The Capstone Committee for Efstathia Polychronopoulou Certifies that this is the approved version of the following capstone:

Data Mining for Signal Detection of Targeted Therapy Related Drug Toxicity in Breast Cancer Patients

Committee:

Kuo Yong-Fang, PhD, Supervisor

Giordano Sharon, MD

Yu Xiaoying, PhD

Dean, Graduate School

Data mining for signal detection of targeted therapy related drug toxicity in breast cancer patients

by

Efstathia Polychronopoulou, MSc

Capstone

Presented to the Faculty of the Graduate School of
The University of Texas Medical Branch
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Public Health

The University of Texas Medical Branch
July, 2019

Acknowledgements

I am grateful to all those with whom I had the pleasure to collaborate during this project. Foremost, I would like to express my sincere gratitude to Dr. Yong-Fang Kuo, my supervisor and committee chair, for her counsel throughout this project and unwavering support during my MPH studies.

I would also like to thank Dr. Sharon Giordano and Dr. Xiaoying Yu, my committee members, for their insightful comments that were instrumental in the completion of this thesis. I would also like to thank Ms Lin-Na Chou for her guidance and assistance with the computational aspects of this project.

I would like to extend my appreciation to Dr Cara Pennel and the faculty of the Department of Preventive Medicine and Community Health for their teaching and research support throughout the years.

Finally, deepest gratitude to my husband for the continuous encouragement and inspiration.

Data mining for signal detection of targeted therapy related drug toxicity in breast cancer patients

Publication No._____

Efstathia Polychronopoulou, MPH

The University of Texas Medical Branch, 2019

Supervisor: Yong-Fang Kuo

Abstract

Application of signal detection methods using claims data can improve post-marketing drug surveillance. The aim of this study is to compare two routinely used approaches, the proportional reporting ratio (PRR) and Gamma Poisson Shrinker (GPS) with a tree-based scan statistic (TBSS). Using data from the Texas Cancer Registry and Surveillance, Epidemiology and End Results linked to Medicare from 2010-2014 we identified 8,949 patients with breast cancer treated with chemotherapy and 2,542 patients treated with trastuzumab in addition to chemotherapy. Inpatient and outpatient visits up to 1 year from start of therapy were used to identify adverse events (AEs). For each method two signaling thresholds were evaluated. Across all methods we found a total of 34 signals associated with use of trastuzumab. Clinical review determined that most identified signals represented known AEs or confounding. GPS on the highest signaling threshold failed to detect a well-established AE when time of follow-up was less than 6 months. Overall there was considerable agreement between methods with GPS being the most conservative. PRR and TBSS may be more appropriate in exploratory drug safety studies using this dataset.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ADE	Adverse Drug Event
WHO	World Health Organization
FDA	Food and Drug Administration
DMM	Data Mining Methods
SRS	Spontaneous Reporting System
PRR	Proportional Reporting Ratio
GPS	Gamma Poisson Shrinker
TBSS	Tree-based Scan Statistic
BC	Breast Cancer
HER2	Human Epidermal Growth Factor Receptor 2
TCR	Texas Cancer Registry
SEER	Surveillance, Epidemiology, and End Results
CCI	Charlson Comorbidity Index
ICD-9	International Classification of Diseases 9th revision
CCS	Clinical Classification System
RR	Relative Risk
CI	Confidence Interval

Chapter 1 Introduction

It has been estimated that adverse drug events (ADEs) occur in about 2 - 7% of annual hospitalizations in the US alone and can lead to increased morbidity, mortality and associated costs[1,2]; thus early detection of unexpected ADEs is of paramount importance for global health. The World Health Organization defines drug safety surveillance or pharmacovigilance as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems”[3]. Drug safety research for medical products is not limited to the later stages of clinical trials; instead it begins at the pre-clinical stage and through phase I – III clinical trials. Drug safety monitoring then continues in the post-approval phase, either by mandated phase IV trials or by adverse event reporting throughout a drug’s life.

Initially post-approval drug surveillance efforts relied on clinical review of adverse event reports collected by organizations such as WHO or the FDA. However, due to the complexity and volume of data that needed to be reviewed, new methods were required to assist in this process. Data mining methods (DMM) are automated computational methods designed to detect drug - event pairs with higher than expected reports that warrant further investigation. A number of DMM have been developed and implemented in drug safety monitoring over the past years[4].

Currently Spontaneous Reporting Systems (SRS) are the main pharmacovigilance data source. The most distinguished SRS in the US is the

Adverse Event Reporting System maintained by the FDA. SRS include mandatory reports from pharmaceutical companies and voluntary reports form health care providers and consumers. While SRS are very useful tools for drug surveillance since they include all marketed drugs and broad patient populations, they also have several limitations. The main concern with SRS is reporting bias; drugs with known ADEs are more likely to be reported, while others are underreported. Other issues with SRS arise from misattributed drug-event pairs, incomplete reporting, and multiple duplicate reports for the same case[5].

In an effort to improve current practice and establish an active surveillance system, the complementary use of secondary data sources has been proposed; electronic healthcare and claims data are not limited to patients with ADEs and include more complete patient information. However, validation of DMM in secondary observational data has been very limited[6–8]. As potential pharmacovigilance data sources are expanding, it becomes imperative to evaluate and refine the current DMM and to design new methods to approach non-traditional drug safety monitoring sources.

The present study aims to compare three data mining methods in the detection of adverse events related with use of trastuzumab (branded name: Herceptin) in a cohort of elderly breast cancer patients using an administrative claims dataset. These methods are the Proportional Reporting Ratio (PRR), the Gamma Poisson Shrinker method (GPS) and the Tree-based Scan Statistic (TBSS). The first two methods are commonly used disproportionality analysis measures; PRR is a frequentist and GPS a Bayesian approach. Currently, the FDA

and the European Medicines Agency use GPS and PRR respectively to detect signals of ADEs[9,10]. TBSS is a relatively new method that can simultaneously evaluate signals at different levels of specificity. Trastuzumab is a targeted breast cancer therapy and was chosen for this study as it has been on the market for more than a decade and therefore related ADEs are fairly well established[11,12].

Chapter 2 Methods

Study design

This study uses a retrospective cohort design. Subjects are categorized as “exposed” (taking the drug of interest) or “not exposed” (not taking the drug of interest). An association between drug exposure and the outcome (adverse events) is established when the outcome occurs with a higher frequency in the exposed than the not exposed group. In this study however there is not a specific outcome of interest; all possible adverse events are considered.

Two approaches are explored for adverse event definition; subject and visit level. The subject level approach utilizes the advantage of observational health data to provide patient-specific information and eliminates any potential bias from repeated visits for the same condition. The visit level approach “resembles” the current SRS approach, as each visit is treated as a distinct spontaneous case. In this sense all occurrences of an ADE are counted (prevalent conditions) in the visit level approach, whereas the subject level approach considers the first occurrence of each ADE only (incident conditions). Zorych et al. (2011) were the first to introduce and evaluate these concepts for longitudinal observational datasets[13].

The aim of the study is to compare different data mining methods to detect signals of adverse events attributed to trastuzumab. The study population is elderly women diagnosed with Breast Cancer (BC) between 2010 – 2014 who received trastuzumab in addition to chemotherapy or received chemotherapy alone. trastuzumab is a targeted therapy for patients with HER2 positive BC. HER2

negative BC patients are typically treated with chemotherapy alone. There is no conclusive evidence in the literature that development of HER2 positive BC versus HER2 negative BC depends on specific patient characteristics, like race, socioeconomic status and place of residence or on prior use of hormonal treatments. In that sense, whether a patient develops HER2 positive or HER2 negative cancer and the subsequent treatment allocation can be considered as natural randomization.

Data source

The linked Texas Cancer Registry (TCR) and Surveillance, Epidemiology, and End Results (SEER) Medicare linkage databases were combined for use in this study. The SEER program, supported by the National Cancer Institute, has been collecting information on newly diagnosed cancer cases in SEER registry areas since 1973. Currently, it is estimated that SEER covers approximately 34.6% of the US population; 32% of Whites, 30% of African Americans, 44% of Hispanics, 49% of American Indians and Alaska Natives, 69% of Pacific Islanders and 58% of Asians[14]. The TCR program was initiated in 1976 and follows the same collection and reporting requirements as SEER. Approximately 120,000 new cases are reported annually in TCR and about 17,000 are breast cancer cases[15]. The registries collect information on patient demographic characteristics, cancer incidence, stage of disease, course of therapy, and survival.

Medicare is the primary health insurance of approximately 98% of the elderly US population[16]. Medicare data include information on hospital admissions,

billed physician services, prescription information and outpatient visits. The linkage of the Medicare and cancer registries is based on matching a person's social security number, name, date of birth and sex. Linkage algorithms are developed by the collaboration of the National Cancer Institute and Medicaid Services.

Cohort selection

This study included all patients aged ≥ 66 years diagnosed with HER-2 positive breast cancer who received trastuzumab in addition to standard chemotherapy and patients with HER-2 negative cancer who received standard chemotherapy alone, between 2010 and 2014.

Exclusion criteria: Patients whose cancer diagnosis originated from an autopsy or death certificate and was not confirmed clinically were excluded. Also, patients were excluded if they did not have continuous enrollment in Medicare Part A or Part B (or were enrolled in HMO) for at least 12 months before diagnosis or if their HER2 status was unknown. Table 1 summarizes the cohort selection flow that was followed.

Inclusion	Exclusion
Breast cancer diagnosis between 2010-2014	Diagnosis before 2010 or after 2014
Diagnosis confirmed clinically	Diagnosis from death certificate / autopsy / other non-confirmed method
Age > or equal to 66 years	Age < 66 years
Treatment with trastuzumab and chemotherapy for HER2+ or chemotherapy alone for HER2-	Treatment with other therapy or unknown HER2 status
Medicare A or B 1 year prior	Enrollment in HMO or non-continuous enrollment in Medicare A or B for 1 year before diagnosis

Table 1: Cohort selection for patients with breast cancer

Study variables

Subjects were categorized in two groups based on their HER2 and treatment status. HER2 information has been collected from SEER and TCR since 2010. Treatment with chemotherapy and/or trastuzumab was determined from Medicare Data using International Classification of Diseases 9th revision (ICD-9) codes from the outpatient (OUTSAF), carrier (NCH), Durable Medical Equipment (DME) and Part D Prescriber Public Use files as previously described[17]. Other variables included in the analysis are shown in Table 2.

