

วิธีจำแนกกลุ่มระยะการเป็นโรคมะเร็งเต้านมด้วยกระบวนการทางสถิติ
CLASSIFICATION GROUPS FOR THE STAGE OF BREAST CANCER WITH
STATISTICS METHOD

เกรียง กิจบำรุงรัตน์

Krieng Kitbumrungrat

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏธนบุรี

E-mail: Kriengstat@yahoo.com (Corresponding Author)

บทคัดย่อ

การวิจัยครั้งนี้เป็นการหาตัวแบบการพยากรณ์การจัดกลุ่มความน่าจะเป็นผู้ป่วยที่ตรวจพบว่าเป็นโรคมะเร็งเต้านมหรือไม่และถ้าเป็นมะเร็งเป็นระยะใด โดยพิจารณาตามลักษณะเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านม ซึ่งมีตัวแปรอิสระคือเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ได้แก่ ความหนาของก้อนเนื้อ, ความสม่ำเสมอของขนาดเซลล์, ความสม่ำเสมอของรูปร่างเซลล์, การเกาะติดขอบของเซลล์, ขนาดเซลล์เดี่ยว, นิวเคลียสไม่ถูกห่อหุ้ม, โครมาตินเฉพาะ, นิวคลีโอไลในภาวะปกติและการแบ่งตัวของเซลล์, ส่วนตัวแปรตามคือความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมตรวจพบว่าเป็นโรคมะเร็งเต้านมในระยะใดหรือไม่เป็นโรคมะเร็งเต้านมโดยใช้เทคนิคตัวแบบการถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) และวิธีวิเคราะห์การจำแนกประเภท (Discriminant Analysis Model) สรุปผลได้ว่าตัวแบบ Ordinal Logistic Regression ใช้ตัวแปรน้อยเพียงบางตัวในการพยากรณ์ในแต่ละระยะของการเป็นมะเร็งและตัวแบบ Ordinal Logistic Regression มีอำนาจจำแนกได้ถูกต้องร้อยละ 55.50 สูงกว่าวิธี Discriminant Analysis มีอำนาจจำแนกได้ถูกต้องร้อยละ 54.10 ตัวแบบการถดถอยเชิงอันดับมีอำนาจจำแนกแบ่งได้เป็นผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 73.60 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 1-4 ได้ถูกต้องร้อยละ 5, ร้อยละ 43.6, ร้อยละ 61.4 และร้อยละ 76.40 ตามลำดับ ส่วนตัวแบบ Discriminant Analysis ใช้ตัวแปรพยากรณ์ทั้ง 9 ชนิดในการพยากรณ์ โดยที่ตัวแบบ Discriminant Analysis มีอำนาจจำแนกผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 56 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 1-4 ได้ถูกต้องร้อยละ 49.10 ร้อยละ 35.6 ร้อยละ 72.6 และร้อยละ 60 ตามลำดับ

คำสำคัญ : โรคมะเร็งเต้านม, ตัวแบบการถดถอยเชิงอันดับ, ตัวแบบการวิเคราะห์การจำแนกประเภท
สถิติทดสอบวอลด์, สถิติทดสอบวิกส์ แลมด้า.

ABSTRACT

This research is prediction of a model of possibility for detection of non- breast cancer and breast cancer and stages of breast cancers. Assessment of characteristics of abnormal growth of breast cancer cells such as Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8), and Mitoses (X_9) are independent variable. The dependent variable is the probability that the patient is detected at

any stage of breast cancer or non-breast cancer based on the tumor cells with abnormal growth of breast cancer by using Ordinal Logistic Regression Model and Discriminant Analysis Model. In conclusions Ordinal Logistic Regression Model can use few variables in prediction of stages of breast cancer and Ordinal Logistic Regression Model has 55.50% classification higher than Discriminant Analysis Model has 54.10% classification. Ordinal Logistic Regression Model showed classification accuracy as following: non-breast cancer 73.60%, breast cancer stage 1- 4 as 5%, 43.6%, 61.4% and 76.4% respectively. Nine predicted variables were used in Discriminant Analysis Model and classification accuracy was recorded as non-breast cancer 56%, breast cancer stage 1- 4 as 49.1%, 35.6%, 72.6% and 60% respectively.

Keywords: Breast cancer, Ordinal Logistic Regression Model, Discriminant Analysis Model, Wald statistic, Wilks' Lambda

1. บทนำ

โรคมะเร็งจัดเป็น 1 ใน 10 ของสาเหตุการตายของประชากรโลก องค์การอนามัยโลก (WHO) รายงานว่าในปี ค.ศ. 2008 มีผู้เสียชีวิตทั่วโลก 58 ล้านคน สาเหตุมาจากโรคมะเร็ง 7.6 ล้านคน คิดเป็นร้อยละ 13 ของจำนวน ผู้เสียชีวิตทั้งหมด ปัจจุบันโรคมะเร็งเป็นโรคที่ก่อความสูญเสียให้กับชีวิตคนเราเป็นจำนวนมาก เช่น มะเร็งเต้านม มะเร็งตับ เป็นต้น มะเร็งเป็นเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ แพร่กระจายไปทั่วร่างกาย แบ่งตัวเร็วกว่าเซลล์ปกติหลายเท่าอยู่นอกเหนือการควบคุมของร่างกาย สามารถแพร่กระจายไปยังบริเวณอื่นของร่างกายที่ห่างไกลออกไป และไม่ติดต่อกับก้อนมะเร็งเดิม โดยแพร่เข้าไปในหลอดโลหิตและน้ำเหลือง คุณสมบัติของก้อนเนื้อเยื่อ มะเร็งโตเร็วมาก บริเวณก่อนตรงกลางมักขาดสารอาหารทำให้เกิดการตายของเนื้อเยื่อมะเร็งเกิดขึ้น ผลต่อมาคือเกิดการติดเชื้อได้ง่าย เพราะเนื้อตายและน้ำเหลืองเป็นอาหารที่ดีสำหรับเชื้อโรค การอักเสบติดเชื้อลุกลามและส่งผลทำให้เกิดการเสียชีวิตในที่สุด มะเร็งทำให้ร่างกายขาดออกซิเจน เพราะมะเร็งแย่งออกซิเจนไปใช้ในการแบ่งตัวของ เซลล์ ทำให้เม็ดเลือดขาวทำงานหนักคอยกำจัดเซลล์มะเร็งจนทำให้ภูมิคุ้มกันต่ำลง ทำให้อวัยวะที่เป็นมะเร็งสูญเสียการทำงาน ผลการขยายตัวทำให้เบียดอวัยวะข้างเคียงให้เสียหายได้ เซลล์มะเร็งจะทำการสร้างสารแองจิโอเจเนซิสมี ผลทำให้เกิดเส้นท่อเลือดมาเลี้ยงก้อนมะเร็งเพิ่มมากขึ้น เพื่อให้เพียงพอต่อการเจริญแบ่งตัวของเนื้อเยื่อมะเร็ง โรคมะเร็งเต้านมเป็นโรคที่พบมากที่สุดอันดับหนึ่งของผู้หญิงสามารถถ่ายทอดทางพันธุกรรม ซึ่งเกิดจากฮอโมนที่ผิดปกติ ความอ้วน อาหารที่รับประทานมีไขมันสูง การกลูกลามของมะเร็งจากอวัยวะอื่นมายังเต้านม (ชัญวลี ศรีสุโข ,2554)

