PRZEGL EPIDEMIOL 2003;57:553–9

Włodzimierz Borkowski, Hanna Mielniczuk

ŁĄCZENIE REKORDÓW NA POTRZEBY ANALIZ EPIDEMIOLOGICZNYCH

Centrum Medyczne Kształcenia Podyplomowego
Dyrektor: Jadwiga Słowińska-Srzednicka

W artykule przedstawiono zagadnienie łączenia rekordów zbiorów publicznych metodą deterministyczną oraz probabilistyczną. Omówiono obciążenia (bias) wyników analiz przeżyciowych wynikające z łączenia. Przedstawiono własną modyfikację łączenia deterministycznego zawierającą komponent losowego wyboru i dokonano analizy jakości tej metody. Przedyskutowano przydatność i ograniczenia stosowania łączenia w badaniach epidemiologicznych.

Słowa kluczowe: *łączenie deterministyczne, łączenie probabilistyczne, obciążenie selekcji, analiza przeżyć*
Key words: *deterministic record linkage, probabilistic record linkage, selection bias, survival analysis*

WSTĘP

Informatyzacja medycyny znajduje odbicie w sposobie gromadzenia i przechowywania danych zdrowotnych. Coraz powszechniejsze jest przechowywanie informacji dotyczących osób lub zjawisk zdrowotnych w odrębnych zbiorach. Łączenie rekordów jest podstawowym mechanizmem integracji informacji w bazach danych wykorzystywanych w systemach szpitalnych, czy też medycznych hurtowniach danych.

Można rozróżnić zasadniczo dwie metody łączenia: deterministyczną i probabilistyczną.

Podstawą łączenia deterministycznego jest identyczność wybranych pól w łączonych rekordach zwanych polami kluczowymi. Przykładowo do łączenia szpitalnej historii choroby z badaniami laboratoryjnymi kluczem może być PESEL, o ile będzie zamieszczony zarazem w historii choroby, jak i w wynikach badań laboratoryjnych. REGON może być podstawą do łączenia danych o szpitalu z kasy chorych i sprawozdawczości państwowej, jeśli będzie się znajdował w obu zbiorach. Możliwe jest osłabienie wymogu całkowitej zgodności kluczy pomiędzy łączonymi rekordami, jak to ma miejsce w metodzie SOUNDEX, gdzie podstawą porównania jest zgodność fonetyczna kluczy (1).

Nie zawsze jest możliwe zdefiniowanie kluczy umożliwiających deterministyczne łączenie. Ważną przyczyną mogą być ograniczenia prawne mające na celu ochronę i poufność danych zdrowotnych. W Polsce w rejestrze zgonów Głównego Urzędu Statystycznego

(GUS) jest pominięty numer PESEL zmarłego. GUS nie udostępnienia także pełnej daty urodzenia osób zmarłych. Szpitalne historie choroby w wersji elektronicznej, udostępniane na zewnątrz, są pozbawione informacji pozwalających na identyfikację pacjenta.

Poważnym problemem utrudniającym łączenie mogą być błędy w kluczach. Błędy literowe uniemożliwiają połączenie rekordów prowadząc do utraty materiału badawczego.

W sytuacjach, kiedy zawodzi metoda deterministyczna, można stosować probabilistyczną metodę łączenia dopuszczającą niezgodności między polami kluczowymi łączonych rekordów. W analizach epidemiologicznych łączenie probabilistyczne ma długą historię (2, 3, 4, 5). Szczególnie często znajdowało ono zastosowanie w rejestrach onkologicznych dla badań przeżyciowych i śledzenia losów pacjenta (6, 1) oraz do kontroli kompletności rejestrów metodą *capture-recapture*, czy też eliminacji błędnie powtarzających się rekordów.

