drugs in context®
The journal of interventions in clinical practice
www.drugsincontext.com

OPEN ACCESS

ORIGINAL RESEARCH FULL TEXT ARTICLE

Probabilistic data linkage: a case study of comparative effectiveness in COPD

Christopher M Blanchette,[1,2] Mitch DeKoven,[3] Ajita P De,[3] Melissa Roberts[4]

[1]University of North Carolina, Charlotte, NC, USA; [2]Otsuka America Pharmaceutical Inc., Princeton, NJ, USA; [3]IMS Health, Alexandria, VA, USA; [4]Lovelace Clinic Foundation, Albuquerque, NM, USA

Abstract

Background: In this era of comparative effectiveness research, new, advanced techniques are being investigated by the research community to overcome the limitations of existing data sources. We describe the approach of probabilistic data linkage as a means to address this critical issue.

Methods: We employed a historical retrospective cohort design. Patients aged 40 and older with a principal or secondary diagnosis of COPD (ICD-9-CM codes 491.xx, 492.xx, and 496) and at least 3 years of continuous enrollment between January 1, 2004 and April 30, 2009 were selected from two US-based commercial administrative claims databases. The index date was designated as the date of the first claim (defined by a 12-month wash-out pre-index period) for the study drugs, for illustration purposes referred to as Treatment 1 or Treatment 2. The primary effectiveness measure was risk of any COPD-related exacerbation observed in the 12-month post-index period, with baseline characteristics being identified in the 12-month pre-index period.

Results: The percentage of the study sample receiving Treatment 1 at index who had an exacerbation was 39.3% for Database A and 39.7% for Database B; for Treatment 2, the percentages were 46.3% and 47.1%, respectively. The event rate of hospitalizations in each database sample was nearly identical as were the odds ratio and corresponding confidence intervals from the adjusted logistic regression models (OR – Database A: 0.72, Database B: 0.74, Database A with imputed outcomes: 0.72).

Conclusions: The probabilistic linkage demonstrated that patients from different databases matched on similar pre-index characteristics may demonstrate similar outcomes in the post-index period.

Keywords: data linkage, medical record linkage, comparative effectiveness research, treatment effectiveness, COPD, outcomes research, ambulatory care, prescription drugs.

Citation	Blanchette CM, DeKoven M, De AP, Roberts M. Probabilistic data linkage: a case study of comparative effectiveness in COPD. Drugs in Context 2013; 212258. doi: 10.7573/dic.212258
Provenance	Submitted; externally peer reviewed
Dates	**Submitted**: 18 September 2013; **Accepted, subject to peer review**: 21 September 2013; **Revised manuscript submitted**: 18 October 2013; **Published**: 31 October 2013

Correspondence address	Christopher Blanchette, PhD, MBA, Associate Dean for Research & Research Associate Professor, College of Health & Human Services, University of North Carolina at Charlotte, 9201 University City Blvd, CHHS 481, Charlotte, NC 28223, USA
Email address	cblanche@uncc.edu
Abbreviations	CER, comparative effectiveness research; COPD, chronic obstructive pulmonary disease; EMR, electronic medical record; ICD-9-CM, International Classification of Disease, 9th Revision, Clinical Modification; ISPOR, International Society for Pharmacoeconomics and Outcomes Research; PHI, personal health information.

Introduction

The ability to compare the relative effectiveness of different technologies and treatment modalities is now imperative in order to meet the growing demand for more rigorous outcomes research. Today, there are many data available to conduct comparative effectiveness research (CER) [1]: 1) data from surveys and 2) data from secondary databases; however, many lack important information. Both of these types of data sources have their own particular limitations. While prospectively collected observational cohorts lend themselves to longitudinal epidemiologic data for research, they are traditionally limited in external validity, specificity of measurement on key exposures and outcomes, and historially dated beyond application to current therapies. Secondary databases, including administrative claims data collected by the infrastructure of the healthcare system for the purposes of payment for services, also have limitations, as they are typically restricted to billing claims and provide limited clinical detail to confirm disease. Additionally, they often contain few patient characteristics (such as race and ethnicity) and have a low proportion of indigent and self-pay populations. These types of databases also lack in details on patients' perception of disease or the impact of disease or treatment on humanistic outcomes. Electronic medical record (EMR) systems, newly launched to replace paper-based chart systems in tracking clinical symptoms and disease, are still in the early stages of development. Moreover, they are frequently positioned in silos of healthcare practice –

traditionally physician practices with limited linkage to inpatient facilities, other specialists or other physician services, emergency and acute services, and pharmacy services.