การรักษาผู้ป่วยมะเร็งเต้านมตามแนวทางของสถาบันมะเร็งแห่งชาติ กรมการแพทย์ กระทรวงสาธารณสุข (2551) มีการรักษา 4 แบบ คือ การผ่าตัด การฉายรังสี การรักษาด้วยระบบยา (systemic therapy) และการรักษา มะเร็งระยะแพร่กระจาย ในการรักษาโรคมะเร็งเต้านมมักจะใช้การรักษาหลายวิธีร่วมกัน เพื่อป้องกันการกลับมาเป็นซ้ำ และผลที่เกิดขึ้นจากการรักษาด้วยวิธีต่าง ๆ เหล่านี้คือ อาการข้างเคียงที่ไม่พึงประสงค์ เช่น คลื่นไส้ อาเจียน ผมร่วง ท้องเสีย โลหิตจาง เป็นต้น จากผลที่เกิดขึ้นดังกล่าวก่อให้เกิดความวิตกกังวล ทุกข์ทรมานทั้งกายและใจต่อผู้ป่วยเป็นอันมาก ค่าใช้จ่ายต่าง ๆ ที่มีราคาสูงมากจากการทั้งการรักษาเอง การต้องหยุดพักงานของญาติที่ดูแลผู้ป่วยก็เป็นเหตุที่ทำให้ขาดรายได้เช่นกัน แม้ว่าผลการรักษาทาง การแพทย์จะทำให้ผู้ป่วยมะเร็งหายขาดได้ ถ้าตรวจพบในระยะเริ่มแรกแต่ สำหรับผู้ป่วยที่เพิ่งตรวจพบว่าเป็นระยะท้าย ๆ แล้ว การรักษาที่ไม่อาจทำให้หายขาดได้เนื่องจากมะเร็งเป็นโรคเรื้อรัง (สถาบันมะเร็งแห่งชาติกรมการแพทย์ กระทรวงสาธารณสุข ,2550).

การสร้างระบบสุขภาพของผู้ป่วยโรคมะเร็งเต้านมนั้น จะพิจารณาจาก ระยะของโรคมะเร็งเต้านม (The Stage of Breast Cancer) ซึ่งเป็นตัวบ่งบอกการลุกลาม ความรุนแรงของโรคมะเร็งเต้านม การที่แพทย์ทราบระยะของ โรคมะเร็งเต้านม แพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำ กับคนไข้ในระยะเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ

ของโรคมะเร็งเต้านมเป็นตัวกำหนดระยะของโรคมะเร็งเต้านม ซึ่งเป็นเซลล์มะเร็งที่แบ่งตัว เติบโตนอกเหนือการควบคุมจากร่างกาย สามารถเติบโตลุกลามเข้าไปในอวัยวะใกล้เคียงโดยเข้ากระแสเลือด ต่อมามีการแพร่กระจาย ซึ่งเรียกว่าการลุกลามแพร่กระจาย (Metastasis)

ด้วยเหตุผลดังกล่าวนี้ การวิจัยนี้จึงมีความสนใจที่จะศึกษาวิธีการสำหรับการพยากรณ์เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมว่าผู้ป่วยถ้าเป็นมะเร็งจะมีความน่าจะเป็นที่เป็นมะเร็งอยู่ในระยะขั้นใดของโรคมะเร็ง โดยใช้วิธีการทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธี ได้แก่ การวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) อย่างไรก็ตามประสิทธิภาพของวิธีการวิเคราะห์การจำแนกประเภทขึ้นอยู่กับลักษณะการแจกแจงของข้อมูลที่นำมาวิเคราะห์ว่าเป็นไปตามข้อสมมติของตัวแบบด้วยหรือไม่ โดยวิธีการวิเคราะห์การจำแนกประเภทมีข้อสมมติเกี่ยวกับตัวแบบค่อนข้างมาก เช่น ความสัมพันธ์ระหว่างกลุ่มและตัวแปรพยากรณ์ต้องเป็นแบบเชิงเส้น ตัวแปรพยากรณ์ควรมีการแจกแจงร่วมแบบปกติพหุและไม่มีความสัมพันธ์ระหว่างกันมากจนก่อให้เกิดปัญหาภาวะแนวร่วมเส้นตรง เป็นต้น สำหรับการวิเคราะห์การถดถอยเชิงอันดับมีข้อจำกัดน้อยกว่า วิธีการวิเคราะห์การจำแนกประเภท