Podstawą łączenia probabilistycznego jest podobieństwo wartości pól składających się na klucz identyfikacyjny. Opierając się na pojęciach analogicznych do czułości i swoistości testów epidemiologicznych ustala się wagi cząstkowe dla poszczególnych pól. Na podstawie tych wag cząstkowych oblicza się wagę ogólną będącą miarą podobieństwa między kluczami łączonych rekordów oraz wartości progowe dla dokonania połączenia. Podstawy matematyczne metodologii łączenia probabilistycznego oraz omówienie czynników warunkujących jakość dopasowania przytacza M. A. Jaro (4).

W trakcie łączenia probabilistycznego dochodzi do utraty materiału (dla fałszywie negatywnych połączeń) oraz zjawisk dających obciążenie selekcji statystyk na połączonych rekordach. Zależnie od rodzaju badań, jakie przeprowadza się na połączonych rekordach, zależy rodzaj i wielkość obciążenia. Przykładowo, badania przekrojowe czynników społeczno-zdrowotnych wśród osób zmarłych w danym roku można dokonać na połączonych rekordach zgonów GUS z rekordami ze spisu powszechnego. Przy założeniu losowego występowania błędów w kluczach nie pojawia się obciążenie selekcji (*selection bias*).

W przypadku analiz przeżycia sytuacja jest inna, ponieważ do wszystkich rekordów zbioru zgłoszeń jest dołączana informacja o losie pacjenta. Nie znalezienie rekordu w zbiorze zgonów dla danego rekordu ze zbioru zgłoszeń jest równoznaczne z informacją o przeżyciu pacjenta. Nawet losowy, fałszywie ujemny wynik dołączania prowadzi do obciążenia wyników analiz przeżycia, wyrażających się zmniejszeniem częstości zgonów i wydłużeniem średniego czasu przeżycia. Z drugiej strony „nadmiar" możliwych do dołączenia rekordów zgonu prowadzi do zwiększenia połączeń fałszywie dodatnich i zawyżenia częstości zgonów oraz skrócenia średniego czasu przeżycia.

MATERIAŁ I METODY

W pracy były łączone dwa zbiory populacyjne. Pierwszym z nich był zbiór zgłoszeń chorób nowotworowych u dzieci i młodzieży w wieku 0–19 lat z lat 1990–1993, o liczebności 5287, uzyskany z Krajowego Rejestru Nowotworów (KRN). Drugim był zbiór zgonów na choroby nowotworowe (przyczyny zgonów kod ICD-9 140-204) dzieci i młodzieży w wieku 0–25 lat, o liczebności 7362, uzyskany ze zbioru wszystkich zgonów GUS z lat 1990–1998, o liczebności 123 963. Danymi używanymi do identyfikacji osoby, wobec braku unikatowego identyfikatora, są data urodzenia, imię i nazwisko. Zarówno w zbiorze KRN jak i GUS nie było identyfikatorów PESEL. GUS powołując się na ustawę o ochronie danych osobowych nie udostępnił imienia i nazwiska zmarłych dzieci oraz utajnił dzień

miesiąca w dacie urodzenia. Dane zawarte w zbiorze KRN takie jak imię, nazwisko i pełna data urodzenia nie mogły być wykorzystane do łączenia. Powstał problem łączenia zbiorów w sposób minimalizujący straty materiału i zapobiegający powstawaniu obciążenia w analizach przeżyciowych. Zakres informacji zawartej w rekordach KRN i GUS ograniczył możliwości utworzenia klucza do łączenia. Ostatecznie na klucz złożyły się: miesiąc i rok urodzenia, płeć, rodzaj nowotworu, województwo zamieszkania (49 województw w/g podziału administracyjnego sprzed 1999 roku). Rodzaj nowotworu w zbiorze KRN był określony na podstawie rozpoznania choroby przy zgłoszeniu w/g ICD-9. W zbiorze GUS rodzaj nowotworu był określony na podstawie przyczyny zgonu w/g ICD-9. Licząc się z różnicami w rozpoznaniu nowotworu w chwili zgłoszenia i zgonu zrezygnowano z dokładnego kodu ICD-9 choroby nowotworowej, przyjmując podział na dwa rodzaje: guzy lite i nowotwory układowe.

