Application of Probabilistic Linkage: Compare Health Care Costs among Menopausal Women with Different Symptoms by Linking Women's Registry & Claims Data

O. Baser[1,2,*], L. Xie[2,#] and J. Du[2,#]

[1]The University of Michigan, Department of Internal Medicine, Ann Arbor, MI, USA

[2]STAT in MED Research, Ann Arbor, MI, 48104, USA

Abstract: *Objectives*: Menopause symptoms are a good disease severity proxy for menopausal women, but are not available in claims database. We applied probabilistic linkage to add symptoms recorded in a registry database to claims data, and compare the healthcare costs among women with various symptoms.

Methods: Women age 45 or older who used estrogen only hormone therapy (HT) were selected from a large U.S. claims database (04/01/2005-09/30/2008). Another group who used estrogen only HT with a menopause diagnosis was selected from the University of Michigan Women's Registry Database. Logistic regression was used to calculate the propensity score for each patient controlling for osteoporosis, gynecological disorders/procedures, genital infection, gynecology system cancer, breast condition, gut condition, hormone disorder, nerve problem, and other individual comorbidities such as rheumatoid disease, depression, and blood clotting. Patients with the closest propensity score from each group were matched, and menopause symptoms for registry patients were added to the claims database records. After repeating probabilistic linkage 250 times, the mean and 95% confidence interval (CI) of healthcare costs during the follow-up period were calculated.

Results: 80 patients from each population were matched after probabilistically linking 20,020 claims database patients with 83 registry database patients. The average cost of patients with at least one symptom was much higher than for patients without symptoms ($13,570 [95% CI: $13,459-$13,680] vs. $3,391 [95%CI: $3,345-$3,436], p-value<0.001). (1 US Dollar= 0.75 Euro) Cost differences were mainly from inpatient, physician visit, and pharmacy costs. Among patients with menopause symptoms, those with hot flashes had the highest costs ($10,127), followed by memory loss ($1,653), vaginal dryness ($864), reduced libido ($568), and mood swings ($358).

Conclusions: Women with menopause symptoms incur higher healthcare costs than those without This study suggests symptoms are important determinants of healthcare expenses and their impact can be assessed by linking registry and claims databases.

Keywords: Probabilistic linkage, women's health, propensity score matching.

INTRODUCTION

Randomized clinical trials (RCTs), although recognized as the "gold standard" to provide solid evidence on causal inference, are carried out using selected populations under idealized, controlled conditions. In the real world, physician practice/prescription patterns and patient preference are influenced by non-clinical factors. For example, practice patterns might be based on physician rules of thumb, experiences and interaction with patients and colleagues. Health plans provide different financial and non-financial incentives to doctors or patients to undergo aggressive treatment as well [1]. Therefore, decision-makers are increasingly seeking information on "real-world" outcomes on which to base their decisions.

Real world data can be in the form of (a) supplements to traditional RCTs; (b) large sample trials; (c) registries; (d) administrative data; (e) health surveys; and (f) electronic health records and medical chart reviews [2]. Each of these data sources collects different variables, and some research questions might require linkage. For example, in order to estimate the health care cost of different treatments, researchers need information on the costs and several severity variables for risk adjustment. Costs can be found in administrative data, but severity measures might be located in electronic health records.

Record linkage involves searching files for records that belong to the same individual. For example, linking the electronic health records and administrative data allow researcher to find relationship between severity and health care costs. Ideal way to link the files is to use deterministic record linkage where we look for exact (dis)agreement on one or more matching variables between files. For example, we might simply use beneficiary id number common to two files. However, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) rules remove the directly identifiable information from the dataset, the

*Address correspondence to this author at the Department of Internal Medicine, Medical School, The University of Michigan, Ann Arbor, MI, 48104, USA; Tel: +1 734-222-5426; Fax: +1 734-864-0359;
E-mail: onur@med.umich.edu

#Co-Authors E-mails: lxie@statinmed.com, jdu@statinmed.com

E-ISSN: 1929-6029/13

only way to link these datasets is to apply methodology that matches two files of individual data under conditions of uncertainty.