2.วัตถุประสงค์ของการวิจัย

เพื่อศึกษาว่าวิธีการพยากรณ์จำแนกกลุ่ม 2 วิธี ได้แก่วิธีการวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) เมื่อนำมาใช้ในการพยากรณ์ความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมตรวจอาการแล้วพบว่าไม่เป็นโรคมะเร็ง หรือถ้าพบว่าเป็นโรคมะเร็งแล้วจะเป็นระยะใดจากขั้นที่ 1 ถึงขั้นที่ 4 เป็นระยะขั้นใดโดยใช้เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติเป็นตัวแปรพยากรณ์ โดยใช้วิธีการวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) กับวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) เพื่อพิจารณาความสามารถในการพยากรณ์การจำแนกกลุ่มได้ถูกต้องและเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติใดบ้างที่สามารถใช้ในการพยากรณ์ได้

กรอบแนวคิดของการวิจัย

การวิเคราะห์การพยากรณ์เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับระยะความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมที่ตรวจพบว่าไม่เป็นโรคมะเร็ง หรือถ้าพบว่าเป็นโรคมะเร็งแล้วจะเป็นระยะขั้นที่ 1 ขั้นที่ 2 ขั้นที่ 3 หรือขั้นที่ 4 โดยใช้ชุดฐาน ข้อมูลจาก Wisconsin Breast Cancer Database (WBCD) ซึ่งลักษณะของตัวแปรของข้อมูลเป็นดังต่อไปนี้

ตัวแปรอิสระ

- เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ
- Clump Thickness (ความหนาของก้อนเนื้อ)
- Uniformity of Cell Size (ความสม่ำเสมอของขนาดเซลล์)
- Uniformity of Cell Shape (ความสม่ำเสมอของรูปร่างเซลล์)
- Marginal Adhesion (การเกาะติดขอบของเซลล์)
- Single Epithelial Cell Size (ขนาดเซลล์เดี่ยว)
- Bare Nuclei (นิวเคลียสไม่ถูกห่อหุ้ม)
- Normal Nucleoli (นิวคลีโอลในภาวะปกติ)
- Bland Chromatin (ของเหลวในเซลล์)
- Mitoses (การขยายตัวของเซลล์)

ตัวแปรตาม

- ความน่าจะเป็นที่ผู้ป่วยจะไม่เป็นมะเร็งเต้านม
- ความน่าจะเป็นที่ผู้ป่วยจะเป็นมะเร็งเต้านม
- เป็นมะเร็งเต้านม ระยะที่ 1
- เป็นมะเร็งเต้านม ระยะที่ 2
- เป็นมะเร็งเต้านม ระยะที่ 3
- เป็นมะเร็งเต้านม ระยะที่ 4

ภาพที่ 1 กรอบแนวคิดของการวิจัย

3.วิธีการดำเนินการทดลอง

การวิจัยนี้เป็นการวิจัยเชิงทดลอง (Experiment Research) เพื่อศึกษาการพยากรณ์เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับระยะความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมที่ตรวจพบว่าไม่เป็นโรคมะเร็งเต้านม หรือถ้าพบว่าเป็นโรคมะเร็งแล้วจะเป็นมะเร็งระยะขั้นที่ 1 ขั้นที่ 2 ขั้นที่ 3 หรือขั้นที่ 4 โดยใช้ตัวอย่างฐานข้อมูลการตรวจรักษาโรคมะเร็งเต้านม Wisconsin Breast Cancer Database (WBCD) จำนวน 683 ตัวอย่าง ซึ่งตัวแปรตาม (Y) คือระยะความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมไม่เป็นโรคมะเร็งเต้านม หรือถ้าพบว่าเป็นโรคมะเร็งเต้านมแล้วจะเป็นมะเร็งระยะขั้นที่ 1 ขั้นที่ 2 ขั้นที่ 3 หรือขั้นที่ 4 ส่วนตัวแปรอิสระคือเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติได้แก่ Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8) และ Mitoses (X_9) โดยใช้เทคนิคตัวแบบการถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) และเทคนิควิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis Model) โดยใช้โปรแกรมสำเร็จรูป SPSS และโปรแกรมสำเร็จรูป SAS ในการวิเคราะห์ข้อมูลทางสถิติ

ทฤษฎีและวิธีการศึกษา

3.1 ตัวแบบการถดถอยเชิงอันดับ (Ordinal Logistic Regression Model)

ตัวแบบวิธีการถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) เป็นเทคนิคของตัวแบบการวิเคราะห์การถดถอยที่ใช้สำหรับการวิเคราะห์ข้อมูลที่มีตัวแปรตามมีระดับการวัดแบบมาตราเรียงลำดับ ในการวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) การจำแนกกลุ่มทำได้โดยการประมาณค่าความน่าจะเป็นของการที่ค่าสังเกตจะอยู่ในกลุ่มที่กำหนด เช่น ในกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมเมื่อค่าของตัวแปรพยากรณ์เป็นค่าที่สังเกตจากค่าประมาณความน่าจะเป็นที่ได้ ค่าสังเกตจะถูกจำแนกให้อยู่ในกลุ่มที่กำหนดหรือกลุ่มอื่นขึ้นอยู่กับค่าความน่าจะเป็นที่เป็นจุดตัด (cut-off probability) โดยในการประมาณสัมประสิทธิ์การถดถอยเชิงอันดับ เพื่อการประมาณความน่าจะเป็นนั้นใช้วิธีประมาณแบบภาวะน่าจะเป็นสูงสุด (maximum likelihood estimation) ตัวแบบของการถดถอยเชิงอันดับคือ

$$\ln \left(\frac{P_i}{1 - P_i} \right) = \alpha_i + \beta x, \text{ เมื่อ } P_i \text{ คือความน่าจะเป็นเหตุการณ์ที่ผู้ป่วยเป็นโรคมะเร็งเต้านมระยะที่ } i; i = 0, 1, 2, 3, 4$$

α_i คือ ค่าคงที่สำหรับกลุ่มของตัวแปรตามที่ i ; $i = 1, 2, 3, \dots, n - 1$

βx คือ $\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$; k คือ จำนวนสัมประสิทธิ์การถดถอย