Tak utworzony klucz nie był unikatowy zarówno w zbiorze KRN jak i GUS, co było przyczyną wprowadzenia modyfikacji do metody deterministycznej. Modyfikacja polegała na losowaniu rekordu w razie powtórzeń rekordów o tej samej wartości klucza w/g następującego algorytmu:

1. Losowy wybór rekordu ze zbioru KRN.
2. Znalezienie w zbiorze GUS rekordów o tej samej wartości klucza jak rekord z pkt. 1 spełniającego dodatkowo warunki: data zgonu nie może być wcześniejsza od daty zgłoszenia w rekordzie KRN z pkt. 1 i nie późniejsza niż 5 lat od tej daty.
3. Przeprowadzenie łączenia wg zasad:
   a) jeśli w pkt. 2 został znaleziony jeden rekord GUS – dołączenie do rekordu KRN z pkt1 faktu zgonu, daty zgonu, pól tego rekordu GUS;
   b) jeśli w pkt. 2 zostały znalezione dwa lub więcej rekordy – losowy wybór jednego z nich ze zbioru pkt. 2 i dołączenie do rekordu KRN z pkt. 1 faktu zgonu, daty zgonu, pól tego rekordu GUS;
   c) Jeśli w pkt. 2 nie został znaleziony żaden rekord GUS – dołączenie do rekordu KRN z pkt. 1 faktu przeżycia.
4. Wyłączenie z dalszego postępowania łączenia rekordów KRN i GUS opracowanych w pkt. 3.

Powyższa procedurę powtarza się do momentu opracowania wszystkich rekordów w zbiorze KRN.

Należy zaznaczyć, że w zbiorze KRN był zanotowany zgon zaistniały maksymalnie 2 lata od zgłoszenia nowotworu oraz data tego zgonu. Zbyt krótki okres obserwacji a również niekompletność rejestracji zgonów w KRN uniemożliwiały przeprowadzenie zaplanowanych pięcioletnich analiz przeżyciowych na tym zbiorze i konieczne było wykorzystanie informacji o zgonach z niezależnego źródła danych a mianowicie rejestru zgonów GUS. Jednak informacja o fakcie i dacie zgonu zawarta w KRN została wykorzystana do weryfikacji zastosowanej metody łączenia. W części materiału KRN, gdzie były odnotowane zgony porównano fakt zgonu i wartości dat w zbiorze KRN i zbiorze wynikowym uzyskanym po zastosowaniu łączenia.

WYNIKI

Sprawdzono unikatowość rekordów przy ustalonym kluczu (miesiąc urodzenia, rok urodzenia, płeć, rodzaj nowotworu, województwo zamieszkania) w łączonych zbiorach.

Zarówno w zbiorze KRN jak i GUS występowały nieunikatowe rekordy: odpowiednio w 19,1% i 15,9% (tab. I).

Tabela I. Liczebność rekordów w zależności od powtarzalności wartości klucza
Table I. Number of records according to recurrence of key values

Powtarzalność wartości klucza	Zbiór KRN		Zbiór GUS	
	liczebność	procent	liczebność	procent
Unikatowe	4275	80,9%	6188	84,1%
Podwójne	810	15,3%	980	13,3%
Trzy i więcej	202	3,8%	194	2,6%
Razem	5287	100%	7362	100%

W wyniku łączenia zbioru KRN zawierającego 5287 rekordów ze zbiorem GUS zawierającym 7362 rekordy otrzymano zbiór wynikowy, w którym status „zmarł" uzyskało 1581 przypadków (tab. II).