Variable	Data Source	Definition
Age	SEER, TCR	Age at time of diagnosis
Race	SEER, TCR	White, Black, Other
Charlson Comorbidity Index (CCI)	Medicare (MEDPAR, OUTSAF, NCH files)	The Charlson Index is derived from an algorithm based on a count of certain comorbid diagnoses in the year before BC diagnosis[18,19].
Breast Cancer stage	SEER, TCR	BC stage at diagnosis: In situ, Localized, Regional, Distant, Unstaged
BC Surgery	SEER, TCR	Received BC surgery in the first course of treatment: Yes, No
Radiation treatment	SEER, TCR	Received radiation in the first course of treatment: Yes, No

Table 2: Independent variables

Outcome: Post-treatment inpatient and outpatient claims from Medicare files were used to determine adverse effects. For inpatient claims, all diagnoses listed in the MEDPAR file were considered, whereas for outpatient claims only the primary encounter diagnosis listed in OUTSAF, NCH or DME files was used.

Diagnoses in the Medicare files are provided in the form of ICD-9 codes. Claims associated with radiation therapy, BC related surgery, administration of chemotherapy, injuries, congenital diseases or other unrelated health factors were excluded (see Appendix A). Separate analyses were conducted for several follow-up times; 3 months, 6 months and 1-year post-treatment initiation were examined.

Two approaches were used for counting adverse events; the subject and the visit level approach. In the subject level approach, for each potential adverse event the number of subjects who had at least one claim during the follow-up period were counted. In the visit level approach, for each potential adverse event the number of visits during the follow-up period were counted, allowing each subject to contribute more than one visit. Distinct visits for the same subject with the same diagnosis were defined as having either different date (even if consecutive dates) or different provider.

To organize the multitude of ICD-9 diagnosis codes - over 14,000 - into clinically meaningful categories the Clinical Classification Software (CCS) was used. The CCS is a categorization scheme that was developed by the Agency for Healthcare Research and Quality to group related ICD-9 codes[20]. The CCS also employs a hierarchical system with four levels. The first level consists of 17 categories of body systems. These categories are split further at each higher level, becoming more specific; level four is the most granular level. Each ICD-9 code is mapped into one category only, but each category can consist of several codes.

Data mining methods

Relative Risk and Proportional Reporting Ratio

The Relative Risk (RR) is a measure of the probability of an event occurring in the exposed group over the probability of the same event occurring in the non-exposed group. The Proportional Reporting Ratio (PRR) is the pharmacovigilancy equivalent of the Relative Risk (RR). It is defined as the ratio of the probability of a specific adverse event given the drug of interest over the probability of the same event under the comparator drug. Routinely, a cut-off value of 2 is used to identify signals of higher than expected frequencies of adverse events[10,21,22]. In this study a cut-off value of 2 with a lower 95% confidence interval of at least 1 and at least 3 cases was used to indicate a signal. Also as a stricter signaling threshold, a PRR value of 2 with a lower 95% CI of 1.5 and at least 3 cases was also considered.

	Adverse Event	All other events	
Drug of interest	A	B	A + B
Other drugs	C	D	C + D
	A + C	B + D	Total

Table 3: Calculation of RR and PRR

Following the notation from Table 3 the formula for PRR (or RR) is:

$$PRR = \frac{A/(A+B)}{C/(C+D)}$$

The distribution of PRR (or RR) is non-normal, but the natural logarithm transformed distribution is approximately normal, therefore the 95% confidence interval (CI) for PRR (or RR) is given by the formula:

$$CI = PRR \cdot e^{\pm 1.96 \cdot s}$$

where $$s = sqrt\ (\frac{1}{A} + \frac{1}{C} - \frac{1}{A+B} - \frac{1}{C+D})$$

The distinction between the two measures in the present study lies in the fact that the denominator for RR refers to the non-exposed group, including subjects without any claims during the study period, whereas the denominator for PRR refers to all other ADEs excluding subjects without claims. In cases where all subjects have at least one claim, the population contributing to RR and PRR will be the same. In this study RR is used in the subject level approach and PRR in the visit level approach.

Gamma Poisson Shrinker

The Gamma Poisson Shrinker (GPS) method, described by DuMouchel[23], assumes that the observed count for each drug-event pair is a random variable that follows the Poisson distribution. Specifically, let Y_{ij} denote the observed count for drug i and adverse event j, where i=1, ..,I and j = 1, … ,J. Then $Y_{ij} \sim$ Pois (μ_{ij}) distribution, where

$\mu_{ij} = \lambda ij * E [Y_{ij}]$

and E [Y_{ij}] is the expected number of events under the null. Expected count is calculated from the marginal counts of adverse event j in both groups and drug i, assuming independence between event and exposure. The method assumes that λ_{ij}'s follow a prior gamma distribution and uses an Empirical Bayesian approach to maximize likelihood and determine the posterior distributions of λ_{ij} 's. GPS then uses the Empirical Bayesian Geometric Mean (EBGM) of the posterior distribution in lieu of the risk ratio. The one sided 95% CI is conventionally used when implementing GPS. The FDA uses a cutoff value of 2 for the lower bound of the one-sided 95% EBGM confidence interval to identify signals on their spontaneous reporting system[24]. In this study a less stringent threshold of EBGM >= 1.5 and a one sided 95% lower CI > 1 was also considered for signal detection.

In practice, the GPS method is designed to avoid spurious false positives that may arise due to small counts, by utilizing a "shrinkage" factor. Figure 1, illustrates the shrinkage effect of GPS compared to using RR[25]. Each point on the figure is a drug-outcome pair; the x-axis consists of the RR estimate on the log scale and the y-axis of the log EBGM estimate. We note that when there is a considerable number of reports for a particular drug-event combination (> 50) the results of the two methods are similar, whereas in cases with few reports GPS estimates "shrink" substantially.

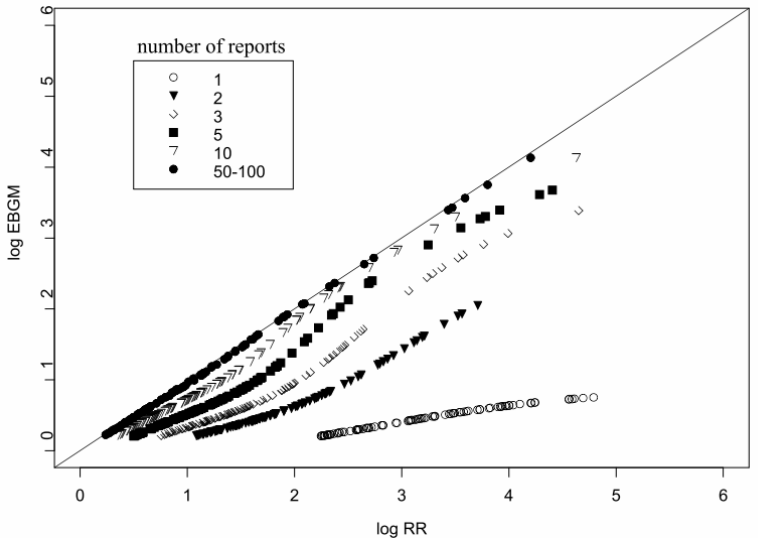

Figure 1: GPS shrinkage in FDA's spontaneous reporting database, retrieved by Madigan et al. (2010).

Tree-based Scan Statistic

In the Tree-based Scan Statistic (TBSS) method, developed by Kulldorff et al.[26], adverse events are classified as a hierarchical tree structure, where related diagnoses are closer together on the tree. For this study the hierarchical structure of diagnosis groups is provided by the Clinical Classification Software used to categorize ICD-9 codes. While the previous methods evaluate each level of diagnoses separately, TBSS evaluates individual and closely related adverse events simultaneously without the need to pre-specify the level of granularity. TBSS evaluates all possible "cuts" on the tree, while formally adjusting for multiple testing of overlapping diagnoses.

Under the null hypothesis, an adverse event is equally likely to occur anywhere in the tree. The alternative hypothesis is that there is at least one branch

of the tree where an adverse event is more likely to occur. The number of events at each node of the tree is assumed to follow a Poisson distribution, where the population of each node is determined by the expected number of adverse events under the null. The method then calculates the likelihood of an event under both the null and the alternative hypotheses for each branch of the tree. The “cut” with the maximum log-likelihood ratio is the one least likely to have occurred by chance. To infer on the statistical significance of the cut’s likelihood ratio, the method conducts random Monte Carlo simulations. The p-value is then calculated as:

$$p= R/(S+1),$$

where R is the rank of the maximum log-likelihood ratio from the original data set compared to the simulated data sets and S in the number of Monte Carlo simulations. A cut-off value of $p < 0.05$ is used in this study to identify signals of excess risk. A cut off-value of $p <0.001$ is considered as the stricter signaling threshold.

Data management

During this study, all analyses were conducted on computers provided by the Rehabilitation Sciences Division and the Office of Biostatistics at the University of Texas Medical Branch. All data management and analyses were conducted in full compliance with HIPPA regulations and data user agreements. Data management and analysis were conducted using SAS 9.4 and R version 3.4.4 (openEBGM package). For TBSS analysis the Tree Scan Software v 1.4 was used.

Chapter 3 Results

The cohort selection process of the study populations from SEER and TCR is summarized in Figures 2 and 3 respectively. Although it is not possible to identify any individuals that may be overlapping between the two datasets, this is highly unlikely to occur as the SEER registry does not cover the Texas population. Of patients who met the overall selection criteria, 8,949 (78%) were in the chemotherapy only group and 2,542 (22%) were in the chemotherapy plus trastuzumab group. 57.8% of total subjects had continuous enrollment at 3 months from the start of treatment, 54.9% at 6 months and 47.3% at 12 months. Table 4 shows the distribution of continuously enrolled subjects per group and Table 5 the mortality distribution. Tables 6 – 8 summarize the characteristics of the two groups at 3, 6 and 12 months from the start of treatment respectively.