และ $X_1, X_2, \dots, X_p$ คือเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติจะเป็นตัวแปรพยากรณ์
ตัวแบบของการวิเคราะห์การถดถอยเชิงอันดับกรณีใช้ Logit เป็น Link function คือ

$$\ln \left(\frac{P_i}{1 - P_i} \right) = \alpha_i + \beta x$$

ถ้าตัวแปรตามมี i ระดับ จะได้ตัวแบบในการพยากรณ์จำแนกกลุ่ม $i - 1$ กลุ่ม ดังนี้

$$\ln\left(\frac{P_1}{1-P_i}\right) = \alpha_1 + \beta x$$

$$\ln\left(\frac{P_2}{1-P_i}\right) = \alpha_2 + \beta x$$

:

$$\ln\left(\frac{P_{i-1}}{1-P_{i-1}}\right) = \alpha_{i-1} + \beta x$$

โดยที่ตัวแบบของกลุ่มสุดท้ายจะเป็นตัวแปรอ้างอิง

3.2 ตัวแบบการวิเคราะห์การจำแนกประเภท (Discriminant Model)

การวิเคราะห์การจำแนกประเภทเป็นวิธีการจำแนกกลุ่มโดยการสร้างฟังก์ชันจำแนกกลุ่มซึ่งเป็นผลบวกเชิงเส้นของตัวแปรอิสระหรือตัวแปรพยากรณ์ $X_1, X_2, \dots, X_p$ เพื่อให้ความแปรปรวนระหว่างกลุ่มเมื่อเทียบกับความแปรปรวนภายในกลุ่มมีค่าสูงสุด นั่นคือหาค่าสัมประสิทธิ์ $a_1, a_2, \dots, a_p$ ของ $X_1, X_2, \dots, X_p$ ที่ทำให้ได้ฟังก์ชันจำแนกที่มีคุณสมบัติดังกล่าวดังนี้ ทำให้ได้ $Z_i = a_0 + a_1X_{i1} + a_2X_{i2} + \dots + a_pX_{ip}$ เป็นคะแนนจำแนก (Discriminant score) สำหรับค่าสังเกตที่ i คะแนนที่ใช้เป็นจุดตัดในการจำแนก (cut-off score) คำนวณจากความน่าจะเป็นก่อน (a-priori probabilities) ของการเป็นสมาชิกกลุ่มและต้นทุนในการจำแนกผิดจากคะแนนจำแนกที่ได้ของแต่ละค่าสังเกตและคะแนนจุดตัดค่าสังเกตจะถูกจำแนกว่าอยู่กลุ่มใด Discriminant Model มีข้อสมมติเบื้องต้นเกี่ยวกับตัวแปรพยากรณ์ว่าในแต่ละกลุ่มตัวแปรพยากรณ์ต้องมีการแจกแจงร่วมแบบปกติพหุที่มีเวกเตอร์ค่าเฉลี่ยแตกต่างกัน แต่เมตริกซ์ความแปรปรวนร่วมเท่ากัน

4. ผลการวิเคราะห์

4.1 การวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis)

ขั้นตอนที่ 1 การวิเคราะห์ข้อมูลระยะการเป็นโรคมะเร็งเต้านมของผู้ป่วย จากการตรวจรักษาโรคมะเร็งเต้านมในแต่ละระยะมีสัดส่วนเท่ากันหรือไม่ผลการตรวจสอบพบว่าผู้ป่วยทั้งหมด 683 คนมีผู้ป่วยมาตรวจแล้วพบว่าไม่เป็นโรคมะเร็งมีจำนวน 182 คน พบว่าเป็นโรคมะเร็งในระยะที่ 1 มีจำนวน 100 คน เป็นโรคมะเร็งในระยะที่ 2 จำนวน 156 คน เป็นโรค มะเร็งในระยะที่ 3 จำนวน 101 คน และเป็นโรคมะเร็งในระยะที่ 4 จำนวน 144 คน เมื่อทำการทดสอบสัดส่วนการตรวจรักษาโรคมะเร็งในแต่ละระยะแล้ว พบว่าไม่เป็นโรคมะเร็งเต้านม หรือเป็นโรคมะเร็งเต้านมในระยะใดของผู้ป่วยแต่ละระยะมีสัดส่วนเท่ากัน ($P\text{-value} = 0.768 > 0.05$ จึงยอมรับสมมติฐาน H_0 : ข้อมูลในแต่ละระดับของตัวแปรตามในแต่ละระยะมีสัดส่วนเท่ากัน) จึงสามารถใช้ Logit เป็น Link function ในการวิเคราะห์ข้อมูลได้

ขั้นตอนที่ 2 ทดสอบความสัมพันธ์ระหว่างชุดของตัวแปรอิสระเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับตัวแปรตามของระยะการเป็นมะเร็งเต้านม โดยใช้ Logit เป็น Link function ในการวิเคราะห์ข้อมูลได้ผลการวิเคราะห์ ดังตารางที่ 1

ตารางที่ 1 การทดสอบความสัมพันธ์ของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับระยะการเป็นมะเร็งเต้านม

Model Fitting Information				
Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1812.290			
Final	918.692	893.598	36	.000

จากตารางที่ 1 พบว่าถ้าตัวแบบประกอบด้วยค่าคงตัวเพียงอย่างเดียวจะให้ค่า -2 Log Likelihood (Intercept Only) = 1812.290 และถ้าตัวแบบประกอบด้วยค่าคงตัวและตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติจะให้ค่า -2Log Likelihood (Final) = 918.692 นั่นคือตัวแบบที่มีตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติรวมอยู่ด้วยมีความเหมาะสมมากกว่าตัวแบบที่มีค่าคงที่เพียงอย่างเดียว นั่นคือมีตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอย่างน้อย 1 ตัวมีความสัมพันธ์กับตัวแปรตามของระยะการเป็นมะเร็งเต้านม หรือจำนวนเหตุการณ์ตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติมีความสัมพันธ์กับสถานะการตรวจ พบว่าไม่เป็นโรค มะเร็งเต้านมหรือเป็นโรคมะเร็งเต้านมในแต่ละระยะของผู้ป่วย (Chi-Square=893.598 ซึ่งได้ค่าP-value=0.000 < 0.05 จึงปฏิเสธ H_0 นั่นคือตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติมีความสัมพันธ์กับตัวแปรตามของระยะการเป็นมะเร็งเต้านม)