While each of these datasets provides unique information, they lack a significant amount of information about the patients themselves and their associated treatments. Given these challenges and data limitations, there is a need for new, advanced techniques to overcome the limitations of existing data sources, thereby enhancing their utility for CER. Herein we describe the approach of probabilistic data linkage in addressing this critical issue.

Overview of data linkage

In the absence of the ideal data source, there is an alternative: a method whereby researchers can maximize the benefits of each individual type of data source, utilizing data linkage methods to fill in the gaps of one database with the strengths of another. Blanchette et al. presented this approach at a workshop during the 13th Annual European ISPOR Congress in Prague in 2010 [2], and discussed preliminary results demonstrating the application linking the two administrative databases, at a workshop during the 16th Annual International ISPOR meeting in Baltimore, in 2011 [3].

Database linkage methods generally employ both deterministic and probabilistic linking algorithms [4,5,6]. Deterministic linkage techniques can be applied when both datasets provide patient-identifying information that can be cross-matched between the two datasets. This is the preferred form of data linkage because the focus is on matching the same patient from both datasets. In the current climate of regulatory and legal restraints on the use of personal health information (PHI), the utility of this approach is limited. In the absence of PHI, a probabilistic method may be applied. However, the focus is not on identifying the same patient but rather on two matching patients with similar behavior and characteristics [7,8,9]. The goal of both of these efforts is to identify the best method for pair-matching between datasets.

Probabilistic linking methods incorporate varying strengths of identifiers, depending on information provided by the identifier. Stronger weights are given to matches for identifiers upon which matching is less likely. For example, matching on gender (male or female) is less likely to produce a matching of equivalent records than matching a record on the combination of birth month and year. Assigned weights are related to frequencies of occurrences of values in linked and unlinked pairs and also to the designated level of agreement [10]. The determination of agreement can be further defined into subcategories: agreement (yes/no) or varying levels of partial agreement (e.g., matching of a date within 2 days before/after or matching within a week before/after). Using probabilistic linking methods generally increases the sensitivity and specificity of models over a deterministic method in linking records from disparate datasets.

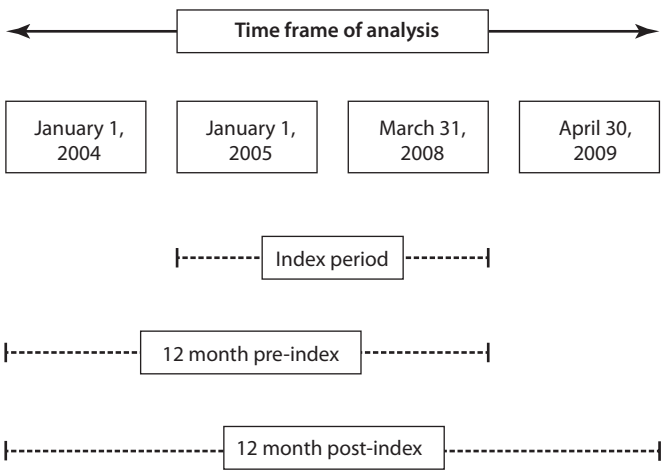
Propensity score methods are increasingly used to match similar individuals in observational studies, along with other approaches to control for channeling bias [11,12]. These methods are similar to the approach used in probabilistic data linkage. In propensity score matching, well-matched pairs are closely matched on the predicted probability of experiencing the dependent variable using relevant covariates. Matching techniques then employ a variety of approaches based on the predicted probability score (ranging from 0 to 1), including nearest neighbor matching heuristics, and often employing sampling without replacement [13,14].

