The use of data mining to investigate abnormal behavior of outpatient drug reimbursement under civil servant medical benefit scheme

Praprai Utama, Eakasit Pacharawongsakda

Big Data Engineering, College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand

Abstract

The improper insurance fund reimbursement occurring from uncommon unappropriated aspects for person entitled to reimbursement is a significant problem of health system which loses a huge amount of money to provide care for this group. At the presents, although claim system can be verified by The Comptroller general's Department (CGD), patient's medical records especially abnormal behavior reports from hospitals are reviewed by CGD's experts with considerably difficulties and taking a long time to check. The objective of this study is to develop patterns of abnormal behavior of outpatient drug reimbursement investigations using Data Mining. 21 million records of outpatient database under Civil Servant Medical Benefit Scheme (CSMBS)'s Direct Billing System in fiscal year 2013 were analyzed by Unsupervised Learning technique. In this reimbursement database, consisting of 4 sections such as outpatient diagnostic (OPDATA), expenses category (OPBILL), dispensing (OPDISP) and medical record (OPDTIEM) which are collected and examined by 2 techniques to compare such as RFM Segmentation Model and Partitional Clustering with K-means, K-medians. The result illustrated that there were 2.6 million patients attended the service with 21 million times of services frequency, and 32 hospitals in total with 710 times are accessed the health services. Additionally, approximately 5.3 hundred thousand people of outpatient reimbursement analyzed by RFM score (555555) showed that there were 112 patients, amounting to 0.02%. As a result of Partitional Cluster, the vast majority of reimbursement fall into one group. To investigate uncommon behavior of reimbursement, this study explored the unusual patterns more efficiently.

Keywords: RFM, K-means, K-medians, unsupervised, euclidean distance.

Received 31 August 2017; Accepted 30 November 2017

Correspondence: Praprai Utama, College of innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand (Tel.: +66-2-954-7300; E-mail address: praprai@hisro.or.th)

กระบวนการตาต้าไมนิ่งเพื่อศึกษาพฤติกรรมผิดปกติ การเบิกจ่ายค่ายานำกลับไปใช้ที่บ้านกรณีผู้ป่วยนอกสิทธิสวัสดิการข้าราชการระบบเบิกจ่ายตรง

ประไพร อุตมา, เอกสิทธิ์ พัชรวงศ์ศักดา
สาขาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ กรุงเทพมหานคร

บทคัดย่อ

การเบิกเคลมเงินประภัยที่ไม่เหมาะสมให้กับผู้มีสิทธิ์เบิกค่ารักษาเนื่องจากการความผิดปกติหรือการใช้ไปในทางที่ผิด ปัญหานี้ได้กลายเป็นปัญหาใหญ่ที่สำคัญของระบบสุขภาพทำให้สูญเสียเงินจำนวนมากในการดูแลกลุ่มคนดังกล่าว ปัจจุบันการตรวจสอบการเคลมสามารถทำได้จากการแจ้งพฤติกรรมผิดปกติของคนดังกล่าวจากโรงพยาบาลแล้วทางกรมบัญชีกลางได้ตรวจสอบจากการอ่านเวชระเบียนคนไข้โดยผู้เชี่ยวชาญ ซึ่งมีความยากลำบาก และใช้เวลาในการตรวจสอบค่อนข้างนาน งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อพัฒนารูปแบบการตรวจสอบความผิดปรกติของการเบิกจ่ายค่ายาผู้ป่วยนอกที่เบิกมาใช้ต่อที่บ้าน โดยการใช้ Data Mining ตรวจสอบวิเคราะห์ข้อมูลด้วยเทคนิค Unsupervised Learning จากฐานข้อมูลผู้ป่วยนอกสิทธิข้าราชการระบบเบิกจ่ายตรงปีงบประมาณ 2556 จำนวน 21 ล้าน records ประกอบด้วย 1) แฟ้มข้อมูลวินิจฉัยโรคของผู้ป่วยนอก (OPDATA) 2) แฟ้มค่าใช้จ่ายรายหมวด (OPBILL) 3) แฟ้มการเบิกจ่ายค่ายา (OPDISP) 4) แฟ้มรายการยา (OPDTIEM) รวบรวมจากการเบิกจ่ายกรมบัญชีกลาง การทดสอบประกอบด้วยเทคนิค 2 เทคนิค เพื่อทำการเปรียบเทียบ คือโมเดล RFM Segmentation และเทคนิค Partitional Clustering ด้วย K-means, K-medians ผลลัพท์ที่ได้พบว่าผู้ป่วยที่มารับบริการจำนวน 2.6 ล้านคน ความถี่ในการมารับบริการจำนวน 21 ล้านครั้ง มาใช้บริการสถานพยาบาลสูงสุด 32 แห่ง จำนวน 710 ครั้ง และจำนวนที่มีการเบิกเคลมค่ายากลับบ้านจำนวน 5.3 แสนคน ผลจาก RFM score (555555) มีจำนวน 112 คน 0.02% ผลจากเทคนิค Partitional Cluster ค่าส่วนใหญ่ตกอยู่ในหนึ่งกลุ่ม เพื่อให้การตรวจสอบพฤติกรรมที่ผิดปกติงานวิจัยชิ้นนี้สามารถช่วยค้นหารูปแบบความผิดปกติที่เกิดขึ้นได้อย่างมีประสิทธิภาพมากขึ้น

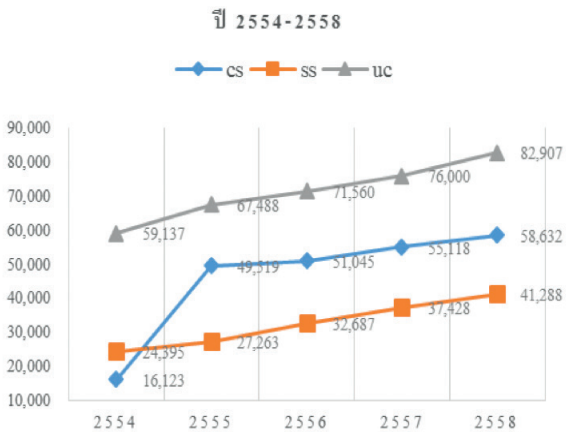
คำสำคัญ: RFM, K-means, K-medians, Unsupervised, Euclidean distance.