Probabilistic linkage makes it feasible and efficient to link different databases in a statistically justifiable manner. Fellegi and Stunner pioneered record linkage theory [3]. Subsequently, the theory was applied to several datasets [4, 5].

In this paper, we will first describe the linkage algorithm using the propensity score matching technique. The linkage will be used to link registry data and administrative data to analyze the effect of symptoms on the health care cost of post-menopausal women. Symptoms from the University of Michigan's Women's Registry database will be linked to commercial claims files where cost information is available.

METHOD

Individual record linkage usually involves two patient-level files: File$_0$ and File$_1$. Let's assume N_0 is the number of patients in File$_0$ and N_1 is the number of patients in File$_1$. Each consists of fields that contain information to be matched. Some of the fixed fields in both files should be equivalent. For example, each file should contain age, gender, comorbid conditions, etc., in order to incorporate the linkage algorithm. Let X be the set of these variables that are available in both data sets. Let $X \in P$, where *P* represents the population. The objective of the linkage process is to classify each pair as belonging to the set of matched record pairs or the set of unmatched record pairs. The matching algorithm will set $w_{N_o,N_1}(i) = \frac{1}{N_1}$, and for each individual *i* in the File$_1$**,** the match in the other file will be picked with $C(X_i) = \min_j \left\|X_i - X_j\right\|, j \in P_0$, where $\|.\|$ is a norm. $C(X_i)$ is the neighborhood for individual *i,* and persons matched to individual *i* are those people in the set A_i, where $A_i = \{j \in P_0 \mid X_j \in C(X_i)\}$. The matching algorithm will take each individual in the first file and search for the individual in the second file with the closest propensity score. Researchers can relate the match weight for a pair of records to the probability that these records are correctly matched. The process needs to be repeated at least 250 times to estimate the standard errors and confidence intervals [6].

A CASE STUDY

Nearly 50 million women each year are projected to reach menopause by 2030 [7]. Menopausal symptoms include vasomotor symptoms, such as night sweats and hot flashes, as well as urogenital atrophy. The prevalence of these vasomotor symptoms, ranging from very mild to severe, is 79% in peri-menopausal and 65% in post-menopausal women aged 40-65 years in the United States [8]. Due to the lack of datasets, there is little information about the association between menopausal symptoms and healthcare costs.

We applied the linkage method to a large U.S. claims database from April 01, 2002 to September 30, 2010 and the University of Michigan Women's Registry database.

To be eligible for the study, subjects from the claims database were required to be female, 45 years of age or older, and have used only HT estrogen from April 01, 2005 to September 30, 2008. The first estrogen only prescription date was designated as the index date. Furthermore, patients were required to have continuous enrollment in a health plan with medical and pharmacy benefits during the 3 years prior to the index date (pre-index period) and during the 2 years after the index date (post-index period). Patients were excluded from the study if they had evidence of any hormone therapy pharmacy claims during the pre-index period or evidence of pregnancy during the post-index period. Menopausal women treated with the estrogen only treatment were selected from the Women's Registry database.

Both cost and utilization variables were available in the claims data. Post-index health care costs were categorized as total, inpatient, outpatient physician, outpatient emergency room (ER), other outpatient, and outpatient pharmacy costs. In terms of health care utilization, the percentage of patients with at least one inpatient stay, outpatient ER or outpatient physician visit was calculated.

Baseline comorbidities, such as osteoporosis, genital infection, bladder/pelvic floor support problems, gynecologic cancer, breast conditions, gut conditions, hormone disorders, nerve problems, rheumatoid disease, depression, and blood clotting, as well as baseline procedures such as gynecological disorders or procedures including hysteroscopy, dilation and curettage, surgical removal of fibroids, uterine artery embolization, endometrial ablation, laparoscopy, surgical treatment for pelvic scar tissue, surgical treatment for endometriosis, along with age. Since estrogen treatment was used in the inclusion criteria, the subjects were perfectly matched by treatment.