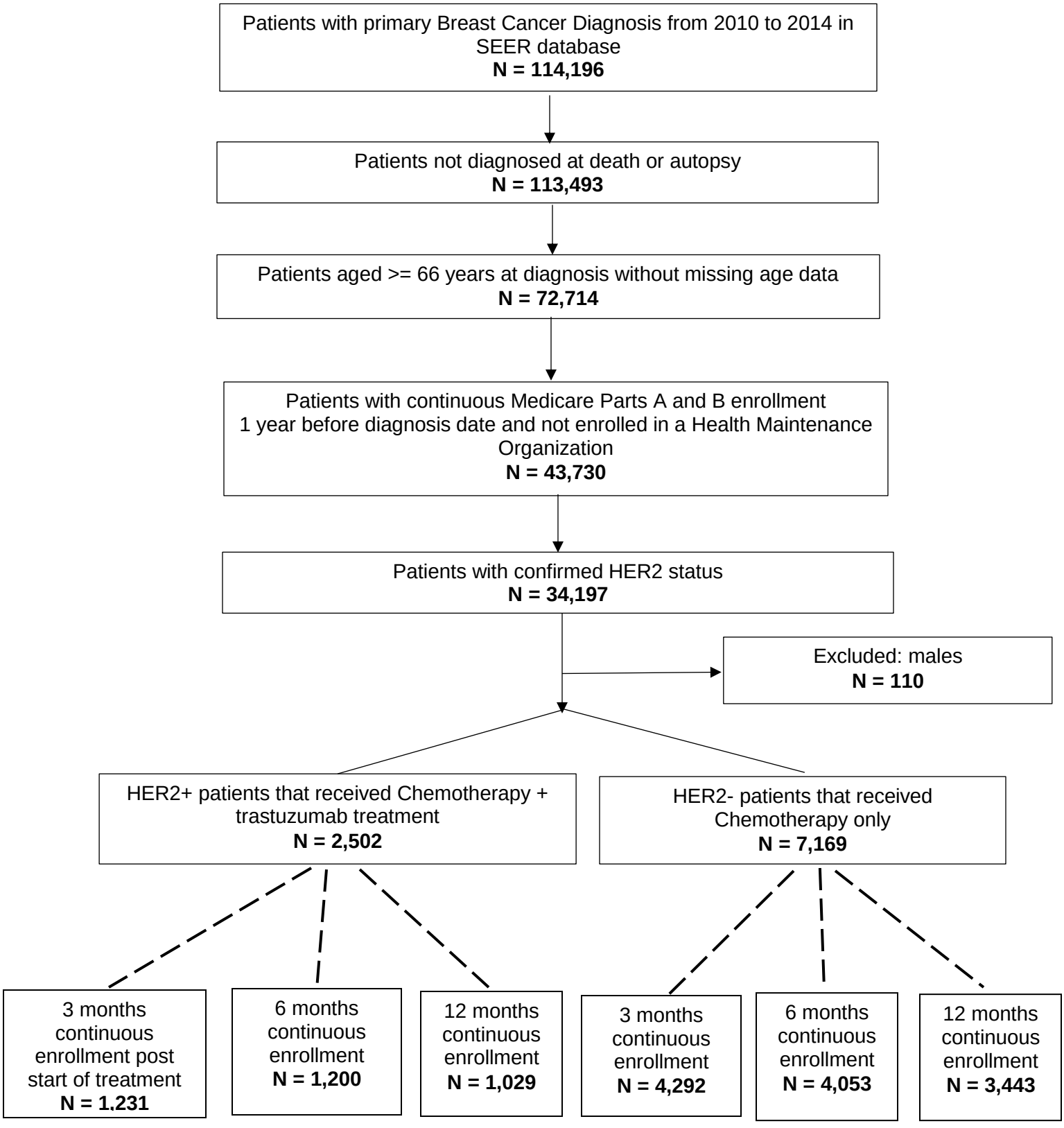

Figure 2: SEER cohort selection flowchart

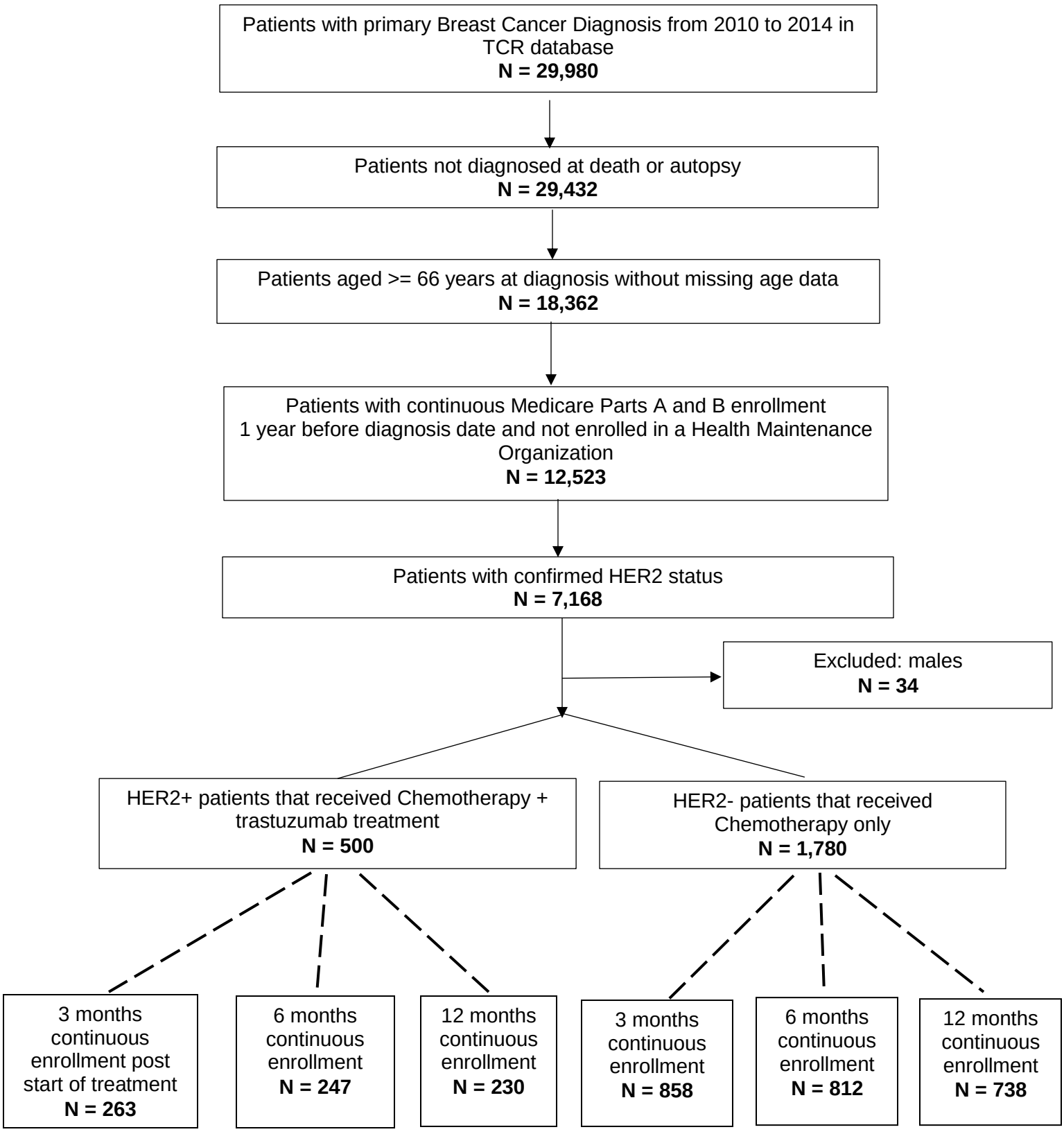

Figure 3: TCR cohort selection flowchart

Continuous ABD enrollment after first treatment			
	N (%) HER2+	**N (%) HER2-**	**p-value***
up to 3 months	1,494 (58.8)	5,150 (57.6)	0.27
up to 6 months	1,447 (56.9)	4,865 (54.4)	**0.0221***
up to 12 months	1,259 (49.5)	4,181 (46.7)	**0.0124***

*Accepted significance level: p-value < 0.05

Table 4: Continuous enrollment for the two groups after start of treatment

Continuous enrollment up to 3 months from the start of treatment was similar between the two groups. At 6 and 12 months a significantly higher proportion of the HER2+ group remained enrolled.

Percent of patients who died after first treatment			
	N (%) HER2+	**N (%) HER2-**	**p-value***
up to 3 months	57 (2.2)	246 (2.8)	0.15
up to 6 months	107 (4.2)	450 (5.1)	0.07
up to 12 months	182 (7.2)	795 (8.9)	**0.0057***

*Accepted significance level: p-value < 0.05

Table 5: Cumulative percent of patients that died during follow-up

To determine whether these differences in enrollment could be attributed to differential overtime mortality between the groups, the percentage of patients (from the overall population) who died at each time period was calculated. At 3 and 6 months, the percentage of patients was similar between the groups. At 1 year from the first dose of treatment, the percentage of patients who died in the HER2- group was 1.7% higher than the HER2+ group accounting perhaps for the difference in enrollment continuity.

3 months		Chemotherapy + trastuzumab (N=1494)	Chemotherapy only (N=5150)	p-value*
		Mean (st. dev.)		
Age at diagnosis		73.3 (6.2)	73.2 (6.2)	0.4473
Charlson Comorbidity Index**		0.82 (1.3)	0.87 (1.3)	0.1768
		Count (%)		
Race	White	1288 (86.2)	4410 (85.6)	**<0.001***
	Black	111 (7.4)	524 (10.1)	
	Other	95 (6.6)	216 (4.2)	
Radiation	Yes	576 (38.6)	2244 (43.6)	**0.0021***
	No	819 (54.8)	2570 (49.9)	
	Unknown	99 (6.6)	336 (6.5)	
BC Surgery	Yes	1273 (85.2)	4447 (86.3)	0.2021
	No	216 (14.5)	673 (13.1)	
	Unknown	5 (0.33)	30 (0.6)	
Stage at diagnosis	In situ	3 (0.2)	19 (0.4)	**0.0317***
	Localized	714 (47.8)	2312 (44.9)	
	Regional	550 (36.8)	2118 (41.1)	
	Distant	213 (14.3)	656 (12.7)	
	Unstaged	14 (1)	45 (0.9)	

**All comorbidities that consist the Charlson Index were tested separately and no stat. sign. differences between groups were found

*Accepted significance level: p-value < 0.05

Table 6: Patient characteristics at 3 months

The two groups at 3 and 6 months were similar in terms of age at time of diagnosis, the number of pre-existing comorbidities indicated by the Charlson Comorbidity Index and whether patients received surgical treatment. All comorbidities that constitute the Charlson Index were tested separately (not shown on Tables) and were not significantly different between groups. There was however a statistically significant difference in race distribution; there were more Black women in the HER2+ group (Chemotherapy + trastuzumab) compared to the HER2- group (Chemotherapy only). In terms of BC stage at diagnosis, the

HER2- group had a higher percentage of regional BC diagnoses. Also, in the HER2- group the proportion of women who received radiation treatment was significantly higher.