วันที่รับต้นฉบับ 31 สิงหาคม 2560; วันที่ตอบรับ 30 พฤศจิกายน 2560

บทนำ

ในปัจจุบันประเทศไทยได้มีการเบิกจ่ายค่ารักษาพยาบาลในการดูแลสุขภาพเป็นจำนวนเพิ่มขึ้นอย่างต่อเนื่อง จากข้อมูลค่ารักษาพยาบาลของทุกสิทธิ์การรักษา จาก**ภาพที่ 1** ระบบหลักประกันสุขภาพถ้วนหน้า (UC) มีค่าใช้จ่าย 82,907 ล้านบาท ซึ่งดูแลผู้มีสิทธิจำนวน 49 ล้านคน กองทุนสวัสดิการรักษาพยาบาลข้าราชการ (CSMBS) มีค่าใช้จ่าย 58,632 ล้านบาท ครอบคลุมบิดา, มารดา, บุตรและข้าราชการบำนาญดูแลผู้มีสิทธิ์ จำนวน 5 ล้านคน และกองทุนหลักประกันสุขภาพแห่งชาติ (SSO) ดูแลผู้มีสิทธิ์จำนวน 11 ล้านคน มีค่าใช้จ่ายในการรักษา พยาบาลผู้ประกันตนทั้งสิ้น 41,288 ล้านบาท

ผู้นิพนธ์ประสานงาน: ประไพร อุตมา, สาขาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ กรุงเทพมหานคร (โทร. +66-2-954-7300; E-mail address: praprai@hisro.or.th)



ภาพที่ 1 ค่าใช้จ่ายในการรักษาพยาบาล
ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

จากการเบิกจ่ายรายคนของสิทธิข้าราชการ ซึ่งน่าสนใจในการนำมาวิเคราะห์ ลงไปถึงการเบิกจ่ายค่าพยาบาลทั้ง 16 หมวด รวมเป็นเงินจำนวน 34 หมื่นล้านบาท ในระบบเบิกจ่ายตรงปีงบประมาณ 2556 ค่าใช้จ่ายสูงสุดอยู่ที่หมวดค่ายาที่ใช้ในสถานพยาบาล คิดเป็นร้อยละ 46 และยากลับไปใช้ต่อที่บ้าน มีการเบิกจ่ายคิดเป็นร้อยละ 24 ของค่าใช้จ่ายทั้งหมด ซึ่งสาเหตุอาจจะมีการเบิกจ่ายยาในปริมาณเกินความจำเป็นหรือไม่เหมาะสม จากข้อมูลที่กล่าวมาข้างต้นเพื่อศึกษาหมวดยาที่นำกลับไปใช้ต่อที่บ้าน พบความผิดปกติของการมารับบริการหลายสถานพยาบาล จำนวนสูงสุด 32 แห่ง และจำนวนครั้งที่ไปรับบริการสูงสุด 710 ครั้ง เพื่อตรวจสอบความผิดปกติดังกล่าวต้องอาศัยผู้เชียวชาญจากการเบิกเคลม และไม่สามารถเห็นพฤติกรรมภาพรวมของผู้เบิกเคลม[1] เดิมวิธีการตรวจเคลมนี้ใช้เวลาในการตรวจสอบนานและไม่มีประสิทธิภาพ หลายประเทศใช้วิธีการนี้[2] อย่างไรก็ตามการตรวจสอบการเคลมของประเทศไทย ทำได้ยากและใช้เวลาในการตรวจสอบนานเช่นกัน ดังนั้นการนำเหมืองข้อมูล Data mining มาช่วยในการตรวจสอบวิเคราะห์ด้วยเทคนิค (unsupervised learning) จะสามารถช่วยค้นหาความผิดปกติในข้อมูลได้ งานวิจัยนี้จึงเกิดขึ้น เพื่อช่วยให้สามารถตรวจสอบรูปแบบผิดปกตินี้ได้มีประสิทธิภาพและรวดเร็วมากขึ้น การศึกษานี้มีวัตถุประสงค์เพื่อค้นหาพฤติกรรมที่ผิดปกติของการไปรับบริการที่สถานพยาบาล โดยการประยุกต์ใช้เทคนิคการแบ่งกลุ่ม

ระเบียบวิธีวิจัย

ศึกษาทฤษฎีและผลการวิจัยที่เกี่ยวข้อง โดยใช้เทคนิคดาต้าไมนิ่ง (Data mining) เพื่อค้นหารูปแบบ (pattern) ของผู้ป่วยนอกที่มารับบริการที่สถานพยาบาลและรับยากลับบ้านที่อาจจะมีพฤติกรรมผิดปกติที่ซ่อนอยู่ กรณีไปหลายสถานพยาบาลในเวลาใกล้เคียงกัน ด้วยเทคนิค RFM Analysis ,K-means, K-medians ในการตรวจสอบเปรียบเทียบกัน 3 เทคนิค เพื่อให้มั่นใจว่าผู้ป่วยที่มีพฤติกรรมผิดปกติจริงตามสมมุติฐาน