ขั้นตอนที่ 3 การทดสอบสัมประสิทธิ์การถดถอยของตัวแบบโดยการทดสอบ Likelihood Ratio Test ผลการทดสอบดังตารางที่ 2

ตารางที่ 2 การทดสอบสัมประสิทธิ์การถดถอยของตัวแบบ (Likelihood Ratio Test)

Effect	-2log Likelihood	Chi-Square	p-value
Intercept	1237.876	319.184	(0.000)*
X ₁ (Clump Thickness)	946.687	27.995	(0.000)*
X ₂ (Uniformity of cell size)	927.703	9.011	0.061
X ₃ (Uniformity of cell shape)	922.282	3.590	0.464
X ₄ (Marginal Adhesion(x4))	926.650	7.958	0.093
X ₅ (Single Epithelial cell size)	929.785	11.093	(0.026)*
X ₆ (Bare nuclei)	941.023	22.331	(0.000)*
X ₇ (Bland Chromatin)	972.989	54.297	(0.000)*
X ₈ (Normal nucleoli)	933.431	14.739	(0.005)*
X ₉ (Mitoses)	973.469	54.777	(0.000)*

*: p-value < 0.05 มีนัยสำคัญทางสถิติระดับ 0.05

จากตารางที่ 2 พบว่าสัมประสิทธิ์ตัวแบบการถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติมีผลต่อระยะการเป็นโรคมะเร็งเต้านมที่มีค่า P-value < 0.05 คือ ความหนาของก้อนเนื้อ (Clump Thickness (X₁)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X₅)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X₆)), โครมาตินเฉพา (Bland Chromatin(X₇)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X₈)) และการแบ่งตัวของเซลล์ (Mitoses (X₉)) ดังนั้นจะได้ตัวแบบในการพยากรณ์ความน่าจะเป็นสะสมเพื่อการจำแนกกลุ่ม

ผู้ป่วยตามสถานะสุขภาพจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลต่อระยะการเป็นมะเร็งเต้านมดังนี้

$$\ln \left(\frac{P_i}{1-P_i} \right) = 1237.876 + 946.687X_1 + 929.785X_5 + 941.023X_6 + 972.989X_7 + 933.431X_8 + 973.469 X_9$$

การทดสอบสัมประสิทธิ์การถดถอยของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับระยะการเป็นมะเร็งเต้านมในแต่ละระยะพบว่า การวิเคราะห์ข้อมูลสัมประสิทธิ์ตัวแบบการถดถอยเชิงอันดับเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลต่อระยะการเป็นมะเร็งเต้านมในแต่ละระยะที่มีค่า P-value < 0.05 จะสามารถสร้างสมการตัวแบบถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) ได้ 4 สมการ ดังนี้

$$\text{สมการไม่พบโรคมะเร็งเต้านม} \ln \left(\frac{P_0}{1-P_0} \right) = 10.044 - 0.464 X_5 + 0.526 X_6 - 0.761 X_7 - 0.434 X_8 - 1.166$$

X_9

$$\text{สมการโรคมะเร็งเต้านมระยะที่ 1} \ln \left(\frac{P_1}{1-P_1} \right) = 7.608 - 0.524 X_1 - 0.340 X_6$$

$$\text{สมการโรคมะเร็งเต้านมระยะที่ 2} \ln \left(\frac{P_2}{1-P_2} \right) = 8.063 - 0.451 X_1 - 0.313 X_4 - 0.322 X_6 - 0.333 X_8$$

$$\text{สมการโรคมะเร็งเต้านมระยะที่ 3} \ln \left(\frac{P_3}{1-P_3} \right) = 1.607 + 0.243 X_7 - 0.140 X_8 - 0.704 X_9$$

โดยมีโรคมะเร็งในระยะที่ 4 ซึ่งเป็นระยะสุดท้ายจะเป็นระยะอ้างอิงตามสภาวะสุขภาพของการมีชีวิตอยู่รอด

ดังนั้นในการตรวจผู้ป่วยแล้วไม่พบว่าเป็นโรคมะเร็งเต้านมควรพิจารณาเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ อาจมีความสัมพันธ์กับขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการแบ่งตัวของเซลล์ (Mitoses (X_9))

การตรวจผู้ป่วยแล้วพบว่าเป็นโรคมะเร็งเต้านมระยะที่ 1 พิจารณาเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอาจมีความสัมพันธ์กับความหนาของก้อนเนื้อ (Clump Thickness (X_1)) และนิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6))

การตรวจผู้ป่วยแล้วพบว่าเป็นโรคมะเร็งเต้านมระยะที่ 2 พิจารณาเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอาจมีความสัมพันธ์กับความหนาของก้อนเนื้อ (Clump Thickness (X_1)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และนิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8))

การตรวจผู้ป่วยแล้วพบว่าเป็นโรคมะเร็งเต้านมระยะที่ 3 พิจารณาเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอาจมีความสัมพันธ์กับความปกติของโครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการแบ่งตัวของเซลล์ (Mitoses (X_9))

โดยมีโรคมะเร็งในระยะที่ 4 ซึ่งเป็นระยะสุดท้ายจะเป็นระยะอ้างอิงตามสภาวะสุขภาพของการมีชีวิตอยู่รอด

ตัวแบบถดถอยเชิงอันดับ (Ordinal Logistic Regression Model) มีอำนาจจำแนกกลุ่มโดยรวมได้ถูกต้องร้อยละ 55.5 แบ่งเป็นการพยากรณ์ผู้ป่วยไม่เป็นมะเร็งได้ถูกต้องร้อยละ 73.6 มะเร็งระยะที่ 1 ได้ถูกต้องร้อยละ 5 มะเร็งระยะที่ 2 ได้ถูกต้องร้อยละ 43.6 มะเร็งระยะที่ 3 ได้ถูกต้องร้อยละ 61.4 และมะเร็งระยะที่ 4 ได้ถูกต้องร้อยละ 76.4 ซึ่งเป็นมะเร็งระยะสุดท้าย