6 months		Chemotherapy + trastuzumab (N=1447)	Chemotherapy only (N=4865)	p-value*
		Mean (st. dev.)		
Age at diagnosis		73.2 (5.9)	73.1 (6)	0.4662
Charlson Comorbidity Index**		0.81 (1.2)	0.85 (1.3)	0.307
		Count (%)		
Race	White	1247 (86.2)	4180 (85.9)	<**0.001***
	Black	107 (7.4)	482 (9.9)	
	Other	93 (6.6)	203 (4.2)	
Radiation	Yes	569 (39.3)	2162 (44.4)	**0.021***
	No	782 (54)	2387 (49.1)	
	Unknown	96 (6.6)	316 (6.5)	
BC Surgery	Yes	1246 (86.1)	4253 (87.4)	0.119
	No	197 (13.6)	585 (12)	
	Unknown	4 (0.3)	27 (0.6)	
Stage at diagnosis	In situ	3 (0.2)	19 (0.4)	**0.005***
	Localized	701 (43.9)	2178 (44.9)	
	Regional	530 (36.7)	2040 (41.9)	
	Distant	199 (13.8)	589 (12.1)	
	Unstaged	14 (1)	39 (0.8)	

**All comorbidities that consist the Charlson Index were tested separately and no stat. sign. differences between groups were found

*Accepted significance level: p-value < 0.05

Table 7: Patient characteristics at 6 months

At one year, the demographic characteristics of the patients remained similar to 3 and 6 months (Table 8). However, among the diagnoses included in the Charlson Comorbidity Index there is a statistically significant difference in the proportion of patients with a prior diabetes diagnosis at 12 months (non-significant comorbidities not shown on Table 8). 23.3 % of the HER2+ group had diabetes compared to 26.4% in the HER2- group. The overall comorbidity index remained

comparable between the groups and no other comorbidity was found to differ significantly.

12 months		Chemotherapy + trastuzumab (N=1259)	Chemotherapy only (N=4181)	p-value*
		Mean (st. dev.)		
Age at diagnosis		73.1 (5.9)	72.8 (5.8)	0.0601
Charlson Comorbidity Index**		0.78 (1.2)	0.83 (1.2)	0.2076
		Count (%)		
Diabetes	No	965 (76.7)	3075 (73.6)	**0.0273***
	Yes	294 (23.3)	1106 (26.4)	
Race	White	1083 (86)	3602 (86.1)	**0.001***
	Black	93 (7.4)	401 (9.6)	
	Other	83 (6.6)	178 (4.3)	
Radiation	Yes	508 (40.4)	1899 (45.4)	**0.0048***
	No	671 (53.3)	2016 (48.2)	
	Unknown	80 (6.3)	266 (6.4)	
BC Surgery	Yes	1091 (86.6)	3702 (88.6)	0.0934
	No	164 (13)	458 (11)	
	Unknown	4 (0.3)	21 (0.5)	
Stage at diagnosis				**0.013***
	In situ	3 (0.2)	16 (0.4)	
	Localized	605 (48)	1867 (44.7)	
	Regional	463 (36.8)	1792 (42.9)	
	Distant	177 (14.1)	474 (11.3)	
	Unstaged	11 (0.9)	32 (0.8)	

**Among Charlson comorbidity index disorders, only diabetes had a significant p-value

*Accepted significance level: p-value < 0.05

Table 8: Patient characteristics at 12 months

Results from Data Mining Algorithms

I. Subject – level approach

Tables 9 - 11 present the signals detected by the 3 methods used for the subject – level analysis at 3, 6 and 12 months from start of treatment. Signals that meet the stricter signaling thresholds for each method are highlighted in red, signals meeting the lower thresholds highlighted in yellow. N refers to the number of cases for each group that were identified. For PRR and GPS, each level of diagnoses was analyzed separately, whereas the Tree Scan Method analyzes all levels concurrently, while adjusting for multiple testing. Because in this analysis each subject could contribute only once in each diagnosis group, the sum of the counts of the higher levels do not equal the count at the lower level; for instance, if a subject was diagnosed with hypertension (level 2 condition) as well as pericarditis (level 2) he will be counted as a case for each of these conditions separately, but will only contribute once for the lower level diagnosis group: Diseases of the circulatory system (level 1)

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
1	Infectious and parasitic diseases					
1.1.2.3	E. Coli septicemia	8	9	3.06		.
7	Diseases of the circulatory system	826	2432			1.11
7.2	Diseases of the heart	481	1290	.	.	1.19
7.2.1	Heart valve disorders	171	236	2.50	1.87	1.88
7.2.1.2	Non-rheumatic mitral valve disorders	88	107	2.83	1.92	2.01
7.2.1.3	Non-rheumatic aortic valve disorders	40	66	2.08	.	.
7.2.1.4	Other heart valve disorders	47	56	2.89	1.91	2.03
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	51	68	2.59	1.85	1.91
7.2.2.1	Cardiomyopathy	40	48	2.93	1.9	2.02
7.2.7	Other and ill-defined heart disease	34	40	2.93	.	2.04
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	8	8	2.76	.	.
9.4.2.2	Duodenal ulcer	8	7	3.94	.	.
9.5.3.8	Incisional hernia without obstruction/gangrene	3	0	24.11	.	.
12	Diseases of the skin and subcutaneous tissue					
12.1.1.1	Cellulitis and abscess of fingers and toes	18	30	2.07	.	.
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	8	10	2.76	.	.
13.6.2	Other acquired deformities	8	10	276	.	.
16	Injury and Poisoning					
16.11	Poisoning	39	55	2.44	.	.
16.11.2	Poisoning by other medications and drugs	37	54	2.36	.	.

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ‘ . ‘: no signal detected, red background: high signal threshold, yellow background: medium signal threshold

Table 9: Signals detected from subject level analysis at 3 months

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
2	Neoplasms					
2.6.1	Cancer of uterus	4	3	4.48	.	.
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders	322	882	.	.	1.19
4	Diseases of the blood and blood-forming organs					
4.4	Other hematologic conditions	9	12	2.52	.	.
7	Diseases of the circulatory system	1097	3094	.	.	1.20
7.2	Diseases of the heart	758	1675	.	.	1.35
7.2.1	Heart valve disorders	330	340	3.26	2.11	2.19
7.2.1.1	Chronic rheumatic disease of the heart valves	47	53	2.98	2.17	2.05
7.2.1.2	Non-rheumatic mitral valve disorders	181	147	4.14	2.21	2.41
7.2.1.3	Non-rheumatic aortic valve disorders	75	102	2.47	2.16	1.85
7.2.1.4	Other heart valve disorders	86	83	3.48	2.19	2.22
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	94	99	3.19	2.11	2.14
7.2.2.1	Cardiomyopathy	79	70	3.79	2.19	2.31
7.2.7	Other and ill-defined heart disease	74	74	3.36	2.11	2.18
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	9	9	3.36	.	.
12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	11	17	2.18		.
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis	9	10	3.03	.	.
16	Injury and Poisoning			.	.	1.19
16.10.2.2	Respiratory complications	4	3	4.48	.	.
16.11	Poisoning	53	72	2.47	.	.
16.11.2	Poisoning by other medications and drugs	50	69	2.44	1.94	1.83

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ‘ . ‘: no signal detected, red background: high signal threshold, yellow background: medium signal threshold

Table 10: Signals detected from subject level analysis at 6 months

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
2	Neoplasms					
2.6.1	Cancer of uterus	8	6	4.43	.	.
2.11.1	Cancer of head and neck	10	13	2.55	.	.
4	Diseases of the blood and blood-forming organs					
4.4	Other hematologic conditions	11	17	2.14	.	.
7	Diseases of the circulatory system	697	2130	.	.	1.16
7.2	Diseases of the heart	814	1849	.	.	1.31
7.2.1	Heart valve disorders	434	428	3.37	2.12	2.26
7.2.1.1	Chronic rheumatic disease of the heart valves	69	73	3.14	2.17	2.14
7.2.1.2	Non-rheumatic mitral valve disorders	251	178	4.68	2.36	2.57
7.2.1.3	Non-rheumatic aortic valve disorders	103	133	2.57	2.08	1.79
7.2.1.4	Other heart valve disorders	133	120	3.68	2.22	2.19
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	128	131	3.24	2.12	2.05
7.2.2.1	Cardiomyopathy	105	105	3.32	2.18	2.10
7.2.2.2	Other peri-; endo-; and myocarditis	25	31	2.68	.	.
7.2.7	Other and ill-defined heart disease	108	121	2.96	2.12	2.13
7.2.11.2	Heart failure	17	22	2.57	.	.
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	8	10	2.66	.	.
9.10.3	Gastroesophageal laceration syndrome	3	0	23.23	.	.
12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	18	21	2.85	.	.
16	Injury and Poisoning					
16.2.4.4	Unclassified fracture of lower limb	4	2	6.64	.	.
16.11	Poisoning					
16.11.2	Poisoning by other medications and drugs	58	94	2.05	.	.

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ‘ . ‘: no signal detected, red background: high signal threshold, yellow background: medium signal threshold

Table 11: Signals detected from subject level analysis at 12 months

At 3 months (Table 9) a total of 18 signals were detected across all methods among 6 different body systems. PRR detected 16 signals (9 at the stricter threshold). The GPS method detected five signals at the lowest signaling threshold. TBSS detected 8 signals (3 at the stricter threshold). TBSS was the only method that detected a signal at the less granular level (Diseases of the circulatory system – level 1). Both GPS and TBSS detected signals relating to heart disease only, while RR detected signals at other body systems as well. PRR is a method sensitive to small event counts (e.g. RR = 24.11 for incisional hernia, with total N = 3). Overall, 5/18 signals were detected by all three methods at 3 months.

Results are similar at 6 months (Table 10). 18 signals are detected, while agreement by all three methods is at 50% (9/18 signals). Heart related conditions are detected by all methods at the stricter signaling threshold. Poisoning by other medications is also detected by all methods. At 1 year (Table 11), all methods agree at 8/20 signals.

In regards to the clinical significance of the detected signals, cardiotoxicity is a well-established adverse event of trastuzumab[27,28]. Other less common adverse reactions, detected mainly by PRR, such as infection, hematologic conditions, gastrointestinal or musculoscelatal disorders have also been associated with trastuzumab use[12,29]. On the other hand some detected signals, like acquired deformities, may be false positives or associated with the underlying condition or other confounding factors (e.g. neoplasms, poisoning, fractures).