Methodology

To demonstrate the utility and subsequent validity of probabilistic data linkage we employed a historical retrospective cohort design. Patients aged 40 and older with a principal or secondary diagnosis of COPD (International Classification of Disease, 9th Revision codes: 491.xx, 492.xx, and 496) and at least 3 years of continuous enrollment in a US health plan contributing to the database between January 1, 2004 and April 30, 2009 were selected from two administrative claims databases representing commercially insured US health plan enrollees. The index date was designated as the date of the first claim (defined by a 12-month wash-out pre-index period) for either study drug, Treatment 1 or Treatment 2. The effectiveness measure was the incidence of COPD-related exacerbation (hospitalization, emergency department visit or oral corticosteroid prescription) observed in the 12-month post-index period, with baseline characteristics being identified in the 12-month pre-index period (Figure 1).

Data linkage was performed by restricting the samples in each database to those with similar pre-index and post-index time periods (as described above). We then pooled data and developed a logistic regression model predicting the propensity to be in Database A using overlapping pre-index variables including age, geographic region, comorbidities, and pre-index utilization of health services and respiratory medications (Table 1). Patients from both databases were matched on the propensity score. We then directly imputed the outcomes of each matched patient from Database A to the matched patient from Database B. Finally, we conducted the comparative effectiveness study in each cohort (A, B, and A with imputed outcomes) (Figure 2).

Figure 1. Comparative effectiveness study design.



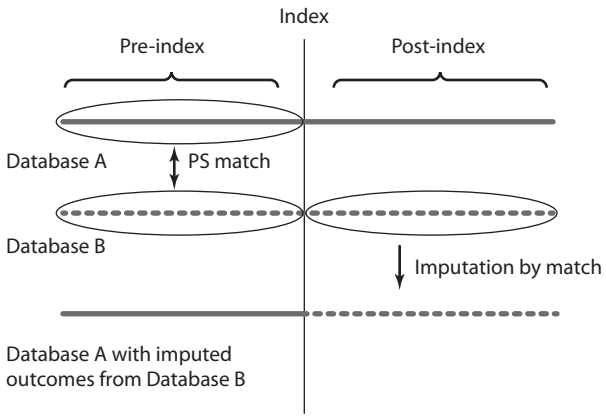
doi: 10.7573/dic.212258.f001

Table 1. Matched sample characteristics.

	Pre-match					Post-match				
	Database A		Database B			Database A		Database B		
	n	%	n	%	*p*-value	n	%	n	%	*p*-value
Total	12,926	100.0	21,334	100.0		11,040	100.0	11,040	100.0	
AGE 40-54	5,528	42.8	6,196	29.0	<0.001	4,419	40.0	4,516	40.9	0.18
AGE 55-64	6,074	47.0	8,148	38.2	<0.001	5,297	48.0	5,341	48.4	0.55
AGE 65-74	505	3.9	3,936	18.4	<0.001	505	4.6	443	4.0	0.04
AGE 75+	819	6.3	3,054	14.3	<0.001	819	7.4	740	6.7	0.04
Atrial fibrillation	632	4.9	1,443	6.8	<0.001	537	4.9	546	4.9	0.78
Arrhythmia	1,371	10.6	2,745	12.9	<0.001	1,162	10.5	1,152	10.4	0.83
Congestive Heart Failure	1,412	10.9	2,747	12.9	<0.001	1,195	10.8	1,181	10.7	0.76
Depression	1,582	12.2	2,044	9.6	<0.001	1,231	11.2	1,213	11.0	0.70
Fibrosis	433	3.3	824	3.9	0.014	388	3.5	399	3.6	0.69
Uncontrolled hypertension	6,189	47.9	10,465	49.1	0.035	5,203	47.1	5,221	47.3	0.81
East	2,277	17.6	2,967	13.9	<0.001	2,112	19.1	2,056	18.6	0.34
Midwest	6,175	47.8	5,285	24.8	<0.001	4,458	40.4	4,319	39.1	0.06
South	3,917	30.3	10,394	48.7	<0.001	3,913	35.4	4,069	36.9	0.03
West	557	4.3	2,682	12.6	<0.001	557	5.0	592	5.4	0.29

doi: 10.7573/dic.212258.t001

Figure 2. Data linkage methodology.



