในระดับประเทศ

ปัญหาและอุปสรรคอีกประการหนึ่งที่ทำให้ข้อมูลจากทั้งสองแหล่งตรงกันน้อย คือ ในระยะแรกโรงพยาบาลกรอกข้อมูลเข้าระบบของทะเบียนมะเร็งไทยด้วยมือ เนื่องจากยังไม่ได้มีระบบฐานข้อมูลโรงพยาบาลเป็นการเฉพาะ ต่อมาจึงเริ่มใช้ระบบฐานข้อมูล ดังนั้นโรงพยาบาลที่ใช้ระบบเวชระเบียนอิเลกทรอนิกส์ ควรกำหนดการดึงข้อมูลที่ถูกต้องพร้อมกับตรวจสอบความถูกต้องของกระบวนการดังกล่าว เพื่อส่งต่อข้อมูลที่ดึงได้ไปยังฐานข้อมูลทะเบียนมะเร็งไทยได้อย่างรวดเร็วตามนโยบายของรัฐ และสนับสนุนระบบการสั่งการรักษาด้วยคอมพิวเตอร์ ซึ่งจะทำให้สามารถสกัด นำข้อมูลไปใช้ได้สะดวกขึ้น นอกจากนี้ยังต้องมีการปรับปรุงชุดคำสั่งในการเชื่อมโยงข้อมูลให้มีความเป็นปัจจุบันเมื่อมีความต้องการข้อมูลเพิ่มขึ้น แม้ว่าการแลกเปลี่ยนข้อมูลทางสุขภาพ (health information exchange) อย่างเช่นการจัดทำทะเบียนมะเร็งไทยไม่พบอุปสรรคจากการเชื่อมโยงข้อมูลด้วยเลขประจำตัวประชาชน 13 หลัก ตามพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562 (13) เพราะมีหน่วยงานรับผิดชอบโดยตรง และเป็นการส่งต่อข้อมูลระหว่างโรงพยาบาลในกระทรวงสาธารณสุขไปยังสถาบันมะเร็งแห่งชาติ กรมการแพทย์ กระทรวงสาธารณสุข ที่มีระบบรักษาความปลอดภัยของข้อมูลดังกล่าว แต่การเชื่อมโยงข้อมูลแบบความน่าจะเป็นเป็นวิธีเสริมซึ่งช่วยลดปริมาณข้อมูลที่หายไป และปรับปรุงการจัดหมวดหมู่ภายในตัวแปรที่เชื่อมโยงกัน สามารถเพิ่มจำนวนผู้ป่วย เพิ่มความละเอียดของข้อมูล และเพิ่มคุณภาพอื่น ๆ

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณท่านผู้บริหารโรงพยาบาลและบุคลากรโรงพยาบาลทุกท่านที่ให้ความอนุเคราะห์ในการเก็บข้อมูลครั้งนี้

เอกสารอ้างอิง

1. Maruekhatat R. Data linking techniques and privacy protection. Journal of King Mongkut's University of Technology North Bangkok 2007;17: 80-5.
2. WHO. Cancer fact sheet [online]. 2021 [cited April 22, 2022]. Available from: www.who.int/en/news-room/fact-sheets/detail/cancer.
3. Ministry of Public Health. Cancer in Thailand [online]. 2021 [cited April 22, 2022]. Available from: www.nci.go.th/e_book/cit_x/index.html.
4. Ministry of Public Health. Definition of key performance indicator for service plan of cancer 2017-2022 [online]. 2017 [cited April 22, 2022]. Available from: tcb.nci.go.th/CWEB/files/ServicePlan61.pdf.
5. National Cancer Institute. Cancer registry manual in Thailand. Bangkok: Information Technology Division, National Cancer Institute; 2015.
6. Merriel SWD, Turner EL, Walsh E, Young G, Metcalfe C, Hounsome L, et al. Validation of the National Cancer Registration and analysis service prostate cancer registry with data from the CAP study. Lancet 2016; 388: S77. DOI: https://doi.org/10.1016/S0140-6736(16)32313-3.
7. Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom. Epidemiology 2018; 29: 308-13. DOI: 10.1097/EDE.0000000000000786
8. Creighton N, Walton R, Roder D, Aranda S, Currow D. Validation of administrative hospital data for identifying incident pancreatic and periampullary cancer cases: a population-based study using linked cancer registry and administrative hospital data in New South Wales, Australia. BMJ open. 2016; 6: e011161. DOI: 10.1136/bmjopen-2016-011161.
9. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. AHRQ methods for effective health care. Rockville, Maryland: Agency for Health care Research and Quality; 2014.
10. Kranker K. dtalink: Faster probabilistic deduplication methods in Stata for record linking and large data files [online]. 2018 [cited April 22, 2022]. Available from: www.stata.com/meeting/columbus18/slides/columbus18_Kranker.pdf.

11. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002; 31: 1246-52.
12. Capuani L, Bierrenbach AL, Abreu F, Takecian PL, Ferreira JE, Sabino EC. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. Cadernos de saude publica. 2014; 30: 1623-32. DOI: 10.1590/0102-311x00024914.
13. Personal Data Protection Act B.E. 2562. Royal Gazette No. 136, Part 69A (May 27, 2019).