II. Visit – level analysis

Tables 12-14 present the results of the visit – level approach for each follow-up period. Similarly to the subject level analysis, the PRR and GPS analyze each diagnosis level separately – therefore the sum of the counts of higher level diagnoses may not necessarily equal to the count of the lower level. Instead, the TBSS method analyzes all levels at the same time.

At 3 months (Table 12) 30 signals were detected across all methods at 12 distinct body systems. The GPS method only detected signals - at the lower threshold - for heart conditions. The PRR and TBSS detected more conditions, but there was no agreement between them, except for the circulatory body system. Agreement across all methods was low, at 5/30 signals. Compared to the subject level approach at 3 months, the visit-level approach returned signals at 6 more body systems, several of which reflect adverse events likely associated with trastuzumab (anemia, nausea/vomiting, respiratory infections).

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
1	Infectious and parasitic diseases					
1.1.2.3	E. Coli septicemia	8	9	3.01	.	.
2	Neoplasms					
2.11.4	Cancer of brain and nervous system	7	3	8.14	.	.
2.11.5	Cancer of thyroid	3	1	10.46	.	.
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders	585	1555	.	.	1.83
4	Diseases of the blood and blood - forming organs					
4.1.3.2	Other deficiency anemia	39	61	2.17		.

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	21	30	2.44	.	.
6.8.1.2	Other otitis media and related conditions	13	19	2.32		.
6.9.2	Other central nervous system disorders	32	53	2.11	.	.
7	Diseases of the circulatory system	1876	5652	.	.	1.12
7.2	Diseases of the heart	1935	4879	.	.	1.12
7.2.1	Heart valve disorders	201	287	2.44	1.73	1.87
7.2.1.2	Nonrheumatic mitral valve disorders	99	120	2.8	1.87	2.06
7.2.1.4	Other heart valve disorders	51	68	2.54	1.85	1.95
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	65	90	2.52	1.79	1.91
7.2.2.1	Cardiomyopathy	51	62	2.79	1.87	2.06
7.2.7	Other and ill-defined heart disease	35	51	2.39	.	.
7.5.4	Other diseases of veins and lymphatics	316	969	.	.	1.36
8	Diseases of the respiratory system					
8.3.2.3	Other asthma with acute exacerbation	11	13	2.87	.	.
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	9	10	3.19	.	.
9.4.2.2	Duodenal ulcer	9	7	4.39	.	.
9.5.3.8	Incisional hernia without obstruction/gangrene	4	0	30.5	.	.
9.12	Other gastrointestinal disorders	353	949	.	.	1.29
9.12.3	Other and unspecified gastrointestinal disorders	229	597	.	.	1.35
12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	15	14	3.74	.	.
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	16	22	2.58	.	.
13.6.2	Other acquired deformities	11	14	2.74	.	.
16	Injury and Poisoning					
16.2.3.2	Fracture of radius and ulna	15	23	2.21	.	.
16.2.3.3	Other fracture of upper limb	6	4	5.09	.	.
16.11	Poisoning	43	63	2.42	.	1.85

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
16.11.2	Poisoning by other medications and drugs	41	61	2.34	.	.
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting	338	793	.	.	1.41
MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ' . ': no signal detected, red background: high signal threshold, yellow background: medium signal threshold						

Table 12: Signals detected from visit level analysis at 3 months

At 6 months (Table 13) results are similar with the 3 months analysis. 32 signals are detected overall, but agreement is higher at 34.4% (11/32 signals). Apart from the heart-related conditions, all methods agreed on anemia, otitis media and poisoning by other medications. Chronic skin ulcers were detected by PRR and TBSS but at different levels (level 3 and level 2 respectively).

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
2	Neoplasms					
2.11.4	Cancer of brain and nervous system	8	9	2.85	.	.
2.11.5	Cancer of thyroid	8	8	3.21	.	.
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders	585	1555	.	.	1.17
3.8.2	Hypovolemia	376	919	.	.	1.24
4	Diseases of the blood and blood-forming organs					
4.1.3.2	Other deficiency anemia	69	104	2.02	1.53	1.71
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	51	70	2.34	1.71	1.79
6.8.1.2	Other otitis media and related conditions	33	42	2.39	.	.

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
7	Diseases of the circulatory system	3806	10383	.	.	1.13
7.2	Diseases of the heart	1935	4879	.	.	1.19
7.2.1	Heart valve disorders	460	468	3.16	2.01	2.09
7.2.1.1	Chronic rheumatic disease of the heart valves	53	59	2.74	1.87	2.03
7.2.1.2	Nonrheumatic mitral valve disorders	227	172	4.02	2.16	2.44
7.2.1.3	Nonrheumatic aortic valve disorders	98	156	.	1.51	1.65
7.2.1.4	Other heart valve disorders	100	114	2.68	1.89	2.00
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	129	153	2.71	2.00	1.97
7.2.2.1	Cardiomyopathy	108	109	3.02	1.95	2.13
7.2.7	Other and ill-defined heart disease	129	153	2.76	2.00	1.98
8	Diseases of the respiratory system					
8.5.3	Empyema and pneumothorax	8	5	5.14	.	.
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	13	9	4.41	.	.
9.5.3.8	Incisional hernia without obstruction/gangrene	6	2	9.15	.	.
9.10.1	Hemorrhage from gastrointestinal ulcer	12	18	2.14	.	.
9.10.3	Gastroesophageal laceration syndrome	3	0	22.47	.	.
12	Diseases of the skin and subcutaneous tissue					
12.3	Chronic ulcer of skin	174	350	.	.	1.42
12.3.3	Other chronic skin ulcer	25	26	3.09	.	.
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	21	20	3.44	.	.
13.6.2	Other acquired deformities	20	30	2.14	.	.
16	Injury and Poisoning					
16.2.3.2	Fracture of radius and ulna	40	55	2.22	.	.
16.2.3.3	Other fracture of upper limb	11	15	2.24	.	.
16.2.5.4	Other and unspecified fracture	13	19	2.09	.	.
16.11	Poisoning	64	95	2.2	.	1.72
16.11.2	Poisoning by other medications and drugs	60	91	2.11	1.51	1.7

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting	338	793	.	.	1.28

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ' . ': no signal detected, red background: high signal threshold, yellow background: medium signal threshold

Table 13: Signals detected from visit level analysis at 6 months

Agreement at 12 months is at 35.3% (12/34) for several heart - related conditions, skin ulcer, otitis media and poisoning. Anemia was only detected by TBSS at 12 months. GPS and PRR also agreed on signals of duodenal ulcer and anal/rectal conditions. GPS at the higher signaling threshold however would detect only 4/12 signals.

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
2	Neoplasms					
2.6.1	Cancer of uterus	10	14	2.36		
2.11.1	Cancer of head and neck	12	17	2.34		
2.11.5	Cancer of thyroid	13	14	3.07		
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders	676	1762	.	.	1.20
3.8.2	Hypovolemia	407	957	.	.	1.30
4	Diseases of the blood and blood-forming organs					
4.1	Anemia	1015	2754	.	.	1.16
4.1.3	Deficiency and other anemia	970	2651	.	.	1.16
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	89	130	2.27	1.87	1.76
6.8.1.2	Other otitis media and related conditions	65	83	2.44	1.86	1.91

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
7	Diseases of the circulatory system	697	2130	.	.	1.12
7.2	Diseases of the heart	3072	7746	.	.	1.20
7.2.1	Heart valve disorders	796	676	3.9	2.28	2.31
7.2.1.1	Chronic rheumatic disease of the heart valves	91	84	3.38	2.12	2.26
7.2.1.2	Nonrheumatic mitral valve disorders	397	233	5.32	2.51	2.74
7.2.1.3	Nonrheumatic aortic valve disorders	164	239	2.14	1.75	1.77
7.2.1.4	Other heart valve disorders	167	165	3.16	2.07	2.19
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	263	254	3.43	2.16	2.20
7.2.2.1	Cardiomyopathy	230	206	3.48	2.15	2.29
7.2.2.2	Other peri-; endo-;and myocarditis	34	52	2.04	.	.
7.2.7	Other and ill-defined heart disease	154	164	3.11	2.10	2.11
7.2.11.2	Heart failure	26	28	2.90	1.64	.
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	19	27	2.37	.	.
9.4.2.2	Duodenal ulcer	20	10	6.24	2.11	.
9.6.5	Anal and rectal conditions	38	50	2.51	1.66	.
9.10.1	Hemorrhage from gastrointestinal ulcer	12	19	2.09	.	.
9.10.3	Gastroesophageal laceration syndrome	4	0	29.78	.	.
12	Diseases of the skin and subcutaneous tissue					
12.3	Chronic ulcer of skin	255	527	.	.	1.46
12.3.3	Other chronic skin ulcer	66	48	4.55	2.20	2.52
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	37	38	3.27	.	.
16	Injury and Poisoning			.	.	1.18
16.2.4.4	Unclassified fracture of lower limb	5	2	7.80	.	.
16.11	Poisoning	96	132	2.44	.	1.82
16.11.2	Poisoning by other medications and drugs	93	121	2.54	2.02	1.89

MLCCS	Diagnosis	N HER2+	N Her2-	PRR	EBGM	Tree-Scan RR
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting	359	854	.	.	1.29

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Reporting Ratio, EBGM: Empirical Bayesian Geometric Mean, Tree-scan RR: relative risk estimated by tree scan, ' . ': no signal detected, red background: high signal threshold, yellow background: medium signal threshold

Table 14: Signals detected from visit level analysis at 12 months

Tables 15 and 16 present a summary of the signals that were detected by all methods from the subject and the visit level analyses respectively for each follow up period. The PRR (or RR) was the method that consistently produced higher signals. Among the two approaches, signal ranges did not differ substantially. Mitral valve disorders and cardiomyopathy were the conditions that had the highest signals on average.

Adverse Events	3 months	6 months	12 months
Heart valve disorders	1.9 - 2.5	2.1 - 3.3	2.1 - 3.4
Chronic rheumatic valve disorders	.	2.1 - 3.0	2.2 - 3.1
Mitral valve disorders	1.9 - 2.8	2.2 - 4.1	2.4 - 4.7
Aortic valve disorders	.	1.9 - 2.5	1.8 - 2.6
Other heart valve disorders	1.9 - 2.9	2.2 - 3.5	2.2 - 2.7
Other heart disease	.	2.1 - 3.4	2.1 - 3.0
Peri-, endo-, myo- carditis	1.9 - 2.6	2.1 - 3.2	2.1 - 3.1
Cardiomyopathy	1.9 - 2.9	2.2 - 3.8	2.1 - 3.3
Poisoning by other medication	.	1.8 - 2.4	.