Abbreviation
PS, propensity score matching on similar observable variables. A multivariable binary logisitic regression model was used to calculate the predicted probability and nearest neighbor matching used to reduce the samples to similar patients.
doi: 10.7573/dic.212258.f001

Results

Of 12,926 patients in Database A and 21,334 patients in Database B, we matched 11,040 patients in each database on the propensity score using a nearest neighbor matching algorithm which matches like pairs of patients from each database without replacement. Prior to matching, all predictors were significantly different at $p<0.05$. Post-matching groups were balanced on all pre-index characteristics with the exception of older age (65–74 and >75 years) as well as representation in the Southern region.

The event rate of exacerbations in each cohort was nearly identical (Treatment 1 [Database A: 39.3%, Database B: 39.7%, and Database A with imputed outcomes: 40.1%] and Treatment 2 [Database A: 46.3%, Database B: 47.1%, and Database A with imputed outcomes: 46.8%]) as were the odds ratio and corresponding confidence intervals from the adjusted logistic regression models (OR – Database A: 0.72, Database B: 0.74, Database A with imputed outcomes: 0.72) (Table 2). This is not entirely surprising since the probability of being in each cohort was determined by the overlapping similarity in characteristics between the two cohorts, just as if the propensity score match was used in a traditional cohort analysis.

Discussion

The probabilistic linkage demonstrated that patients from different databases matched on similar pre-index characteristics may demonstrate similar outcomes in the post-index period. The implication of these findings and this approach is significant as it provides a case example for linkage of simulated cohorts from varying databases (e.g., EMR and administrative datasets, administrative datasets and surveys, EMR and surveys, etc.).

The creation of simulated cohorts could help explain a more complete picture of patient outcomes and allow researchers to

Table 2. Effectiveness outcomes.

Data	Rate of Exacerbation		Adjusted Results*		
	Treatment 1	Treatment 2	OR	95% CI	
Database A	39.3%	46.3%	0.72	0.66	0.79
Database B	39.7%	47.1%	0.74	0.67	0.81
Database A with Imputed Outcomes	40.1%	46.8%	0.72	0.65	0.80

*Risk associated with Treatment 1 compared to Treatment 2; adjusted for age, geographic region, comorbidities, and pre-index utilization of health services and respiratory medications using logistic regression.
doi: 10.7573/dic.212258.t002

explore all aspects of a patient experience – economic, clinical, and humanistic through the use of retrospective data analysis. For example, the use of administrative claims databases for comparative effectiveness research is often lacking key variables of interest such as income, race, lab results, clinical detail identified through notes, etc. Through this linking technique, one could impute these characteristics into a similar matched-patient cohort and assess the associations with the outcome. Similarly, a survey of patients could be augmented with key administrative and clinical detail. Even a clinical trial sample could be matched to administrative or electronic health record data to populate long-term outcomes. This technique has the ability to provide a view into the world of more complete data with further validation of the methods and use in diverse data sources.

There are several limitations worth noting related to this type of data linkage. The divergence of the measures between the databases may also cause a divergence in "like" patients thereby introducing error in the outcomes assessment. Care should be taken to ensure that similar measures are used to build the logistic regression model to create the propensity score; the assumption is that probabilistic matching produces two individuals with similar attributes and not the same individual. However, in the absence of complete data, the approach offers a good start and paves the way for further examples of this type of linkage method to be explored with various databases.

These data represent US-based administrative claims submitted from healthcare providers to health insurance plans and therefore there is a potential for miscoded claims. It is assumed that these patients fully utilize their health insurance benefits to avoid the full cost of the health service; however unlikely, there is a potential that the patient may be receiving health benefits from a provider who does not submit a claim for reimbursement to the patient's health insurance provider.

Conclusions

The probabilistic linkage method for simulated cohorts demonstrated that patients from different US-based health plan administrative claims databases matched on similar pre-index characteristics may demonstrate similar outcomes in the post-index period. This is the first application of this method for comparative effectiveness research and further validation in diverse datasets is required.

Contributions

Substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data (CMB, MR).

Drafting the article or revising it critically for important intellectual content (CMB, MD, APD, MR).

Final approval of the version to be published (CMB, MD, APD, MR).