Serhat Peker, Altan Kocyigit และ Erhan Eren[3] ได้ศึกษาเรื่อง LRFMP model for customer segmentation in the grocery retail industry: a case study โดยพัฒนาโมเดล RFM ในรูปแบบใหม่เรียกว่า LRFMP ประกอบด้วย leagth, recency, frequency, monetary, periodicity มาประยุกต์ใช้ในการจัดกลุ่ม ลูกค้าวงการค้าปลีกร้านขายของชำในประเทศตุรกี การแบ่งกลุ่มลูกค้าใช้เทคนิค data mining โดยใช้ K-means ผลการวิจัยสามารถแบ่งกลุ่มลูกค้าได้ 5 กลุ่ม คือ ลูกค้าประจำแฟนพันธ์แท้, ลูกค้าประจำ, ลูกค้าที่ไม่แน่นอน, ลูกค้ารายได้สูง, ลูกค้ารายได้น้อย จากความแตกต่างในพฤติกรรมการซื้อสินค้าตามโมเดล LRFMP ทำให้กำหนดกลยุทธ์ทางการตลาดได้

Hsin-HungWu, Shih-Yen Lin และ Chih-Wei Liu[4] ได้ศึกษาเรื่อง Analyzing Patients' Values by Applying Cluster Analysis and LRFM Model in a Pediatric Dental Clinic in Taiwan โดยศึกษาโมเดล LRFM (leagth , recency, frequency, monetary) ในคลีนิคทันตกรรมเด็กของประเทศไต้หวัน ใช้เทคนิค SOM และใช้ K-means อัลกอลิทรึม ในการแบ่งกลุ่ม 12 กลุ่ม ด้วยค่าเฉลี่ยของ L, R, F, M และใช้เมทริกซ์พิจารณาความสัมพันธ์จากระยะเวลา และผู้ป่วยใหม่ ผลที่ได้คือสามารถแบ่งกลุ่มผู้ป่วยได้ 3 กลุ่ม คือ กลุ่มผู้ป่วยคนไข้ประจำแฟนพันธ์แท้ ได้จากค่า L,R,F มีค่ามากกว่าค่าเฉลี่ยรวมกลุ่มที่เป็นผู้ป่วยเก่า จากค่า L มีค่ามากๆ และอาจจะกลายเป็นผู้ป่วยประจำแฟนพันธ์แท้ได้ในอนาคต และกลุ่มผู้ป่วยที่หายไปเนื่องจากค่า L, R มีค่าต่ำ จากผลการวิเคราะห์ที่ได้สามารถนำมาวางแผนกลยุทธ์การตลาดเพื่อออกแบบความต้องการของผู้ป่วยที่แตกต่างกันไปและพัฒนาผู้ป่วยให้กลายเป็นลูกค้าประจำในอนาคต

Bijan Geraili, Mahdi Nasiri, Mohammad Arab[5] ได้ศึกษาเรื่อง Improving Fraud and Abuse Detection in General Physician Claim A Data Mining Study โดยศึกษาตัวชี้วัด 13 ตัว ของสิทธิการรักษาประกันสังคมมี สถานพยาบาลในเครือ 70 แห่ง คลินิก 280 แห่ง เพื่อวัดประสิทธิภาพในการตรวจสอบแพทย์ ที่ต้องสงสัยว่าทุจริต 98% และแพทย์จ่ายยาไม่เหมาะสม (abuse) 85% วิธีการวิจัยใช้การทำเหมืองข้อมูล เทคนิค supervised และ unsupervised แล้วนำไปสร้างแบบจำลอง ได้แบ่งเป็น 2 กลุ่ม 1) พฤติกรรมแพทย์จ่ายยาไม่เหมาะสมหรือเกินความจำเป็น วิธีนี้แพทย์อาจจะสร้างรายได้ทางอ้อม และแพทย์อาจจะให้ยาที่แพทย์ ได้รับผลประโยชน์จากหน่วยงาน ภายนอก 2) พฤติกรรมที่แพทย์ไม่เขียนรายการยาในใบสั่งยา แต่ให้คนไข้ไปซื้อยาจากร้านขายยาข้างนอก แล้วนำใบเสร็จไปเบิกกับหน่วยงานที่ต้นสังกัด วิธีการนี้อาจจะมีการเขียน ใบสั่งยาปลอม และนำไปซื้อร้านขายยาที่สมรู้รวมคิด และนำบิลไปเบิกจากหน่วยงานหรือบริษัทประกัน แล้วนำเงินที่ได้มาแบ่งกัน ผลการวิจัยพบว่า มีตัวชี้วัดที่มีความสัมพันธ์กับเรื่องค่าใช้จ่าย คือ ความถี่มารับบริการซึ่งผู้ป่วยไปรับบริการกับแพทย์มากกว่า 1 ครั้งในช่วงเวลาสั้น และแพทย์ ได้เขียนใบสั่งยาให้ไปซื้อที่ร้านขายยาข้างนอกจากบันทึกการรักษารายการยา ในใบสั่งยา 30% ให้ไปซื้อจากร้านยาข้างนอกสถานพยาบาล ทั้งนี้ แพทย์ 54% มีพฤติกรรมไม่เหมาะสม และ 2% สงสัยว่าทุจริต

Dallas Thornton และคณะ[6] ได้ศึกษาเรื่อง Outlier - based Health Insurance Fraud Detection for U.S. Medicaid

โดยใช้เหมืองข้อมูลในการตรวจสอบการทุจริตด้วยเทคนิค Unsupervised เพื่อพยากรณ์การทุจริตข้อมูลการเคลมทันตกรรม โดยใช้ข้อมูล 2012 - 2013 จำนวน 11 เดือน พบว่า คนไข้ไม่ใช่ทั้งหมดที่ทำทุจริต แต่ที่ต้องสงสัยคือสถานพยาบาลที่มีผู้ป่วยมารับบริการมากกว่า 300 คนต่อสัปดาห์ อาจต้องสงสัยว่าทุจริต เช่น มีการใส่ข้อมูลปลอมหรืออาจจะเป็นการให้บริการทันตกรรมเคลื่อนที่ วิธีการตรวจสอบการทุจริต จะหาผู้ให้บริการที่มีค่าผิดปกติจาก multiple predictors โดยการวิเคราะห์ตัวชี้วัด 14 ตัว วิเคราะห์ข้อมูลด้วยโปรแกรม R กรณีที่มีการเบิกเงินเคลม 10,000 ดอลล่าร์ต่อเดือน และการวิเคราะห์ด้วยเทคนิค K - means clustering เพื่อดูค่าที่ผิดปกติจากข้อมูลผู้ให้บริการจำนวน 500 แห่ง ผลการวิจัยพบว่าผู้ให้บริการ 71% มีพฤติกรรมที่น่าสงสัยและผิดปกติ จากหลักฐานทำให้สามารถเข้าตรวจสอบได้ตามกฎหมาย