4.2 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

การวิเคราะห์ข้อมูลด้วยการวิเคราะห์จำแนกประเภท (Discriminant Analysis) ในการศึกษาปัจจัยเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลต่อระยะความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมที่ตรวจพบว่าเป็นโรคมะเร็งเต้านมหรือถ้าพบว่าเป็นโรคมะเร็งเต้านมแล้วจะเป็นมะเร็งระยะขั้นที่ 1 ขั้นที่ 2 ขั้นที่ 3 หรือขั้นที่ 4 เมื่อทำการทดสอบว่าแต่ละปัจจัยที่นำมาใช้จำแนกกลุ่มผู้ป่วยมีความแตกต่างกันระหว่างกลุ่มหรือไม่ด้วยสถิติทดสอบ Wilks' Lambda พบว่าปัจจัยที่แตกต่างกันระหว่างกลุ่มที่ระดับนัยสำคัญ 0.05 ได้แก่ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติทั้ง 9 ชนิด แสดงได้ในตารางที่ 3

ตารางที่ 3 ค่าสถิติทดสอบการเท่ากันของค่าเฉลี่ยของแต่ละปัจจัย

Effect	Wilks' Lambda	F-test	p-value
X ₁ (Clump Thickness)	0.499	170.398	(0.000)*
X ₂ (Uniformity of cell size)	0.347	319.202	(0.000)*
X ₃ (Uniformity of cell shape)	0.346	320.586	(0.000)*
X ₄ (Marginal Adhesion(x4))	0.518	157.637	(0.000)*
X ₅ (Single Epithelial cell size)	0.522	155.322	(0.000)*
X ₆ (Bare nuclei)	0.350	314.798	(0.000)*
X ₇ (Bland Chromatin)	0.434	221.420	(0.000)*
X ₈ (Normal nucleoli)	0.479	184.063	(0.000)*
X ₉ (Mitoses)	0.701	72.237	(0.000)*

*: p-value < 0.05 มีนัยสำคัญทางสถิติระดับ 0.05

จากตารางที่ 3 เมื่อคัดเลือกปัจจัยที่ใช้จำแนกกลุ่มพบว่าปัจจัยที่ใช้จำแนกกลุ่มคือความหนาของก้อนเนื้อ (Clump Thickness (X₁)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X₂)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X₃)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X₄)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X₅)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X₆)), โครมาตินเฉพาะ (Bland Chromatin (X₇)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X₈)) และการแบ่งตัวของเซลล์ (Mitoses (X₉)) ซึ่งจะได้ตัวแบบปัจจัยการวิเคราะห์จำแนกประเภทแบ่งกลุ่มของกลุ่มผู้ป่วยไม่เป็นโรคมะเร็งเต้านม และกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมในแต่ละระยะด้วยตัวแบบปัจจัยที่ใช้จำแนกกลุ่ม Fisher's linear discriminant functions จะได้

ตัวแบบกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมดังต่อไปนี้

ตัวแบบกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมคือ

$$D_0 = -3.989 + 0.768X_1 + 0.301X_2 + 0.037X_3 + 0.129X_4 + 0.624X_5 + 0.099X_6 + 0.459X_7 + 0.003X_8 + 0.224X_9$$

ตัวแบบกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมระยะที่ 1 คือ

$$D_1 = -4.24 + 0.603X_1 + 0.396X_2 + 0.031X_3 + 0.009X_4 + 0.678X_5 + 0.145X_6 + 0.825X_7 + 0.042X_8 + 0.246X_9$$

โดยที่มีค่า Eigenvalue = 4.430, Wilks' Lambda = 0.143, Chi-square = 1311.499, P-value = 0.000 <

0.05

ตัวแบบกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมระยะที่ 2 คือ

$$D_2 = -4.044 + 0.625X_1 + 0.393X_2 + 0.040X_3 + 0.025X_4 + 0.707X_5 + 0.144X_6 + 0.677X_7 + 0.007X_8 + 0.301X_9$$

โดยที่มีค่า Eigenvalue = 0.226, Wilks' Lambda = 0.778, Chi-square = 169.474, P-value = 0.000 < 0.05

ตัวแบบกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมระยะที่ 3 คือ

$$D_3 = -18.853 + 1.389X_1 + 0.278X_2 + 0.247X_3 + 0.278X_4 + 0.796X_5 + 1.082X_6 + 1.241X_7 + 0.290X_8 + 0.110X_9$$

โดยที่มีค่า Eigenvalue = 0.046, Wilks' Lamda = 0.954, Chi-square = 31.810, P-value = 0.004 < 0.05

ตัวแบบกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมระยะที่ 4 คือ

$$D_4 = -22.361 + 1.424X_1 + 0.056X_2 + 0.468X_3 + 0.209X_4 + 1.003X_5 + 1.214X_6 + 0.831X_7 + 0.612X_8 + 0.826X_9$$

โดยที่มีค่า Eigenvalue = 0.002, Wilks' Lamda = 0.0998, Chi-square = 1.209, P-value = 0.976 > 0.05

จะได้ว่า ตัวแบบ D_4 จะไม่มีความสัมพันธ์กับตัวแบบ เพราะว่าโรคมะเร็งในระยะที่ 4 ซึ่งเป็นระยะสุดท้ายจะเป็นระยะอ้างอิงตามสภาวะสุขภาพของการมีชีวิตอยู่รอด

4.3 ผลการเปรียบเทียบประสิทธิภาพ

ประสิทธิภาพของตัวแบบ Discriminant Analysis และ Ordinal Logistic Regression Analysis พิจารณาจากอัตราการจำแนกกลุ่มได้ถูกต้องซึ่งแสดงไว้ในตารางที่ 4

ตารางที่ 4 จำนวนและร้อยละของผู้ป่วยที่จำแนกกลุ่มได้ถูกต้องสำหรับตัวแบบ Discriminant Analysis และ Logistic Regression Analysis เมื่อใช้กับกลุ่มตัวอย่างผู้ป่วย 683 คน