Potential conflicts of interest

International Committee of Medical Journal Editors' (ICMJE) Potential Conflicts of Interests forms for each of the authors are summarized below. Original forms for each of the authors are available for download at:
http://drugsincontext.com/download/dic.212258_COI.pdf

CMB: Payment or receipt of services for any aspect of the submitted work: AstraZeneca funded the original study but did not fund the methodological process on which this paper focuses; financial activities outside the submitted work: consultancy for Centecor (not related); employee of Otsuka America Pharmaceutical Inc; grants/grants pending from Ethicon (not related), Bristol Myers Squib (not related), Auxilium (not related), Eli Lilly (not related), GlaxoSmithKline (COPD); stock/stock options from Celgene, Cubist, Gilead; no other relationships/conditions/circumstances that present a potential conflict of interest.

MK: No payment or receipt of services for any aspect of the submitted work; financial activities outside the submitted work: employee of IMS Health, Alexandria, VA, USA; no other relationships/conditions/circumstances that present a potential conflict of interest.

APD: No payment or receipt of services for any aspect of the submitted work; financial activities outside the submitted work: employee of IMS Health, Alexandria, VA, USA; no other relationships/conditions/circumstances that present a potential conflict of interest.

MR: No payment or receipt of services for any aspect of the submitted work; financial activities outside the submitted work: grants/grants pending from Astra Zeneca, GlaxoSmithKline, Boehringer Ingelheim, Pfizer; Lovelace Clinic Foundation, Albuquerque,

NM, USA has received funding for respiratory related healthcare research studies; no other relationships/conditions/circumstances that present a potential conflict of interest.

References

1. Federal Coordinating Council For Comparative Effectiveness Research – Report to the President and the Congress. June 30, 2009. Available at: http://www.med.upenn.edu/sleepctr/documents/FederalCoordinatingCoucilforCER_2009.pdf. [Last accessed: October 25 2013].
2. Exuzides AK, Blanchette CM, de Moor C. Imputation Techniques to Improve Data Availability from Electronic Medical Records. International Society for Pharmacoeconomics and Outcomes Research 13th Annual European Congress. November 6–9, 2010. Prague, Czech Republic.
3. Blanchette CM, Exuzides AK, Saunders WB, Stemkowski S. Advanced Missing Data Techniques in Observational Research: Case Studies in Data Linkage and Imputations. International Society for Pharmacoeconomics and Outcomes Research 16th Annual European Congress. May 21–25, 2011. Baltimore, MD, USA.
4. Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: Systematic review. Rev Saude Publica 2009;43:875–82. http://dx.doi.org/10.1590/S0034-89102009005000060
5. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14:491–8. http://dx.doi.org/10.1002/sim.4780140510
6. Lyons RA, Jones KH, John G, et al. The SAIL databank: Linking multiple health and social care datasets. BMC Med Inform Decis Mak 2009;9:3. http://dx.doi.org/10.1186/1472-6947-9-3
7. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81. http://dx.doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B
8. Rose S, van der Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. Int J Biostat 2009;5:Article 1. http://dx.doi.org/10.2202/1557-4679.1127
9. van der Laan MJ. Estimation based on case-control designs with known prevalence probability. Int J Biostat 2008;4:Article 17. http://dx.doi.org/10.2202/1557-4679.1114
10. Mason CA, Tu S. Data linkage using probabilistic decision rules: A primer. Birth Defects Res A Clin Mol Teratol 2008;82:812–21. http://dx.doi.org/10.1002/bdra.20510
11. Zhou Z, Lam P. Discussion of: Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. J Biopharm Stat 2007;17:25–27. http://dx.doi.org/10.1080/10543400601044782
12. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using sub-classification on the propensity score. J Am Statist Assoc 1984;79:516–24.
13. Gonzalez Smith ME, Storer JA. Parallel algorithms for data compression. JACM 1985;32:344–373. http://dx.doi.org/10.1145/3149.3152
14. Parsons LS. Reducing bias in a propensity score matched-pair sample using Greedy Matching techniques. 26th Annual SAS User Group International Conference. April 22–25, 2001. Cary, NC, USA. Paper 214–26.