Qi Liu และ Miklos Vasarhelyi[7] ได้ศึกษาเรื่อง Healthcare fraud detection A survey and a clustering model incorporating Geo-location information ของประเทศอเมริกาในปีงบประมาณ 2010 วัตถุประสงค์เพื่อตรวจสอบการทุจริตให้ประสิทธิภาพมากขึ้นและช่วยลดค่าใช้จ่ายด้านสุขภาพ โดยให้ความสนใจในการป้อนข้อมูลที่ผิดพลาดหรือไม่สมบูรณ์ในการเคลมเงินและการให้บริการทางการแพทย์ที่ไม่ซ้ำซ้อน วิธีการวิจัยใช้ข้อมูล SSO ได้แบ่งกลุ่มผู้มารับบริการเป็น 3 ประเภท 1) อายุมากกว่า 65 ปี 2) อายุน้อยกว่า 65 ปีและคนพิการ 3) ผู้ป่วยโรคไตวายเรื้อรังระยะสุดท้าย และแบ่งกลุ่มผู้ให้บริการเป็น 3 ประเภท 1) Hospital insurance 2) Medical insurance 3) prescription drug coverage (เฉพาะผู้มีรายได้น้อย) ได้แบ่งการทุจริตเป็น 4 ประเภท 1) การทุจริต ของผู้ให้บริการ 2) การทุจริตสมาชิกประกันภัย 3) การโกงผู้ให้บริการประกันภัย 4) สมรู้ร่วมคิดทุจริต การทุจริตที่เกี่ยวข้องกันมากกว่าหนึ่งฝ่าย การนำเทคนิคดาต้าไมนิ่งมาใช้ในการตรวจสอบทุจริตด้วยเทคนิค Supervised ซึ่งใช้ neural network ในการตรวจสอบผลที่ได้พบว่ามีการทุจริตเงินจำนวน 47.9 พันล้านดอลล่าห์ของค่าใช้จ่ายประกันสุขภาพภาครัฐอย่างไรก็ตามงานวิจัยนี้สามารถพัฒนาวิธีการตรวจสอบการทุจริตเพื่อป้องกันไม่ได้เกิดขึ้นได้ในอนาคต

RFM Model

แบบจำลอง RFM ผู้ค้นพบ Hughes (1996) เพื่อวิเคราะห์พฤติกรรมความต้องการของลูกค้า ซึ่งใช้ตัวแปร (R) วันล่าสุดที่มา (F) ความถี่ (M) จำนวนเงิน หากลูกค้ามีระยะเวลาสั้นๆ แสดงโอกาสกลับมาซื้อสินค้าสูง และความถี่ในการซื้อบ่งบอกได้ว่าเป็นลูกค้าประจำ และจำนวนเงินที่ใช้ไปเมื่อนำมาวิเคราะห์ต้องทำการ แบ่งข้อมูล R, F, M เป็น 5 กลุ่ม มีจำนวนเท่าๆ กัน (quintile) 20% ซึ่งการจัดกลุ่มลูกค้าที่มีพฤติกรรมคล้ายๆ กันไว้ด้วยกันจะช่วยให้สามารถดูภาพรวมของลูกค้าได้ง่ายขึ้น เช่น กลุ่มลูกค้าที่มีการซื้อสินค้าของเรา เมื่อมาซื้อสินค้าของเราบ่อยๆ และมีการใช้จ่ายเยอะหรือกลุ่มลูกค้าที่ไม่ค่อยมาซื้อสินค้าของเราแต่เมื่อมาซื้อแต่ละครั้งจะซื้อในจำนวนที่มาก วิธีการจัดกลุ่มตามพฤติกรรมการซื้อสินค้าโดยดูจาก 1) ระยะเวลา (จำนวนวัน) จากการซื้อล่าสุดที่ผ่านมา 2) ความถี่ของการซื้อสินค้า 3) การใช้จ่ายของลูกค้า วิธีการนี้นำเสนอ โดย Fader ในปี 2005 เรียกแบบย่อว่า RFM ซึ่งวิธีการนี้จะแบ่งค่า Recency (R), Frequency (F) และ Monetary (M) ออกเป็น 5 ส่วนเท่าๆ กัน โดยเลขที่มีค่ามากที่สุด (คือเลข 5) จะมีความสำคัญที่สุดและใช้ค่าตัวเลข 3 หลักเป็นตัวแทน ของแต่ละกลุ่ม เช่น 555 คือ กลุ่มที่มีค่า R=5, F=5 และ M=5 หมายความว่าเป็นลูกค้าที่มักจะมาซื้อสินค้าบ่อยๆ และมีการใช้จ่ายที่สูง ต่อมามีการประยุกต์ใช้แบบจำลอง LRFM มีตัวแปร (L) ระยะเวลาที่เป็นลูกค้า โดย Reinartz and Kumar (2000) ซึ่งแสดงถึงเป็นลูกค้าที่มีความภักดี