Table 15: Signal range for conditions detected by all methods from subject-level approach

Adverse Events	3 months	6 months	12 months
Heart valve disorders	1.7 - 2.4	2.0 - 3.2	2.3 - 3.9
Chronic rheumatic valve disorders	.	1.9 - 2.8	2.1 - 3.4
Mitral valve disorders	1.9 - 2.8	2.2 - 4.0	2.6 - 5.3
Aortic valve disorders	.	.	1.8 - 2.1
Other heart valve disorders	1.9 - 2.5	1.9 - 2.7	2.1 - 3.2
Other heart disease	.	2.0 - 2.8	2.1 - 3.1
Peri-, endo-, myo- carditis	1.8 - 2.5	2.0 - 2.7	2.2 - 3.4
Cardiomyopathy	1.9 - 2.8	2.0 - 3.0	2.2 - 3.5
Anemia	.	1.5 - 2.0	.
Otitis media/related complications	.	1.7 - 2.3	1.8 - 2.3
Chronic skin ulcer	.	.	2.2 - 4.6
Poisoning by other medication	.	1.5 - 2.1	1.9 - 2.5

Table 16: Signal range for conditions detected by all methods from visit-level approach

Chapter 4 Summary and Conclusions

Observational health databases have the potential to complement current pharmacovigilance practice by addressing inherent problems, such as underreporting and lack of proper control group. In this study we used an administrative claims dataset to evaluate the performance of three different data mining algorithms; the Proportional Reporting Ratio (and its equivalent Relative Risk), the Gamma Poisson Shrinker and the Tree Scan Statistic.

Previous studies have mainly focused on applying data mining methods on spontaneous reporting databases[21,22,24,30]. Brown et al. were the first to compare the GPS and TBSS methods using health plan data, while Curtis et al. demonstrated the feasibility of GPS when applied in claims data[6,7]. Two more studies that utilized observational health datasets evaluated the use of TBSS alone[8,26]. In these studies the unit of analysis was at the subject level; our study also assesses the use of this methodology at the visit level. Zorych et al. where the first to use PRR, GPS and other disproportionality methods to evaluate different approaches to event counting[13].

We detected between 16 – 34 signals across all methods depending on approach and follow-up period. These numbers may overestimate the adverse events related with trastuzumab, since every detected signal at each level was counted as unique. Agreement between all methods ranged from 17% to 50%. Brown et al. have reported agreement between GPS and TBSS that ranged from 32-35% based on the choice of signaling threshold[6]. In terms of variation of the

detected signals, PRR was the least robust method especially with small counts (Appendix B). Also, the range of the detected confidence intervals was considerably larger with PRR and TBSS compared to GPS, despite the fact that the calculated CIs for the latter were one-sided and two-sided for the former. As expected, Bayesian “shrinkage” resulted in GPS being the more conservative, yet more stable method.

There is no “golden standard” in regard to signaling thresholds in data mining. Not only is there an inherent “trade-off” between sensitivity and precision with the choice of signaling criteria, but also these measures may vary considerably among different datasets[21]. Our approach in this study was to use thresholds utilized by the FDA or pharmaceutical companies, as well as less strict thresholds, considering the lack of prior data mining research with this dataset. Therefore, we evaluated two signaling thresholds for each method. GPS was the method less likely to detect a signal at the higher threshold, especially when the event counts were low. In the study of Curtis et al., that also utilized claims data, none of the GPS detected signals had a value > 2, which was the cut-off we used for the stricter threshold[7]. This is consistent with the theoretical basis of the method and also an indication that more flexible criteria may be appropriate in implementing GPS with Medicare claims data. PRR was sensitive to small event counts, but whether these findings represent true rare adverse events or “false positives” needs to be evaluated clinically prior to determining the appropriateness of the signaling criteria used. TBSS was more likely than the other two methods to detect signals at a less granular level; it would therefore be more appropriate to be

used as a first “screening” method to determine body systems more likely to be affected by the drug of interest.

In this study two approaches were evaluated with different units of analysis-subject and visit. Results were generally consistent between the two methodologies, although in the visit-level approach more signals pertaining to a higher number of body systems were detected. Although this could be the reflection of the “artificial” inflation of count numbers, as each subject may contribute more than once for each diagnosis, several of the newly detected signals are indeed associated with trastuzumab use. These results are consistent with the study of Zorych et al. who showed that a distinct-patient approach to longitudinal data had worse performance than an SRS-like approach[13]. The visit level approach may therefore be more appropriate when evaluating rare, but potentially severe, adverse events.

Cardiotoxicity is the most well-established severe ADE of trastuzumab in breast cancer patients, more commonly manifested as cardiomyopathy[27,28,31]. In this study this known ADE was detected by all methods and approaches, albeit not always at the stricter signaling threshold. This finding suggests that claims data can be a credible supplement to the current drug safety surveillance practice.

Heart valve disorders, particularly mitral, were also consistently detected in this study. The risk of mitral valve disease in the trastuzumab group was found to be between two to five times as high as the non-exposed group, depending on the method and follow-up period. A prior population-level study did not find any association between heart valve disease and trastuzumab[28]. Instead, valvular

disease in breast cancer patients has been reported as an adverse effect of radiation therapy[32]. In this study however, the percent of patients that received radiation therapy was higher in the non-exposed group. Whether the risk of heart valve disease increases in patients that receive radiation therapy plus trastuzumab is unclear. Further study is needed to evaluate whether this finding may be a true ADE of trastuzumab, the result of unadjusted confounding factors, or a reflection of increased cardiac monitoring in the exposed group.

Signals detected in body systems other than the circulatory were not as consistent in this study. However, the majority of those signals are associated with use of trastuzumab; chronic skin ulcers or mouth sores have been reported after prolonged trastuzumab use, as well as anemia, gastrointestinal disorders, arthralgia, infections or electrolyte disorders[12,29,33]. Otitis media has not been reported previously as an ADE of trastuzumab, it could however occur as a complication of other respiratory infections that are known to be increased with this drug. We also found signals for poisoning by other medications in the exposed group. This is unlikely to be associated with trastuzumab; it is unclear whether this group of ICD-9 codes “captures” true adverse events only or also includes cases of medication misuse.

One limitation of this study arises from the use of administrative claims data and the fact that billing/coding errors or discrepancies cannot be accounted for; the same disease may be coded with different ICD-9 codes among different providers. In addition, the use of CCS to classify and group related diagnoses may not be the most appropriate for this study. Also, the sample size may have been

small for some rare ADEs to be detected. Due to the relatively small sample size, we chose not to stratify the population by race or breast cancer stage, which would account for these confounders.

In contrast to previous studies, we did not limit our analysis to allow only one diagnosis per subject[6,26], since many of the ADEs related to trastuzumab may manifest after a longer follow-up, preceded by other acute events. Therefore a further limitation is the lack of independence between different ADE counts. Unfortunately, the current data mining methods cannot readily adapt to include the dependency structure of the data.

Finally, data mining methods are used as a first step in drug safety surveillance in order to identify excess risk. Because these methods evaluate at the same time a large number of potential outcomes, it is not possible to adjust for all possible confounders. Instead, the next step is by clinical review to determine which signals are indicated by confounding and which would warrant further review and rigorous epidemiologic study.

In conclusion, administrative claims data can enhance current drug safety surveillance, but further research is needed to determine the optimal methods and signaling criteria. The choice of approach and method depends largely on the purpose of each study; the GPS method is less likely to report false positives and may be used to detect more common ADEs. A combination of two data mining algorithms, the TBSS and PRR may be more appropriate when the interest is to also detect signals of less common ADEs. Also, for initial exploratory purposes the

visit level approach may provide superior performance, while the subject level approach can serve a more confirmatory role to already suspected ADEs.

References

1. Lazarou J, Pomeranz BH, Coney PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *J Am Med Assoc*. 1998;279(15):1200-1205.

2. Classen DC, Pestonik SL, Scott Evans R, Lloyd JF, Burke JP. Adverse Drug Events in Hospitalized Patients. Excess Length of Stay, Extra Costs and Attributable Mortality. *JAMA*. 1997;277:301-306. doi:10.1097/00006254-199705000-00016

3. World Health Organization. *The Importance of Pharmacovigilance - Safety Monitoring of Medicinal Products*.; 2002. doi:10.1002/0470853093

4. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clin Pharmacol Ther*. 2012;91(6):1010-1021. doi:10.1038/clpt.2012.50.Novel

5. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol Drug Saf*. 2007;16(4):359-365. doi:10.1002/pds.1323

6. Brown J, Petronis K, Bate A, et al. Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic. *Pharmaceutics*. 2013;5(1):179-200. doi:10.3390/pharmaceutics5010179

7. Curtis JR, Cheng H, Delzell E, et al. Adaptation of Bayesian Data Mining Algorithms to Longitudinal Claims Data : Coxib Safety as an Example. *Med Care*. 2008;46(9):969-975. doi:10.1097/MLR.0b013e318179253b.Adaptation

8. Schachterle SE, Hurley S, Liu Q, Petronis KR, Bate A. An Implementation and Visualization of the Tree-Based Scan Statistic for Safety Event Monitoring in Longitudinal Electronic Health Data. *Drug Saf*. 2019;(0123456789). doi:10.1007/s40264-018-00784-0

9. Data Mining at FDA -- White Paper. US Food and Drug Administration. https://www.fda.gov/science-research/data-mining-fda/data-mining-fda-white-paper#changepoint. Published 2018. Accessed May 5, 2019.

10. Guideline on the Use of Statistical Signal Detection Methods in the Eudravigilance Data Analysis System. European Medicines Agency. https://www.ema.europa.eu. Published 2006. Accessed May 5, 2019.

11. Cameron D, Piccart-Gebhart MJ, Gelber RD, et al. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet (London, England)*. 2017;389(10075):1195-1205. doi:10.1016/S0140-6736(16)32616-2

12. Goldhirsch A, Gelber RD, Piccart-Gebhart MJ, et al. 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): An open-label, randomised controlled trial. *Lancet*. 2013;382(9897):1021-

1028. doi:10.1016/S0140-6736(13)61094-6

13. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Stat Methods Med Res*. 2013;22(1):39-56. doi:10.1177/0962280211403602

14. Overview of the SEER Program. National Cancer Institute. https://seer.cancer.gov/about/overview.html. Published 2019. Accessed June 14, 2019.