Logistic regression was used to calculate the propensity scores for each patient in two cohorts, controlling for baseline confounders. Patients with the closest propensity score from each cohort were matched, and the menopause symptoms for patients from the Women's Registry database were added to the patients' records in the claim database. Healthcare costs and utilizations were calculated for the linked sample. After randomly sorting the sample from the claims database, the same method was applied to link the two cohorts as described above. Healthcare costs and utilizations were calculated for each new linked sample. The probabilistic linkage process was repeated 250 times. The mean and 95% confidence interval (CI) of healthcare cost and utilizations during the follow-up period were calculated. Healthcare costs and utilizations were compared between patients with at least one symptom versus those without any symptoms.

After applying inclusion and exclusion criteria, 20,020 patients were selected from the claims database, and 83 patients were selected from the Women's Registry. As shown in Figure **1**, out of 83 patients from the Women's Registry, 41 had hot flashes, 13 had memory problems, 2 had mood swings, 5 had a reduced libido, 6 had vaginal dryness, and 16 had no symptoms at all.

After probabilistic linkage, 80 patients with symptom variables as well as healthcare costs and utilizations were linked from the two databases. Figure **2** shows the healthcare cost comparison between patients with at least one symptom versus those without symptoms. The average total cost of patients with at least one symptom was much higher than for patients without symptoms ($13,570 [95% CI: $13,459-$13,680] vs. $3,391 [95%CI: $3,345-$3,436], p-value<0.001). Significantly higher costs were also found for patients with at least one symptom in inpatient costs ($1,997 [95% CI: $1,925-$2,070] vs. $247 [95% CI: $239-$254], p-value<0.001), outpatient ER costs ($167 [95% CI: $160-$174] vs. $118 [95% CI: $116-$120], p-value<0.001), outpatient physician costs ($967 [95% CI: $961-$974] vs. $248 [95% CI: $246-$251], p-value<0.001), pharmacy costs ($3,676 [95% CI: $3,648-$3,704] vs. $903 [95% CI: $890-$916], p-value<0.001), and other outpatient costs ($6,762 [95% CI: $6,696-$6,827] vs. $1,874 [95% CI: $1,840-$1,908], p-value<0.001), compared to patients without symptoms.

Figure **3** shows the healthcare utilization comparison between patients with at least one symptom versus those without. Patients with at least one symptom had a significantly higher percentage of inpatient stays (12.78% [95% CI: 12.57%-12.99%] vs. 4.05% [95% CI: 3.98% vs. 4.12%], p-value<0.001), outpatient ER visits (20.60% [95% CI: 20.29%-20.91%] vs. 6.31%[95% CI: 6.18%-6.43%], p-value<0.001), and outpatient physician visits (79.86% [95% CI: 79.80%-79.91%] vs. 18.71% [95% CI: 18.68%-18.73%], p-value<0.001), compared to patients without symptoms.

Among patients with menopause symptoms, those with hot flashes had the highest total cost ($10,127),