LRFMP Model

โดยมีการนำ (P) มาใช้ในกับ LRFM เพื่ออธิบายลักษณะความสม่ำเสมอของลูกค้าเนื่องจากแบบจำลอง RFM เดิมไม่ได้สะท้อนถึงการซื้อตามความเป็นจริงของลูกค้า เป็นเพียงการทำธุรกรรมครั้งล่าสุดที่อาจจะทำให้เข้าใจผิดได้ การเพิ่มตัวแปร (P) เข้าไปในโมเดลเพื่อแก้ปัญหานี้ การคำนวณค่า Periodicity (P) แสดงลักษณะว่าเป็นลูกค้าประจำ โดยกำหนดระยะเวลาเป็นส่วนเบี่ยงเบนมาตรฐานของการมาซื้อสินค้า

$$\text{Periodicity} = \text{stdev}(IVT_1, IVT_2, \ldots, IVT_{n-1}, IVT_n) \quad (1)$$

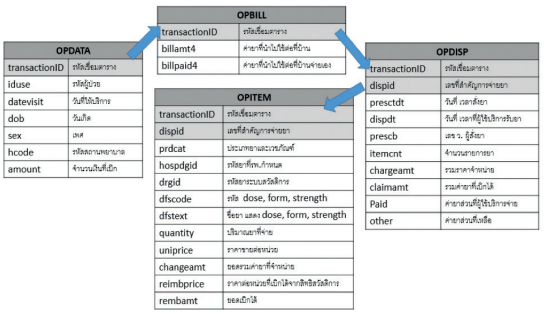
IVT หมายถึง ระยะเวลาห่างระหว่างวันที่เข้ามาซื้อสินค้าล่าสุดลบด้วยวันถัดมา ถ้าค่า IVT ต่ำแสดงว่าลูกค้าเข้ามาซื้อสินค้าเป็นประจำ

$$IVTi = \text{date\_diff}(t_{i+1}, t_i) \quad (2)$$

การเตรียมข้อมูล

ข้อมูลที่นำมาใช้เป็นการเบิกค่ายากลับไปใช้ที่บ้านของผู้ป่วยนอกรายบุคคล ประกอบด้วยแฟ้มข้อมูล 4 แฟ้ม คือ OPDATA, OPBILL, OPDISP และ OPITEM ซึ่งเป็น Normalized form โดยมี Primary key field เป็น transactionID ดังภาพที่ 3 เนื่องด้วยข้อมูลดิบมีทั้งหมด 2 ฐาน จากหน่วยงานกรมบัญชีกลาง ต้องรวมข้อมูลเข้าด้วยกันเพื่อให้ได้ตัวแปรครบพร้อม

ทำการวิจัยได้โดยสำนักสารสนเทศบริการสุขภาพเป็นหน่วยงานตรวจเคลมและนำข้อมูลเข้าระบบ GFMIS แต่ไม่ได้จ่ายจริง ส่วนการจ่ายเคลมอนุมัติโดยกรมบัญชีกลางหลังจากนั้นก็เข้าสู่กระบวนการตรวจสอบความถูกต้องของข้อมูลโดยตรวจสอบความซ้ำซ้อนของข้อมูล, ตรวจสอบจำนวนเงินเบิกจ่ายให้ตรงกับยอดเบิก, วันที่มารับบริการ (datevisit) และความถี่ของการมารับบริการ



ภาพที่ 3 ฐานข้อมูลผู้ป่วยนอกลักษณะ Normalized form

วิธีดำเนินการวิจัย

การคำนวนตัวแปร R (Recency) จากวันสุดท้ายที่มารับบริการสถาน พยาบาลโดยกำหนดวันสุดท้าย (30 กันยายน 2556) ลบวันที่มารับบริการครั้ง สุดท้าย (datevisit) จำนวนครั้งในการมารับบริการที่สถานพยาบาล (F), จำนวนสถานพยาบาลที่ไปรับบริการรายบุคคล (H) และค่าเฉลี่ยค่ายาที่เบิกรายคนต่อครั้ง (M) จากตัวแปร (billamt4) และทำการแบ่งกลุ่ม Cluster RFM, K-means และ K-medians เพื่อแบ่งกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน มีความคล้ายคลึงกันมากที่สุดและข้อมูลที่อยู่ต่างกลุ่มกันมีความแตกต่างกันมากที่สุดโดยแบ่งเป็น 5 กลุ่ม ด้วยโปรแกรม STATA

Partitional clustering

ก่อนทำการแบ่งกลุ่มต้อง Normalize ข้อมูลให้มีสเกลเท่ากันก่อน หลังจากนั้นก็การแบ่ง Cluster ด้วย K-means, K-medians ใช้วิธีการวัดระยะห่าง Euclidean distance ในการทำ RFM, RFMPHQ โมเดลตัวแปรที่นำมาแบ่ง Cluster ประกอบด้วยตัวแปรดังต่อไปนี้

- M (Monetary) ค่าเฉลี่ยการเบิกจ่ายค่ายารายครั้ง
- R (Recency) จำนวนวันจากการมารับบริการล่าสุด
- F (Frequency) ความถี่ของการมารับบริการ
- P (Periodicity) ส่วนเบี่ยงเบนมาตรฐานของระยะเวลาที่มาใช้บริการ
- H (HospVisit) จำนวนสถานพยาบาลที่ไปรับบริการ
- Q (Quantity) ปริมาณยาที่เบิกจ่าย
- Age อายุ

ผลการทดลอง

ค่าสถิติและความสัมพันธ์

จาก**ตารางที่ 1** พบว่า (Q) มีความสัมพันธ์กับค่าเฉลี่ยการเบิกจ่ายค่ายารายครั้ง (M) เท่ากับ 9% โดยที่ (H) มีความสัมพันธ์กับความถี่ (F) เท่ากับ 32% ซึ่งมีตัวแปรทั้งหมดมีความสัมพันธ์กันค่อนข้างน้อย

ตารางที่ 1 แสดงค่าสัมพันธ์ค่าสถิติ Correlation ระหว่างตัวแปร

	Q	M	R	F	P	H	Age
Q	1						
M	0.090	1					
R	-0.032	0.011	1				
F	-0.005	-0.037	-0.142	1			
P	0.049	0.025	-0.074	-0.301	1		
H	-0.028	-0.090	-0.140	0.322	-0.156	1	
Age	0.125	0.072	-0.108	0.143	-0.089	0.125	1

ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

จาก**ตารางที่ 2** ข้อมูลการมารับบริการที่สถานพยาบาลจำนวน 21 ล้านครั้ง จำนวนผู้มารับบริการ 2.6 ล้านคน ตัดข้อมูลผู้มารับบริการที่มีการเบิกจ่ายค่ายากลับบ้านเหลือ 5.3 แสนคน จำนวนเงินที่มีการเบิกจ่ายค่ายารายครั้ง (M) เฉลี่ยต่อคนมากสุด 340,906 บาทต่อครั้ง มีความถี่ (F) ในการมามากสุดที่ 710 ครั้ง ช่วงระยะเวลาที่กลับมารับบริการ (P) สูงสุด 255 วัน ถ้าค่าน้อยแสดงถึงผู้ป่วยมารับบริการเป็นประจำ จำนวนสถานพยาบาล ที่ไปใช้บริการมากสุด 32 แห่ง มีการเบิกค่ายากลับบ้านสูงสุด (Mdrug) 2.9 ล้านบาท ปริมาณยาที่ได้รับจำนวนมากสุด 5,700 หน่วย

ตารางที่ 2 แสดงหาค่าสถิติของตัวแปร

Stats	Mean	Median	Max	Min	SD
M	890	184	340,907	1	3,269
Mdrug	8,580	1,410	2,943,455	1	37,133
R	98	62	365	1	89
F	12	8	710	1	17
P	34	29	255	0	28
H	2	2	32	1	1
Q	67	30	5,700	1	99
Age	57	59	100	0	19

ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

RFMPHQ score

การจัดอันดับผู้มารับบริการด้วยการแบ่งกลุ่มตามโมเดล RFM Segmentation ค่า score ซึ่งแบ่งได้ 8,693 กลุ่ม ที่มีผู้มารับบริการสูงสุด (RFMPHQ = 112512) จำนวน 3,145 คน คิดเป็น 0.58% ความหมายคือผู้รับบริการมาไม่บ่อยแต่มีการเบิกจ่ายค่ายากลับบ้านจำนวนมากและไปแค่สถานพยาบาลเดียว) ค่า score (RFMPHQ = 555555) มีจำนวน 112 คน คิดเป็น 0.02% ความหมายคือ ผู้มารับบริการมาเป็นประจำ ,มีการเบิกจ่ายค่ายาจำนวนมาก ,ไปหลายสถานพยาบาล และปริมาณการเบิกยาสูง

การประยุกต์ใช้โมเดล LRFMP เป็นเทคนิคที่เกิดจาก Ha and Park (1998) ซึ่งใช้สัญญาลักษณ์ (↑) กับค่าเฉลี่ยของตัวแปร L, R, F, M, IVT หรือ P ที่มีค่ามากกว่าค่าเฉลี่ยรวม และจะใช้ (↓) ถ้าค่าที่ได้น้อยว่าค่าเฉลี่ยรวมจาก**ตารางที่ 3, 4** เป็นการประยุกต์ใช้ค่า (H) แทนค่า (L), (P) เพื่อดูสถานะผู้มารับบริการเป็นประจำ ค่าที่ได้จะต่ำและ (H) จำนวนสถานพยาบาลที่ไปรับบริการเข้าใน RFMPH

RFMPHQ Score ด้วยการแบ่งกลุ่ม K-means, K-medians

จาก**ตารางที่ 3** การคำนวณค่า RFMPHQ score ค่าที่มีลูกศรขึ้น (↑) จะปรากฏตัวแปรในค่า score ส่วนค่าที่มีลูกศร (↓) จะไม่ปรากฏตัวแปรใน Cluster จากการแบ่งกลุ่มด้วยเทคนิค K-means สามารถจำแนกได้ 4 ประเภท ประเภทแรก (FHQ) คิดเป็น 50% ใน Cluster 1 มีค่าเฉลี่ยความถี่มาสถานพยาบาล (F) 15 ครั้ง ค่าเฉลี่ยจำนวนสถานพยาบาลที่ไป (H) 1.93 โรงพยาบาล และปริมาณยาที่เบิกกลับบ้านเฉลี่ย (Q) จำนวนเฉลี่ย 69.7 หน่วย ประเภทที่ 2 (R) ใน Cluster 4,5 จำนวนวันเฉลี่ยที่ผู้ป่วยมารับบริการ ล่าสุด 299, 107 วัน คิดเป็น 9% และ 19% ประเภทที่ 3 (RM) ใน Cluster 3 คิดเป็น 16% มีการเบิกจ่ายค่ายามากกว่าค่าเฉลี่ย เท่ากับ 1,082 บาท ประเภทที่ 4 (P) ใน Cluster 2 คิดเป็น 7% ระยะเวลาเฉลี่ยที่มาใช้บริการประจำ 106 วัน

ตารางที่ 3 ค่าเฉลี่ยโมเดล RFMPHQ score ด้วย K-means

Cluster	N	R	F	M	P	H	Q	RFMPHQ	Size
C1	267,968	31.59	15.16	861.36	28.81	1.93	69.76	FHQ	50%
C2	35,102	65.97	3.47	820.51	106.98	1.47	63.57	P	7%
C3	87,961	189.04	8.79	1082.72	33.81	1.63	66.03	RM	16%
C4	48,387	299.61	8.94	855.30	22.65	1.60	56.25	R	9%
C5	100,501	107.07	10.91	838.72	30.01	1.79	66.52	R	19%
Total	539,919	97.55	12.01	890.01	34.38	1.80	66.93		100%

ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

จากตารางที่ 4 การคำนวณค่า RFMPHQ score แบ่งด้วย K-medians สามารถจำแนกได้ 5 ประเภท ประเภทแรก (FH) ใน Cluster 1 คิดเป็น 37% โดยมีค่า F สูงสุด 18 ครั้ง ค่าเฉลี่ยจำนวนสถานพยาบาลที่ไป (H) 2.05 แห่ง ประเภทที่ 2 (MPQ) ใน Cluster 2 คิดเป็น 18% โดยมี (M) เท่ากับ 1,010 บาท และ (Q) ปริมาณยาเฉลี่ยที่ได้รับ 75 หน่วย ประเภทที่ 3 (RMP) ใน Cluster 3 คิดเป็น 17% โดยมี (M) เฉลี่ยสูงสุด 1,064 บาท ประเภทที่ 4 (R) ใน Cluster 4 คิดเป็น 10% ประเภทที่ (RH) ใน Cluster 5 คิดเป็น 18% โดยมีค่า (H) เฉลี่ยเท่ากับ 1.83 แห่ง

ตารางที่ 4 ค่าเฉลี่ยโมเดล RFMPHQ score ด้วย K-medians

Cluster	N	R	F	M	P	H	Q	RFMPHQ	Size
C1	197,783	29.17	18.15	787.63	22.01	2.05	65.97	FH	37%
C2	98,618	42.81	5.37	1010.36	67.71	1.53	75.71	MPQ	18%
C3	92,327	179.48	8.51	1064.37	37.22	1.61	66.58	RMP	17%
C4	56,055	291.20	8.98	865.08	23.57	1.60	56.74	R	10%
C5	95,136	102.85	11.32	823.56	29.16	1.83	66.20	RH	18%
Total	539,919	97.55	12.01	890.01	34.38	1.80	66.93		100%

ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

จากค่าที่ได้**ตารางที่ 3, 4** เพื่อเลือกกลุ่มคนที่ลักษณะคล้ายกัน กลุ่มที่สนใจคือ (FHQ) และ (FH) ทำการ intersect ระหว่าง K-means = C1, K-medians= C1 พบกลุ่มคนที่มีลักษณะเดียวกัน 197,783 คน ใน**ตารางที่ 5** และเพื่อตรวจสอบว่ากลุ่มคนดังกล่าวอยู่ในกลุ่มเดียวกับโมเดล RFMPHQ จึงทำการ intersect อีกครั้ง

ตารางที่ 5 Intersect K-means, K-medians

	K-means					
K-medians	C1	C2	C3	C4	C5	Total
C1	197,783	64,865	0	0	5,320	267,968
C2	0	30,309	3,568	0	1,225	35,102
C3	0	0	80,292	7,668	1	87,961
C4	0	0	0	48,387	0	48,387
C5	0	3,444	8,467	0	88,590	100,501
Total	197,783	98,618	92,327	56,055	95,136	539,919

ที่มา: สำนักงานวิจัยเพื่อการพัฒนาหลักประกันสุขภาพไทย (สวปก)

ทำการ intersect กลุ่ม FHQ กับโมเดล RFMPHQ จากกลุ่ม FHQ=555 มีจำนวน 4,594 คน (F=5) คือ ความถี่ของการมารับบริการ อยู่ระหว่าง 17-710 ครั้ง (H=5) จำนวนสถานพยาบาลที่ไปรับบริการ 3-32 แห่ง และ (Q=5) ปริมาณยาที่เบิกจ่าย 101-5700 หน่วย

ตารางที่ 6 รูปแบบพฤติกรรมการมารับบริการ

กลุ่ม	คำอธิบาย
FHQ	มาบ่อย, ไปหลายรพ, ปริมาณยาจำนวนมาก
FM	มาบ่อย, เบิกเคลมจำนวนมาก
MPQ	เบิกเคลมจำนวนมาก, มาประจำ, ปริมาณยาจำนวนมาก
RMP	นานครั้งมา, เบิกเคลมจำนวนมาก

จาก**ตารางที่ 6** แสดงรูปแบบพฤติกรรมการมารับบริการที่ได้จากการวิเคราะห์โมเดล RFMPQH สามารถเห็นความสัมพันธ์ระหว่างตัวแปรและอ่านตีความหมายได้เป็นคู่ พบว่าความถี่ (F) มีความสัมพันธ์กับจำนวนเงินที่เบิก (M) กลุ่มที่มาบ่อยและเบิกค่ายาจำนวนมาก (F↑ M↑) กลุ่มที่นานครั้งมาแต่เบิกค่ายามาก (F↓M↑) กลุ่มที่นานครั้งมาและเบิกค่ายาน้อย (F↓M↓) และกลุ่ม (F↑M↓) มาบ่อยแต่เบิกค่ายาน้อย อีกในหนึ่งผู้มารับบริการมาล่าสุด (R) และความถี่ (F) กลุ่ม (R↑F↓) นานๆ มาครั้งความถี่ต่ำ และ (R↓F↓) มาใช้บริการเร็วๆ นี้ แต่มีความถี่ต่ำ กลุ่มที่ (R↓F↑) มาใช้บริการเมื่อเร็วๆ นี้และมีความถี่สูง ความสัมพันธ์อีกในหนึ่งระหว่างความถี่ (F) และจำนวนสถานพยาบาล ที่ไปรับบริการ (H) กลุ่ม (F↑ H ↑) มาใช้บริการบ่อยแล้วไปหลายโรงพยาบาล กลุ่มที่ไม่ค่อยมาโรงพยายาบาล (F↓ H ↓) และกลุ่มที่เบิกจ่ายเงินมากไปรักษาหลายแห่ง (M↑H ↑) กลุ่ม (M↓H ↓) ผู้ป่วยเบิกค่ายาน้อยและไปรักษาแค่แห่งเดียว และกลุ่มผู้ป่วยที่เบิกค่ายามากและไปแค่แห่งเดียว (M↑H↓)