15. 2018 Texas Cancer Registry Annual Report. Texas Department of State Health Services. https://www.dshs.texas.gov/tcr/. Published 2018. Accessed June 8, 2019.

16. Davis K., Schoen C. BF. *Medicare: 50 Years of Ensuring Coverage and Care*.; 2015.

17. Huo J, Giordano SH, Smith BD, Shaitelman SF, Smith GL. Contemporary toxicity profile of breast brachytherapy versus external beam radiation after lumpectomy for breast cancer. *Int J Radiat Oncol Biol Phys*. 2016;94(4):709-718. doi:10.1016/j.ijrobp.2015.12.013

18. SEER-Medicare: Medicare Claims Files. National Cancer Institute. https://healthcaredelivery.cancer.gov/seermedicare/medicare/claims.html. Published 2017. Accessed May 4, 2018.

19. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-383.

20. Clinical Classifications Software (CCS) for ICD-9-CM. Agency for Healthcare Research and Quality. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed January 6, 2018.

21. Candore G, Juhlin K, Manlik K, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Saf*. 2015;38(6):577-587. doi:10.1007/s40264-015-0289-5

22. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf*. 2001;10(6):483-486. doi:10.1002/pds.677

23. Dumouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *Am Stat*. 1999;53(3):177-190.

24. Szarfman A, Machado SG, O'Neill RT. Use of Screening Algorithms and Computer Systems to Efficiently Signal Higher-Than-Expected Combinations of Drugs and Events in the US FDA's Spontaneous Reports Database. *Drug Saf*. 2002;25(6):381-392. doi:10.2165/00002018-200225060-00001

25. Madigan D., Ryan P., Simpson S. ZI. Bayesian Methods in Pharmacovigilance. In: *Oxford University Press*. ; 2010.

26. Kulldorff M, Fang Z, Walsh SJ. A Tree-Based Scan Statistic for Database Disease Surveillance. *Biometrics*. 2003;59(2):323-331. doi:10.1111/1541-

0420.00039

27. Wittayanukorn S, Qian J, Johnson BS, Hansen RA. Cardiotoxicity in targeted therapy for breast cancer: A study of the FDA adverse event reporting system (FAERS). *J Oncol Pharm Pract*. 2017;23(2):93-102. doi:10.1177/1078155215621150

28. Chavez-MacGregor M, Zhang N, Buchholz TA, et al. Trastuzumab-related cardiotoxicity among older patients with breast cancer. *J Clin Oncol*. 2013;31(33):4222-4228. doi:10.1200/JCO.2013.48.7884

29. Cabiddu M, Guarneri P, Petrelli F, Barni S, Lonati V. The Risk for Anemia with Targeted Therapies for Solid Tumors. *Oncologist*. 2012;17(5):715-724. doi:10.1634/theoncologist.2012-0024

30. Li Y, Ryan PB, Wei Y, Friedman C. A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions. *Drug Saf*. 2015;38(10):895-908. doi:10.1007/s40264-015-0314-8

31. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med*. 2001;344(11):783-792. doi:10.1056/NEJM200103153441101

32. Yusuf SW, Sami S, Daher IN. Radiation-Induced Heart Disease: A Clinical Update. *Cardiol Res Pract*. 2011;2011(April 2007):1-9. doi:10.4061/2011/317659

33. Al-Dasooqi N, Bowen JM, Gibson RJ, Sullivan T, Lees J, Keefe DM. Trastuzumab induces gastrointestinal side effects in HER2-overexpressing breast cancer patients. *Invest New Drugs*. 2009;27(2):173-178. doi:10.1007/s10637-008-9152-1

Appendix A

Diagnoses and procedures that were excluded from analysis and associated codes.

Conditions / Procedures	ICD – 9 Codes	CPT codes and NDC codes
BC surgery	'854' '852'	'19180' '19182' '19200' '19220' '19240' '19110' '19120' '19125' '19126' '19160' '19162' '19301' – '19307'
BC chemotherapy	'V5811' 'V662' 'V672' 'V5812'	'00004110020' '00004110051' '00004110116' '00004110150' '00004110175' '00015050241' '00015050301' '00015050401' '00015050541' '00015054841' '00015347630' '00026848858' '00054412925' '00054413025' '00069055038' '00069098038' '00075800120' '00075800180' '00078040105' '00078040134' '00078040215' '00078043815' '00085124402' '00085124801' '00085125901' '00085125902' '00172375377' '00173075200' '00378326694' '00703315401' '10019095501' '10139006202' '50242005121' '50242006001' '50242006301' '50242013468' '55390023110' '55390023701' '55390023801' '58406064007' '61703034209' '61703034222' '61703034250' '62856060210' '50242014501' '50242005656' '50242013460' '50242013468'

Conditions / Procedures	ICD – 9 Codes	CPT codes and NDC codes
BC radiation	'V580' 'V661' 'V671' 'V581' '922' '9220' '9221' '9222' '9223' '9224' '9225' '9226' '9227' '9229' '923' '9230' '9231' '9232' '9233' '9234' '9235' '9236' '9237' '9238' '9239' '924' '9241'	'19296' '19297' '19298' '20555' '31643' '32553' '41019' '43241' '49411' '55875' '55876' '55920' '57155' '57156' '58346' '61770' '61793' '76000' '76001' '76370' '76872' '76873' ''76950' '76965' '77002' '77012' '77014' '77021' '77261' '77262' '77263' '77280' '77285' '77290' '77295' '77299' '77300' '77301' '77305' '77310' '77315' '77321' '77326' '77327' '77328' '77331' '77332' '77333' '77334' '77336' '77338' '77370' '77371' '77372' '77373' '77399' '77401' '77402' -'77427' '77431' '77432' '77435' '77469' '77470' '77499' '77520' '77522' '77523' '77525' '77600' '77605' '77610' '77615' '77620' '77750' '77761' '77762' '77763' '77776' '77777' '77778' '77781' - '77790' '77799' '79900' '77761' – '77799' '70010' - '79999' '0073T' '0182T' '0190T' '0197T' 'A4650' 'A9527' 'C1715' 'C1716' 'C1717' 'C1718' 'C1719' 'C1728' 'C1879' 'C2616' 'C2634' -'C2699' 'C9714' 'C9715' 'C9725' 'C9726' 'C9728' 'G0173' 'G0174' 'G0251' 'G0339' 'G0340' 'Q3001' 'S8030'

Conditions that were excluded from analysis and associated CCS codes

Conditions / Procedures	CCS codes
Injuries	16.1, 16.3 – 16.9
Diseases originating in the perinatal period	11, 15
Congenital diseases	14
Other codes relating to admission/ aftercare/ immunizations	17.2, 1.5

Appendix B

MLCCS	Diagnosis	PRR 95% LCI	PRR 95% UCI	EBGM 95% LCI	EBGM 95% UCI	Tree-Scan p-value
1	Infectious and parasitic diseases					
1.1.2.3	E. Coli septicemia	1.18	7.93			
7	Diseases of the circulatory system					0.003
7.2	Diseases of the heart					<0.001
7.2.1	Heart valve disorders	2.07	3.01	1.85	1.89	<0.001
7.2.1.2	Nonrheumatic mitral valve disorders	2.15	3.74	1.91	1.93	<0.001
7.2.1.3	Nonrheumatic aortic valve disorders	1.42	3.08			
7.2.1.4	Other heart valve disorders	1.97	4.25	1.91	1.93	0.003
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	1.81	3.70	1.84	1.89	0.005
7.2.2.1	Cardiomyopathy	1.90	4.35	1.90	1.93	0.012
7.2.7	Other and ill-defined heart disease	1.86	4.60			0.033
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	1.10	6.97			
9.4.2.2	Duodenal ulcer	1.43	10.90			
9.5.3.8	Incisional hernia without obstruction/gangrene	1.25	466.70			
12	Diseases of the skin and subcutaneous tissue					
12.1.1.1	Cellulitis and abscess of fingers and toes	1.16	3.70			
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	1.10	6.97			
13.6.2	Other acquired deformities	1.10	6.97			
16	Injury and Poisoning					
16.11	Poisoning	1.63	3.70			
16.11.2	Poisoning by other medications and drugs	1.60	3.60			

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.

Confidence Intervals for each method for the subject level analysis at 3 months

MLCCS	Diagnosis	PRR 95% LCI	PRR 95% UCI	EBGM 95% LCI	EBGM 95% UCI	Tree-Scan p-value
2	Neoplasms					
2.6.1	Cancer of uterus	1.00	20.01			
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders					0.036
4	Diseases of the blood and blood-forming organs					
4.4	Other hematologic conditions	1.06	5.97			
7	Diseases of the circulatory system					<0.001
7.2	Diseases of the heart					<0.001
7.2.1	Heart valve disorders	2.84	3.75	2.11	2.12	<0.001
7.2.1.1	Chronic rheumatic disease of the heart valves	2.02	4.40	2.10	2.27	0.004
7.2.1.2	Nonrheumatic mitral valve disorders	3.36	5.10	2.13	2.29	<0.001
7.2.1.3	Nonrheumatic aortic valve disorders	1.85	3.31	2.09	2.25	<0.001
7.2.1.4	Other heart valve disorders	2.59	4.68	2.11	2.27	<0.001
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	2.42	4.21	2.11	2.12	<0.001
7.2.2.1	Cardiomyopathy	2.77	5.21	2.11	2.28	<0.001
7.2.7	Other and ill-defined heart disease	2.45	4.62	2.11	2.12	<0.001
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	1.34	8.45			
12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	1.02	4.63			
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	1.23	7.43			
16	Injury and Poisoning					0.021
16.10.2.2	Respiratory complications	1.00	20.01			
16.11	Poisoning	1.74	3.51			
16.11.2	Poisoning by other medications and drugs	1.70	3.49	1.04	2.11	0.024

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.