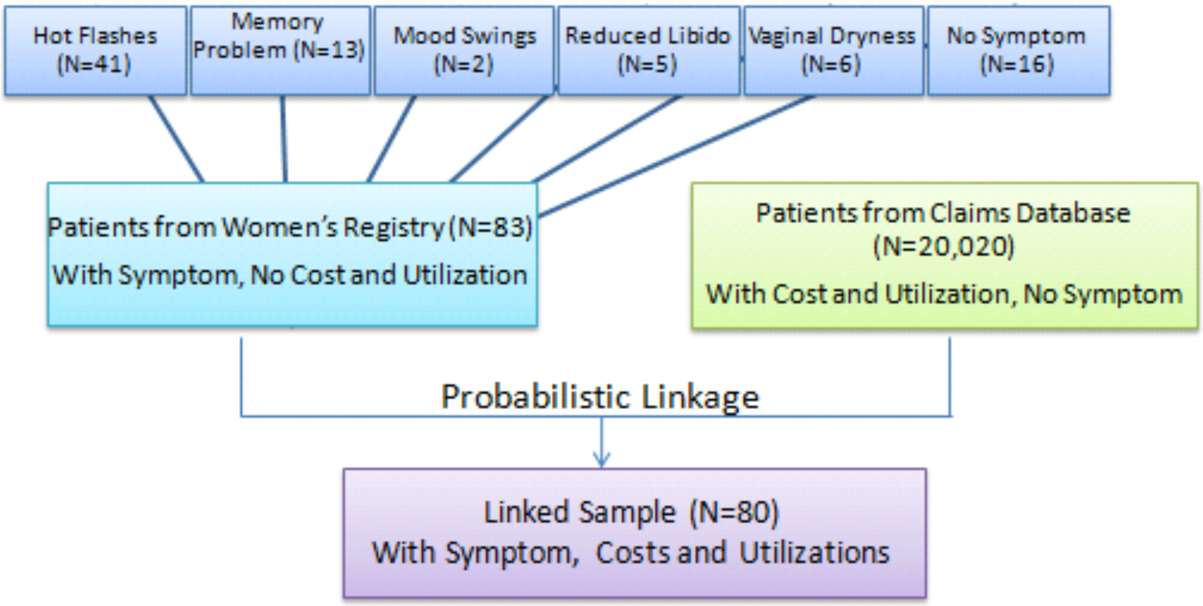
Hot Flashes (N=41)
Memory Problem (N=13)
Mood Swings (N=2)
Reduced Libido (N=5)
Vaginal Dryness (N=6)
No Symptom (N=16)
Patients from Women's Registry (N=83)
With Symptom, No Cost and Utilization
Patients from Claims Database (N=20,020)
With Cost and Utilization, No Symptom
Probabilistic Linkage
Linked Sample (N=80)
With Symptom, Costs and Utilizations

Figure 1: Symptom Distribution of Patients from the Women's Registry.

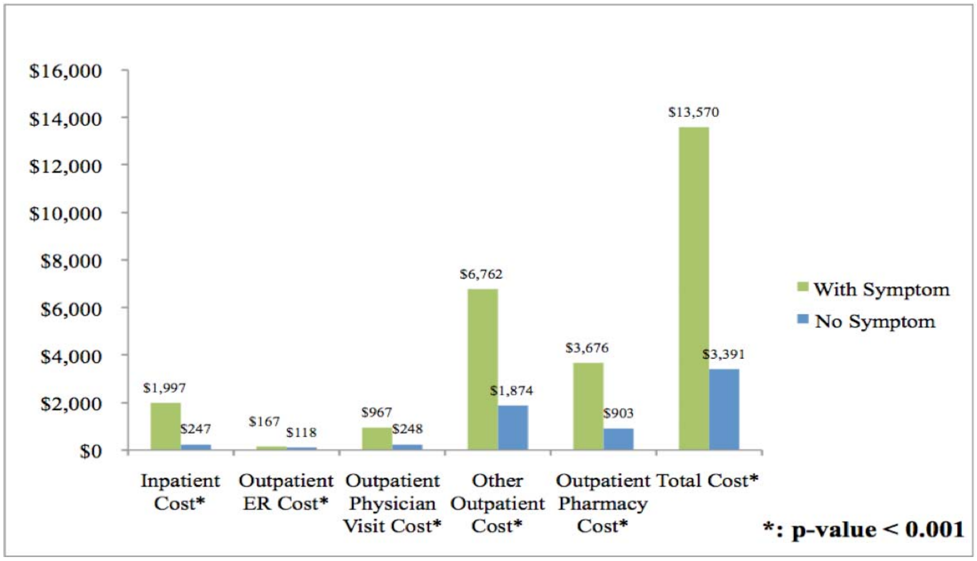

Figure 2: Follow-Up Healthcare Costs Between Patients With At Least One Symptom Versus Those Without.