Tabela II. Wyniki łączenia zbioru KRN i zbioru GUS
Table II. Results of the linkage of KRN and GUS files

Status pacjenta	Liczba przypadków	Procent
Żyje	3706	70,1%
Zmarł	1581	29,9%
Razem	5287	100%

Sprawdzono w jak dużym stopniu zastosowana metoda dołącza zgony do rekordów, w których był zarejestrowany zgon w zbiorze KRN (tab. III).

Tabela III. Dołączenie faktu zgonu do rekordów z odnotowanym w zbiorze KRN zgonem
Table III. Linked facts of death to records flagged in KRN file as death

Kryterium	Wynik	Liczba	Procent
Dołączenie faktu zgonu	tak	848	68,1
	nie	397	31,9
	razem zgonów w KRN	1245	100%

Porównywano zgodność dat zgonów zbiorze KRN i w zbiorze wynikowym, ażeby sprawdzić czy zgony zostały dołączone do należnego im osobnika (tab. IV).

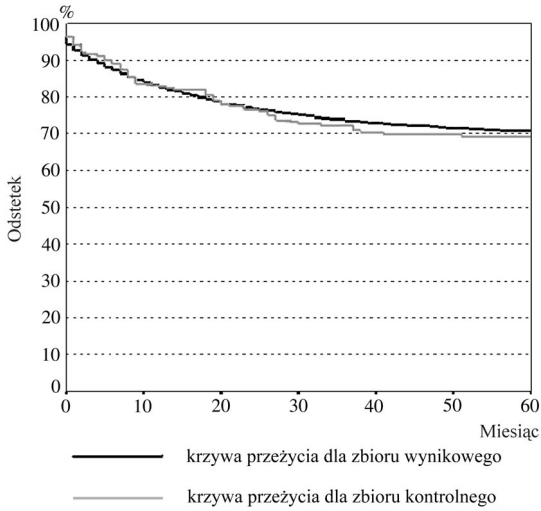
Tabela IV. Zgodność dat zgonów dołączonych i odnotowanych w zbiorze KRN
Table IV. Identity of death dates linked and flagged in KRN file

Kryterium	Wynik	Liczba	Procent
Zgodne daty	tak	731	58,7%
	nie	514	41,3%
	razem zgonów w KRN	1245	100%

Otrzymany 68,1% zgodności zgonów oraz 58,7% zgodności daty jest konsekwencją braku unikatowości klucza dotyczącego blisko 20% rekordów zarówno zbioru KRN jak i GUS. Należy zaznaczyć, że uzyskanych rekordów wynikowych nie można rozpatrywać jednostkowo dla śledzenia losów pojedynczego pacjenta. Natomiast wnioskowanie populacyjne jest uprawnione pod warunkiem uniknięcia obciążenia selekcji.

O poprawności łączenia wnioskowano empirycznie poprzez porównanie statystyk analiz przeżycia pięcioletniego na naszym materiale ze statystykami uzyskanymi na materiale kontrolnym. Materiałem kontrolnym był zbiór 189 przypadków dziecięcych z rejestru chorób nowotworowych Województwa Mazowieckiego z roku 1996. W zbiorze tym starannie odnotowano zarówno zgłoszenia z roku 1996 jak i dalsze losy tych pacjentów do roku 2001. Krzywe przeżycia Kaplana-Mayera oraz 5-letnie przeżycia obliczone na zbiorze wynikowym i na materiale kontrolnym okazały się bardzo podobne (ryc. 1).

Materiał badawczy uzyskany w wyniku łączenia został wykorzystany do dalszych analiz pięcioletnich przeżyć dzieci chorych na choroby nowotworowe (7).



Ryc. 1. Porównanie krzywych przeżycia uzyskanych na zbiorze wynikowym i kontrolnym.
Fig. 1. Comparison of survial curves obtained from linked and control files.