ตัวแบบ	Discriminant Analysis	Ordinal Logistic Regression
ตัวแปรพยากรณ์		
-ไม่เป็นโรคมะเร็งเต้านม	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_5, X_6, X_7, X_8, X_9
-เป็นโรคมะเร็งเต้านมระยะที่ 1	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_1, X_6
-เป็นโรคมะเร็งเต้านมระยะที่ 2	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	$X_1, X_4, X_6, X_8,$
-เป็นโรคมะเร็งเต้านมระยะที่ 3	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_7, X_8, X_9
-เป็นโรคมะเร็งเต้านมระยะที่ 4	ระยะที่ 4 เป็นตัวแปรอ้างอิง	ระยะที่ 4 เป็นตัวแปรอ้างอิง
อำนาจการจำแนกกลุ่ม (%)		
-ไม่เป็นโรคมะเร็งเต้านม	56.0%	73.6%
-เป็นโรคมะเร็งเต้านมระยะที่ 1	49.1%	5.0%
-เป็นโรคมะเร็งเต้านมระยะที่ 2	35.6%	43.6%
-เป็นโรคมะเร็งเต้านมระยะที่ 3	72.6%	61.4%
-เป็นโรคมะเร็งเต้านมระยะที่ 4	60.0%	76.4%
รวม	54.1%	55.5%

จากตารางที่ 4 พบว่าตัวแบบ Discriminant Analysis ต้องใช้ตัวแปรทุกตัวในการพยากรณ์ทุกระยะการเป็นโรค มะเร็งเต้านมหรือไม่เป็นมะเร็งเต้านม โดยมีอำนาจจำแนกกลุ่มผู้ป่วยภาพรวมได้ถูกต้องร้อยละ 54.10 แบ่งเป็นผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 56 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 1 ได้ถูกต้องร้อยละ 49.10 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 2 ได้ถูกต้องร้อยละ 35.6 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 3 ได้ถูกต้องร้อยละ 72.6 และผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 4 ได้ถูกต้องร้อยละ 60

ตัวแบบ Ordinal Logistic Regression ใช้ตัวแปรน้อยเพียงบางตัวในการพยากรณ์ในแต่ละระยะของการเป็นมะเร็ง ซึ่ง Ordinal Logistic Regression Model มีอำนาจจำแนกกลุ่มสูงกว่าวิธี Discriminant Analysis Model โดยสามารถจำแนกกลุ่มผู้ป่วยภาพรวมได้ถูกต้องร้อยละ 55.5 แบ่งเป็นผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 73.6 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 1 ได้ถูกต้องร้อยละ 5 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 2 ได้ถูกต้องร้อยละ 43.6 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 3 ได้ถูกต้องร้อยละ 61.4 และผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 4 ได้ถูกต้องร้อยละ 76.4

5. สรุปและอภิปรายผล

การพยากรณ์ความน่าจะเป็นเพื่อการจำแนกกลุ่มผู้ป่วยที่ตรวจพบว่าเป็นโรคมะเร็งเต้านมระยะใดหรือไม่เป็นโรคมะเร็งเต้านมซึ่งถือว่าเป็นสิ่งจำเป็น เพราะจะเป็นเครื่องมือในการสร้างระบบสภาวะสุขภาพจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ในการศึกษานี้ใช้เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ 9 ชนิด โดยใช้วิธีการทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธี ได้แก่วิธีการวิเคราะห์การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) ผลการวิเคราะห์พบว่าตัวแบบ Ordinal Logistic Regression Analysis มีอำนาจจำแนกได้ถูกต้องร้อยละ 55.50 สูงกว่าวิธี Discriminant Analysis มีอำนาจจำแนกได้ถูกต้องร้อยละ 54.10 ตัวแบบการถดถอยเชิงอันดับมีอำนาจจำแนกแบ่งได้เป็นผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 73.60 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 1 ได้ถูกต้องร้อยละ 5 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 2 ได้ถูกต้องร้อยละ 43.6 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 3 ได้ถูกต้องร้อยละ 61.4 และผู้ป่วยที่เป็นมะเร็งเต้านมระยะที่ 4 ได้ถูกต้องร้อยละ 76.40 ในการเลือกตัวแปรพยากรณ์เข้าสู่ตัวแบบ จะได้ว่าผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมมีความสัมพันธ์กับเซลล์เยื่อบุผิวขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการแบ่งตัวของเซลล์ (Mitoses (X_9)) ผู้ป่วยที่เป็นโรคมะเร็งเต้านมระยะที่ 1 มีความสัมพันธ์กับความหนาของก้อนเนื้อ (Clump Thickness (X_1)) และนิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) ผู้ป่วยที่เป็นโรคมะเร็งเต้านมระยะที่ 2 มีความสัมพันธ์กับความหนาของก้อนเนื้อ (Clump Thickness (X_1)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of Cell Shape (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และนิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) ผู้ป่วยที่เป็นโรคมะเร็งเต้านมระยะที่ 3 มีความสัมพันธ์กับโครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการแบ่งตัวของเซลล์ (Mitoses (X_9)) โดยมีโรคมะเร็งในระยะที่ 4 ซึ่งเป็นระยะสุดท้ายจะเป็นระยะอ้างอิงตามสภาวะสุขภาพของการมีชีวิตอยู่รอด

ส่วนตัวแบบ Discriminant Analysis ใช้ตัวแปรพยากรณ์ทั้ง 9 ชนิด ตัวแบบ Discriminant Analysis ต้องใช้ตัวแปรทุกตัวในการพยากรณ์ทุกระยะการเป็นโรคมะเร็งเต้านมหรือไม่เป็นมะเร็งเต้านม โดยมีอำนาจจำแนกกลุ่มผู้ป่วยภาพรวมได้ถูกต้องร้อยละ 54.10 แบ่งเป็นจำแนกผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 56 ผู้ป่วยที่เป็นมะเร็งเต้านมระยะต่าง ๆ สามารถจำแนกได้ถูกต้องดังนี้ ระยะที่ 1 ร้อยละ 49.10 ระยะที่ 2 ร้อยละ 35.6 ระยะที่ 3 ร้อยละ 72.6 และระยะที่ 4 ร้อยละ 60