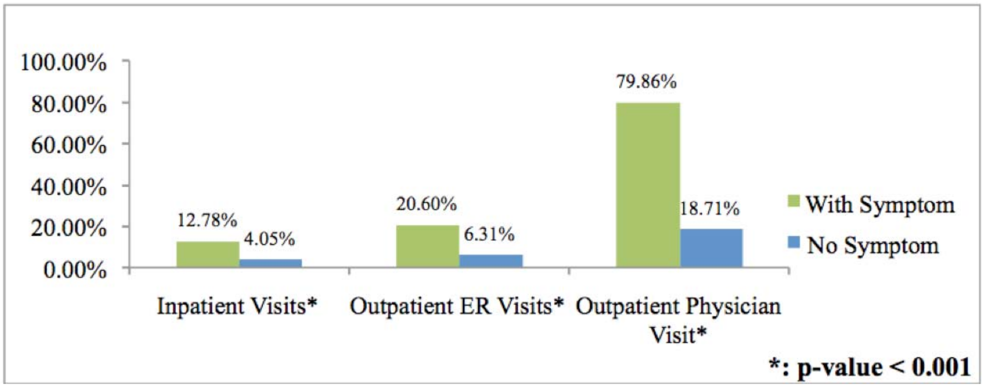

Figure 3: Follow-Up Healthcare Utilization between Patients With At Least One Symptom Versus Those Without any Symptoms.

followed by patients with memory loss ($1,653), vaginal dryness ($864), reduced libido ($568), and mood swings ($358). The same trend was also found in the healthcare utilization of patients with symptoms.

DISCUSSION

Investigators often have to rely on observational data to estimate the effect of treatments on outcomes, such as healthcare costs, which are difficult or impossible to estimate in the artificial setting of a clinical trial. Therefore, there is a broad surge of interest in the fundamental approaches of drawing causal inference from observational data which has emerged during the last decade.

Administrative data (typically retrospective) are frequently used in outcomes research studies when measuring resource use and costs in the real world setting. However, this data is collected primarily for reimbursement purposes and is missing clinical measures, such as symptoms and disease severity scores. However, clinical measures can be available in electronic medical records or registry databases.

The HIPAA Privacy and Security Rules protect the privacy of individually-identifiable health information and prevent linking these datasets in deterministic matching. Since directly identifiable information such as names, social security numbers, and addresses are not available in these datasets, the only way to link these datasets under the HIPAA rules is through probabilistic matching. Enhanced variable lists from this linkage allow investigators to answer the research questions that would otherwise not be possible.

In this paper, we linked the University of Michigan Women's health registry database with a US claims database to analyze the effect of symptoms on health care costs. Many equivalent fields in both datasets, such as age, comorbid conditions, and medical procedures made the probabilistic matching feasible. The comparison of numerous data fields lead to a judgment of whether two records refers to same patient. This judgment was based on the cumulative weight of agreement and disagreement among field values. For example, agreement on age field alone would not determine that two records to refer to same patients but agreement that is based on the index score which determined by age, comorbid conditions and medical procedures nearly guarantees that two records refer to the same individual. Probabilistic linkage utilized propensity score matching algorithm to determine whether two records should be linked based on the information in each record.

Probabilistic linkage is increasingly used in health research [5, 9-11]. Recently, Baser *et al*. probabilistically linked a large hospital database with outpatient claims from 2005 through 2007 to analyze anticoagulant bridging therapy use in patients undergoing hip and knee replacement surgery [12]. The hospital database contained information pertaining to medication use within the hospital, while the outpatient claims contained information about medication use outside of the hospital. The only way to analyze the bridging therapy was to incorporate a probabilistically linked database. Newgard *et al*. also evaluate the use of existing data sources, probabilistic linkage, and multiple imputations to build population based injury databases across phases of trauma. The performed probabilistic linkage of emergency medical system records to four hospitals and post discharge data sources and then handled missing values using multiple imputation [13].