วิจารณ์พลและสรุป

งานวิจัยชิ้นนี้นำเสนอผลการศึกษาพฤติกรรมการมารับบริการที่สถานพยาบาลโดยทำการเบิกจ่ายค่ายานำกลับไปใช้ต่อที่บ้านบ่อยครั้งเกินความจำเป็น โดยการวิจัยได้ประยุกต์ใช้โมเดล RFM กับการเคลมด้านสุขภาพ ซึ่งเดิมโมเดล RFM ใช้จัดกลุ่มลูกค้าที่มาซื้อสินค้า เพื่อนำไปกำหนดกลยุทธ์ทางการตลาด การนำโมเดล RFM มาประยุกต์ใช้ครั้งนี้ใช้ตัวแปร วันล่าสุดที่มารับบริการ, ความถี่, และการเบิกค่ายาเฉลี่ยต่อครั้ง อาจพบรูปแบบการมารับบริการจากค่า RFM score แต่เพื่อให้ชัดเจนมากขึ้น ได้นำเทคนิค K-means, K-medians แบ่งกลุ่ม 5 กลุ่ม (k=5) เพื่อให้สามารถเปรียบเทียบกับโมเดล RFM 5 กลุ่มได้ พบกลุ่มที่น่าสนใจคือ FHQ คือ กลุ่มที่ไปรับบริการบ่อย, ไปหลายสถานพยาบาล และปริมาณยาที่ได้รับกลับบ้านจำนวนมาก เพื่อให้ทราบว่าเป็นกลุ่มคนเดียวกัน ทำการ intersect กลุ่ม K-means, K-medians ได้กลุ่มคนที่มีลักษณะเดียวกัน 197,783 คน อย่างไรก็ตามกลุ่มดังกล่าวยังกว้าง เมื่อเทียบกับการทำด้วยโมเดล RFM ที่สามารถชี้กลุ่มได้ชัดเจนกว่า ดังนั้นผู้วิจัยจึงทำการ intersect อีกครั้งกับโมเดล RFM ทำให้ได้ค่า FHQ score = 555 จำนวน 4,594 คน สามารถแยกกลุ่มที่มีลักษณะเดียวกันได้ชัดเจนมากขึ้น และเพื่อเพิ่มประสิทธิภาพในการศึกษารูปแบบที่ชัดเจนขึ้น การประยุกต์ใช้โมเดล RFMPHQ ทำการวิเคราะห์และอ่านตีความค่าเป็นคู่ สามารถเห็นรูปแบบลักษณะของผู้มารับบริการได้ชัดเจนมากยิ่งขึ้น ที่กล่าวมาทั้งหมดนี้การใช้เทคนิคดาต้าไมนิ่งในการแบ่งกลุ่มผู้มารับบริการที่มีปริมาณข้อมูลจำนวนมาก สามารถทำให้เห็นรูปแบบที่ซ้อนอยู่ในข้อมูลได้ และตรวจสอบข้อมูลได้ง่ายขึ้น ใช้เวลาไม่มาก และลดค่าใช้จ่ายในการจ้างผู้เชียวชาญมาทำการตรวจสอบ

ข้อจำกัดในการวิจัยครั้งนี้คือ หน่วยของยาที่แตกต่างออกไป เช่น ยาเม็ด ยาขวด ยาแผง และอื่นๆ ซึ่งการนำมาใช้ในงานวิจัยครั้งนี้เป็นการรวมยอดจำนวนหน่วยของยา ถ้ามีการแบ่งหน่วยยาให้ชัดเจนกว่านี้ อาจจะทำให้การวิเคราะห์แม่นยำมากขึ้น แนวทางการพัฒนาในอนาคตได้มีการเพิ่มเทคนิคดาต้าไมนิ่ง เช่น เทคนิค DBSCAN Clustering เพื่อแบ่งกลุ่มข้อมูลที่แปลกๆ ให้อยู่ในกลุ่มเดียวกัน และสร้างโมเดล Decision tree ซึ่งต้องหาตัวอย่างกลุ่มคนที่มีพฤติกรรมผิดปกติเพื่อใช้ในการ

รันโมเดล เดิมทีโมเดล RFMPQH ก็ให้ผลที่ดีและแม่นยำแล้ว แต่ผู้วิจัยต้องการเปรียบเทียบหลายเทคนิค เพื่อค้นหากลุ่มผู้มีพฤติกรรมผิดปกติที่ซ่อนอยู่ในข้อมูลโดยใช้เวลาไม่มากและมีประสิทธิภาพในการตรวจสอบมากที่สุด

เอกสารอ้างอิง

1. กรมบัญชีกลาง, คู่มือสวัสดิการรักษาพยาบาลข้าราชการ เล่ม 1, 2551, หน้า 6-38
2. เอกสิทธิ์ พัชรวงศ์ศักดา, Introduction to Business Analytics with RapidMiner Studio 6, 2558, หน้า 67-90
3. Peker S, Kocyigit A, Eren PE. LRFMP model for customer segmentation in the grocery retail industry a case. Market Intel Plan. 2017;35(4) :544-59.
4. Wu HH, Lin SY, Liu CW. Analyzing Patients' Values by Applying Cluster Analysis and LRFM Model in a Pediatric Dental Clinic in Taiwan. Sci World J. 2014; DOI: 10.1155/2014/685495.
5. Geraili B, Nasiri M, Arab M. Improving Fraud and Abuse Detection in General Physician Claim A Data Mining Study. Int J Health Politic Manag. 2015;5(3):165-72.
6. Thornton D, van Capelleveen G, Poel M, van Hillegersberg J, Mueller RM. Outlier-based Health Insurance Fraud Detection for U.S. Medicaid. In Proceedings of the 16th International Conference on Enterprise Information Systems. 2014:684-94.
7. Liu Q, Vasarhelyi M. Healthcare fraud detection A survey and a clustering model incorporating Geo-location information. in 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia. 2013:1-10.