Confidence Intervals for each method for the subject level analysis at 6 months

MLCCS	Diagnosis	PRR 95% LCI	PRR 95% UCI	EBGM 95% LCI	EBGM 95% UCI	Tree-Scan p-value
2	Neoplasms					
2.6.1	Cancer of uterus	1.54	12.74			
2.11.1	Cancer of head and neck	1.12	5.81			
4	Diseases of the blood and blood-forming organs					
4.4	Other hematologic conditions	1.01	4.58			
7	Diseases of the circulatory system					<0.001
7.2	Diseases of the heart					<0.001
7.2.1	Heart valve disorders	2.99	3.79	2.12	2.13	<0.001
7.2.1.1	Chronic rheumatic disease of the heart valves	2.27	4.34	1.96	2.39	<0.001
7.2.1.2	Nonrheumatic mitral valve disorders	3.91	5.61	2.19	2.55	<0.001
7.2.1.3	Nonrheumatic aortic valve disorders	2.00	3.30	1.89	2.28	0.002
7.2.1.4	Other heart valve disorders	2.90	4.67	2.03	2.43	<0.001
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	2.56	4.11	2.12	2.13	<0.001
7.2.2.1	Cardiomyopathy	2.55	4.32	1.98	2.39	<0.001
7.2.2.2	Other peri-; endo-; and myocarditis	1.59	4.52			
7.2.7	Other and ill-defined heart disease	2.30	3.81	2.12	2.13	<0.001
7.2.11.2	Heart failure	1.37	4.82			
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	1.05	6.72			
9.10.3	Gastroesophageal laceration syndrome	1.20	449.51			
12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	1.52	5.33			
16	Injury and Poisoning					
16.2.4.4	Unclassified fracture of lower limb	1.22	36.22			
16.11	Poisoning					
16.11.2	Poisoning by other medications and drugs	1.49	2.83			

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.

Confidence Intervals for each method for the subject level analysis at 12 months

MLCCS	Diagnosis	PRR 95% LCI	PRR 95%UCI	EBGM 95% LCI	EBGM 95% UCI	Tree-Scan p-value
1	Infectious and parasitic diseases					
1.1.2.3	E. Coli septicemia	1.16	7.80			
2	Neoplasms					
2.11.4	Cancer of brain and nervous system	2.10	31.46			
2.11.5	Cancer of thyroid	1.09	100.56			
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders					0.045
4	Diseases of the blood and blood-forming organs					
4.1.3.2	Other deficiency anemia	1.45	3.23			
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	1.40	4.26			
6.8.1.2	Other otitis media and related conditions	1.15	4.69			
6.9.2	Other central nervous system disorders	1.36	3.26			
7	Diseases of the circulatory system					<0.001
7.2	Diseases of the heart					0.024
7.2.1	Heart valve disorders	2.04	2.92	1.61	2.97	<0.001
7.2.1.2	Nonrheumatic mitral valve disorders	2.15	3.64	1.87	1.88	<0.001
7.2.1.4	Other heart valve disorders	1.77	3.65	1.87	1.88	0.003
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	1.83	3.46	1.47	2.01	<0.001
7.2.2.1	Cardiomyopathy	1.93	4.03	1.87	1.88	0.002
7.2.7	Other and ill-defined heart disease	1.56	3.68			
7.5.4	Other diseases of veins and lymphatics					0.017
8	Diseases of the respiratory system					
8.3.2.3	Other asthma with acute exacerbation	1.29	6.40			
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	1.30	7.85			
9.4.2.2	Duodenal ulcer	1.62	11.70			
9.5.3.8	Incisional hernia without obstruction/gangrene	1.64	566.50			
9.12	Other gastrointestinal disorders					0.022
9.12.3	Other and unspecified gastrointestinal disorders					0.042

12	Diseases of the skin and subcutaneous tissue					
12.3.3	Other chronic skin ulcer	1.80	7.74			
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	1.35	4.91			
13.6.2	Other acquired deformities	1.24	6.03			
16	Injury and Poisoning					
16.2.3.2	Fracture of radius and ulna	1.15	4.23			
16.2.3.3	Other fracture of upper limb	1.44	18.01			
16.11	Poisoning	1.64	3.56			0.044
16.11.2	Poisoning by other medications and drugs	1.58	3.48	1.01	1.94	
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting					<0.001
MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.						

Confidence Intervals for each method for the visit level analysis at 3 months

MLCCS	**Diagnosis**	**PRR 95% LCI**	**PRR 95% UCI**	**EBGM 95% LCI**	**EBGM 95% UCI**	**Tree-Scan p-value**
2	Neoplasms					
2.11.4	Cancer of brain and nervous system	1.10	7.40			
2.11.5	Cancer of thyroid	1.21	8.55			
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders					0.013
3.8.2	Hypovolemia					0.009
4	Diseases of the blood and blood-forming organs					
4.1.3.2	Other deficiency anemia	1.49	2.74	1.08	1.97	0.009
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	1.63	3.35	1.00	2.09	0.031
6.8.1.2	Other otitis media and related conditions	1.52	3.78			

7	Diseases of the circulatory system					<0.001
7.2	Diseases of the heart					<0.001
7.2.1	Heart valve disorders	2.78	3.58			<0.001
7.2.1.1	Chronic rheumatic disease of the heart valves	1.89	3.97	1.59	2.17	0.002
7.2.1.2	Nonrheumatic mitral valve disorders	3.31	4.90	2.00	2.37	<0.001
7.2.1.3	Nonrheumatic aortic valve disorders			1.12	1.89	0.002
7.2.1.4	Other heart valve disorders	2.05	3.50	1.68	2.12	0.001
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	2.14	3.42			<0.001
7.2.2.1	Cardiomyopathy	2.32	3.94	1.74	2.19	<0.001
7.2.7	Other and ill-defined heart disease	2.06	3.68			<0.001
8	Diseases of the respiratory system					
8.5.3	Empyema and pneumothorax	1.68	15.70			
9	Diseases of the digestive system					
9.4.2.2	Duodenal ulcer	1.88	10.30			
9.5.3.8	Incisional hernia without obstruction/gangrene	1.85	45.34			
9.10.1	Hemorrhage from gastrointestinal ulcer	1.03	4.44			
9.10.3	Gastroesophageal laceration syndrome	1.16	436.00			
12	Diseases of the skin and subcutaneous tissue					
12.3	Chronic ulcer of skin					0.004
12.3.3	Other chronic skin ulcer	1.78	5.34			
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	1.86	6.34			
13.6.2	Other acquired deformities	1.22	3.77			
16	Injury and Poisoning					
16.2.3.2	Fracture of radius and ulna	1.48	3.33			
16.2.3.3	Other fracture of upper limb	1.03	4.87			
16.2.5.4	Other and unspecified fracture	1.03	4.22			
16.11	Poisoning	1.61	3.03			0.016
16.11.2	Poisoning by other medications and drugs	1.53	2.93			0.032
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting					0.005

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.

Confidence Intervals for each method for the visit level analysis at 6 months

MLCCS	Diagnosis	PRR 95% LCI	PRR 95% UCI	EBGM 95% LCI	EBGM 95% UCI	Tree-Scan p-value
2	Neoplasms					
2.6.1	Cancer of uterus	1.05	5.32			
2.11.1	Cancer of head and neck	1.12	4.89			
2.11.5	Cancer of thyroid	1.44	6.54			
3	Endocrine; nutritional; metabolic diseases; immunity disorders					
3.8	Fluid and electrolyte disorders					<0.001
3.8.2	Hypovolemia					<0.001
4	Diseases of the blood and blood-forming organs					
4.1	Anemia					<0.001
4.1.3	Deficiency and other anemia					0.003
6	Diseases of the nervous system and sense organs					
6.8.1	Otitis media and related conditions	1.73	2.97	1.25	2.18	<0.001
6.8.1.2	Other otitis media and related conditions	1.77	3.38	1.26	2.22	<0.001
7	Diseases of the circulatory system					<0.001
7.2	Diseases of the heart					<0.001
7.2.1	Heart valve disorders	3.52	4.31	2.17	2.40	<0.001
7.2.1.1	Chronic rheumatic disease of the heart valves	2.51	4.54	1.85	2.42	<0.001
7.2.1.2	Nonrheumatic mitral valve disorders	4.53	6.24	2.33	2.71	<0.001
7.2.1.3	Nonrheumatic aortic valve disorders	1.76	2.61	1.51	1.97	<0.001
7.2.1.4	Other heart valve disorders	2.55	3.91	1.86	2.31	<0.001
7.2.2	Peri-; endo-; and myocarditis; cardiomyopathy	2.89	4.07	2	2.33	<0.001
7.2.2.1	Cardiomyopathy	2.89	4.20	1.95	2.37	<0.001
7.2.2.2	Other peri-; endo-; and myocarditis	1.32	3.14			
7.2.7	Other and ill-defined heart disease	2.49	3.87	1.93	2.29	<0.001
7.2.11.2	Heart failure	1.70	4.94	1.00	2.38	
9	Diseases of the digestive system					
9.2	Disorders of teeth and jaw	1.32	4.25			
9.4.2.2	Duodenal ulcer	2.92	13.33	1.20	2.64	
9.6.5	Anal and rectal conditions	1.65	3.83	1.04	2.27	

9.10.1	Hemorrhage from gastrointestinal ulcer	1.01	4.30			
9.10.3	Gastroesophageal laceration syndrome	1.60	554.00			
12	Diseases of the skin and subcutaneous tissue					
12.3	Chronic ulcer of skin					<0.001
12.3.3	Other chronic skin ulcer	3.14	6.60	1.99	2.43	<0.001
13	Diseases of musculoskeletal system and connective tissue					
13.1	Infective arthritis/osteomyelitis (expect caused by TB or STD)	2.10	5.14			
16	Injury and Poisoning					<0.001
16.2.4.4	Unclassified fracture of lower limb	1.51	40.20			
16.11	Poisoning	1.90	3.18			<0.001
16.11.2	Poisoning by other medications and drugs	1.94	3.33	1.83	2.23	<0.001
17	Symptoms; signs; ill-defined conditions /factors influencing health status					
17.1.6	Nausea and vomiting					0.003

MLCCS: Multi-level Clinical Classification System; PRR: Proportional Risk Ratio, EBGM: Empirical Bayesian Geometric Mean, LCI: Lower confidence interval, UCI: Upper confidence interval. Confidence intervals for PRR are two-sided, for GPS one-sided.

Confidence Intervals for each method for the visit level analysis at 12 months

Vita

Efstathia Polychronopoulou was born to Ioannis and Anthi Polychronopoulou on 10 March 1984 in Athens, Greece. She graduated from the National Technical University of Athens with a Bachelor of Science in Applied Mathematics and Physical Sciences in 2007 and a Master's of Science in Special Education in 2011 from the University of Athens. In August 2016, she began her Master in Public Health with a concentration in Biostatistics at the Department of Preventive Medicine and Community Health at the University of Texas Medical Branch in Galveston. She currently resides in Galveston, Texas.

This capstone was typed by Efstathia Polychronopoulou.