Due to lack of independent gold-standard source of information against which the linkage can be judged, there are limited studies with reliability and accuracy of probabilistic record linkage, however, existing literature shows good validity identifying few false positives [14-18]. In outcomes research advantages and disadvantages of linkage must be weighted against each other before applying the technique. The results should be interpreted with caveat and provide guidelines for future data collection where information coming from the different datasets can be collected for same individual.

CONCLUSION

Probabilistic linkage can have important applications in outcomes research. In our case study, by linking registry data containing symptoms with claims data containing health care costs and utilization, we showed that women with menopause symptoms incur higher healthcare costs than those without symptoms.

AKNOWLEDGEMENT

Editorial support was provided by Elyse Fritschel of STATinMED Research.

REFERENCES

[1] McClellan M, Uncertainty, health-care technologies, and health-care choices. Am Econom Rev 1995; 38-44.

[2] Garrison LP, *et al*. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. Value Health 2007; 10(5): 326-35. http://dx.doi.org/10.1111/j.1524-4733.2007.00186.x

[3] Fellegi IP, Sunter AB. A theory for record linkage. J Am Statist Assoc 1969; 64(328): 1183-10. http://dx.doi.org/10.1080/01621459.1969.10501049

[4] Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Statist Assoc 1989; 84(406): 414-20. http://dx.doi.org/10.1080/01621459.1989.10478785

[5] Jaro MA. Probabilistic linkage of large public health data files. Statist Med 2007; 14(5-7): 491-98.

[6] Baser O, Crown WH, Pollicino C. Guidelines for selecting among different types of bootstraps. Curr Med Res Opin 2006; 22(4): 799-808. http://dx.doi.org/10.1185/030079906X100230

[7] CB W. Hormone therapy for the management of menopausal symptoms: pharmacotherapy update. J Pharm Pract 2010; 6(23): 540-47.

[8] RE W. Frequency and severity of vasomotor symptoms among peri-and post menopausal women in the United States. Climacteris 2008; 2008(11): 32-43.

[9] Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998; 20(1): 112-21. http://dx.doi.org/10.1093/oxfordjournals.epirev.a017966

[10] Adams MM, *et al*. Constructing reproductive histories by linking vital records. Am J Epidemiol 1997; 145(4): 339-48. http://dx.doi.org/10.1093/oxfordjournals.aje.a009111

[11] Whiteman D, *et al*. Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford Record Linkage Study Register. Am J Epidemiol 2000; 152(9): 823-28. http://dx.doi.org/10.1093/aje/152.9.823

[12] Baser O, *et al*. Anticoagulation prophylaxis practice patterns in patients having total hip, total knee replacement in a US health plan. Am Health Drug Benefits 2011; 4(4): 240-48.

[13] Newgard C, *et al*. Evaluating the Use of Existing Data Sources, Probabilistic Linkage, and Multiple Imputation to Build Population-based Injury Databases Across Phases of Trauma Care. Acad Emerg Med 2012; 19(4): 469-80. http://dx.doi.org/10.1111/j.1553-2712.2012.01324.x

[14] Coeli CM, *et al*. Probabilistic linkage in household survey on hospital care usage. Revista de Saúde Pública 2003; 37(1): 91-99. http://dx.doi.org/10.1590/S0034-89102003000100014

[15] Ford I. Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. J Clin Epidemiol 1995; 48(12): 1441-52. http://dx.doi.org/10.1016/0895-4356(95)00530-7

[16] Fair M, *et al*. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. Chronic Dis Can 2000; 21(1): 8-13.

[17] Ramsay C, Campbell M, Glazener C. Linking Community Health Index and Scottish morbidity records for neonates: the Grampian experience. Health Bull 1999; 57: 70-75.

[18] Shannon H, *et al*. Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. Can J Public Health. Revue canadienne de sante publique 1989; 80(1): 54.

Received on 17-11-2012 Accepted on 05-12-2012 Published on 31-01-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.01.04