ดังนั้นในการสร้างระบบสภาวะสุขภาพของผู้ป่วยโรคมะเร็งเต้านมนั้น จะพิจารณาจาก ระยะของโรคมะเร็งเต้านม (The Stage of Breast Cancer) ซึ่งเป็นตัวบ่งบอกการลุกลาม ความรุนแรงของโรคมะเร็งเต้านม การที่แพทย์ทราบระยะของโรคเต้านม แพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ในระยะเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมเป็นตัวกำหนดระยะของโรคมะเร็งเต้านม ซึ่งเป็นเซลล์มะเร็งที่แบ่งตัว เติบโตนอกเหนือการควบคุมจากร่างกาย สามารถเติบโตลุกลามเข้าไปในอวัยวะใกล้เคียงโดยเข้ากระแสเลือด ต่อมำน้ำเหลืองได้ ซึ่งเรียกว่าการลุกลามแพร่กระจาย (Metastasis)

ระยะโรคแบบ TNM เป็นกำหนดระยะของโรคแบบมาตรฐานสากลทั่วโลก กำหนดขึ้นโดย International Union Against Cancer (UICC) และ The American โดยที่ระยะโรคแบบ TNM กำหนดจาก T (Tumor) คือ ขนาดก้อนจำนวนก้อน ตำแหน่งที่ก้อนอยู่ N (Lymph nodes) คือ ต่อมำน้ำเหลืองที่มะเร็งลุกลามไปและ M (Metastasis) คือ การขยายตัวของเซลล์ จะได้ว่าความเป็นมะเร็งเต้านม ระยะ 0 คือ มะเร็งระยะต้น ๆ (Carcinoma in situ) ที่อยู่บนชั้นของเซลล์ปกติยังไม่แทรกเข้าไปในเนื้อเยื่อปกติ ส่วนการเป็นมะเร็งเต้านม ระยะ I ระยะ II ระยะ III คือ ขนาดต่าง ๆ กันของก้อนจากเล็กไปใหญ่ จำนวนต่อมน้ำเหลืองและตำแหน่งต่อมน้ำเหลืองที่แพร่ไป และส่วนการเป็นมะเร็งเต้านม ระยะ IV คือเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติแพร่กระจายไปยังอวัยวะอื่น ตัวอย่าง มะเร็งเต้านมระยะ T3N2M0 หมายถึง มะเร็งก้อนใหญ่ที่มีการลุกลามเข้าต่อมน้ำเหลือง แต่ยังไม่มีการแพร่กระจายไปยังอวัยวะอื่นของร่างกาย

การสร้างตัวแบบจำแนกทางสถิติเป็นการช่วยให้แพทย์ทราบว่าการตรวจเลือดที่เจริญเติบโตผิดปกติของโรค มะเร็งเต้านมชนิดใดที่มีผลกระทบต่อระยะของโรค มะเร็งเต้านม ซึ่งแพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้ และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ในระยะเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการ รักษาได้ดียิ่งขึ้น การสร้างตัวแบบจำแนกทางสถิติจะมีประสิทธิภาพก็ต่อเมื่อข้อมูลที่วิเคราะห์มีลักษณะการแจกแจง สอดคล้องกับข้อสมมติ วิธีการจำแนกกลุ่มแบบพาราเมตริกทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธีนี้ คือ การวิเคราะห์ การถดถอยเชิงอันดับ (Ordinal Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) ในภายหลังควรมีการนำวิธีการทางพาราเมตริกมาใช้กับระบบสภาวะสุขภาพของผู้ป่วย เช่น เครือข่าย ประสาทเทียม (Neural Networks) ต้นไม้การตัดสินใจ (Decision Trees) หรือการรู้จำลักษณะเฉพาะ (Trait Recognition Approach) เป็นต้น

6.บรรณานุกรม

ชัยวุฒิ ศรีสุโข (2554). โรค มะเร็งเต้านม: อมรินทร์พริ้นติ้งแอนด์พับลิชชิ่ง กรุงเทพฯ; 2554; หน้า 5-10.

สถาบันมะเร็งแห่งชาติ (2550). แนวทางการตรวจวินิจฉัยและรักษาโรค มะเร็งเต้านม. พิมพ์ครั้งที่ 1 พ.ศ. 2550.

กรุงเทพฯ :

กรมการแพทย์ กระทรวงสาธารณสุข (2546). การประเมินเทคโนโลยีการตรวจวินิจฉัยมะเร็งเต้านมในระยะเริ่มแรก ที่เหมาะสมสำหรับประเทศไทย: ตุลาคม 2546. หน้า 28-30.

ชุดฐานข้อมูลจาก Wisconsin Breast Cancer Database (WBCD). <http://www.archive.ics.uci.edu/ml/dataset.html>.

วีรานันท์ พงศาภักดี (2541). การวิเคราะห์ข้อมูลเชิงกลุ่ม : ทฤษฎีและการประยุกต์ (GLIM และ SPSS/FW).

พิมพ์ครั้งที่ 2. นครปฐม: ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

วิทยาเขตพระราชวังสนามจันทร์, 2541.

กัลยา วานิชย์บัญชา. (2551) การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร:

โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

กัลยา วานิชย์บัญชา. (2551) การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. พิมพ์ครั้งที่ 6.

กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี (2550) . การวิเคราะห์พหุระดับ (Multi-Level Analysis) . พิมพ์ครั้งที่ 4. กรุงเทพมหานคร:

โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย

Agresti, A., (2002). Categorical Data Analysis. 2nd rev-ed., New York: John Wiley & Sons, 2002.

Lachin JM. (2000). Biostatistical methods: the assessment of relative rate. New York: Wiley-

Interscience,

Wallenstein S, Bodian C.. (1987). Inferences on odds ratios, relative risks, and risk differences

based on standard regression programs. Am J Epidemiol 1987;126:346-55.

Wacholder S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences.

Am J Epidemiol 1986;123:174-84.

Rahman, Mezbuhar, Cortes, Judy and Pardis, Cyrus (2001), A Note on Logistic Regression,

Journal of Statistics & Management Systems, Vol. 4, No.2 p 175-187

Jiahe Qian, (2003), Application of Logistic Regression in Analysis of e-rater Data, Education

Testing Service, ETS MS 02-T, Princeton, NJ 08541