DYSKUSJA

Wraz z rozwojem hurtowni danych medycznych coraz częstsza jest sytuacja łączenia rekordów w sposób zautomatyzowany. Odnosi się to do dużych zbiorów publicznych w opiece zdrowotnej (8) w tym do rejestrów onkologicznych. Uprzednie podejście odwołujące się do czynnika ludzkiego przy łączeniu tzw. ręcznym, jest zastępowane zautomatyzowanymi działaniami mającym dobre podstawy matematyczne (9), realizowanymi przez powstające programy komercyjne (10, 4).

Problem łączenia jest złożonym zagadnieniem teoretycznym, szczególnie trudne są sprawy związane z obciążeniami. Powtarzające się wartości klucza są uciążliwością dla badaczy nawet, gdy występują w małej skali (11). W USA łączenia zbiorów urodzeń i zgonów

niemowląt na podstawie identyfikatorów pacjentów w opiece zdrowotnej są wykonywane przez służby państwowe a zbiory do analiz udostępniane bezpłatnie poprzez Internet (12).

W naszym przypadku łączenie probabilistyczne nie rozstrzyga problemu powtórzeń wartości kluczy w materiale. Zastosowana metoda łączenia rozwiązująca problem powtórzeń pozwala uzyskać materiał do analiz statystycznych pod warunkiem uniknięcia obciążeń selekcji. Teoretyczne oszacowanie wielkości obciążenia wymagałoby rozpatrzenia złożonego modelu matematycznego uwzględniającego rozkłady występowania chorób nowotworowych u dzieci, rozkłady czasu przeżycia, rozkłady wieku i miejsca zamieszkania. Rozkłady te były nieznane i trudne do oszacowania. Dlatego zastosowano empiryczną ocenę poprawności łączenia dla potrzeb analiz statystycznych.

Należy jednak rozważyć czynniki warunkujące powstawanie obciążenia. W pracy braliśmy pod uwagę zachorowania na nowotwory w latach 1990–1993. Ponieważ interesowały nas przeżycia 5-letnie uwzględniliśmy zgony na nowotwory z lat 1990–1998. Jednakże ten zbiór zgonów dotyczy również nowotworów, dla których zachorowanie nastąpiło przed rokiem 1990 oraz po 1993. W porównaniu do liczby zgłoszeń nowotworów z lat 1990–1993 powstał nadmiar potencjalnych do połączenia zgonów. Przy braku unikatowości klucza mogło to prowadzić do fałszywie dodatniego wyniku łączenia, przez to do „przeszacowania" liczby przeżyć kończących się zgonem.

Rekord zgłoszenia rozpatrywany w pierwszej kolejności ma większą szansę na dołączenie rekordu zgonu niż kolejne. Zastosowana metoda łączenia eliminuje obciążenie wywołane przez to zjawisko (zapobiega nierównomiernemu rozkładowi w materiale takich sytuacji).

W przypadku analiz dla szczegółowych rozpoznań, użyty klucz uwzględniający tylko podział na dwa rodzaje nowotworów nie jest właściwy, bo powoduje obciążenie związane z mechanizmem selekcji – fałszywie dodatnie wyniki łączenia występują częściej dla nowotworów rzadziej występujących. Fakt ten ilustruje konieczność uwzględnienia rodzaju planowanych analiz przy konstruowaniu klucza, ewentualnie przeprowadzanie łączenia na podzbiorach danych.

WNIOSKI

1. Zastosowanie prostej metody będącej rozszerzeniem podejścia deterministycznego o składową losową na zbiorach zgłoszeń chorób i zgonów na choroby nowotworowe u dzieci dało materiał pozwalający na wiarygodne epidemiologiczne analizy przeżyciowe.
2. Poprawne przeprowadzenie procesu łączenia niedeterministycznego wymaga zrozumienia mechanizmów selekcji w celu uniknięcia obciążeń wyników, co wymaga rozważań teoretycznych i doświadczenia praktycznego.

W Borkowski, H Mielniczuk

USAGE OF RECORD LINKAGE IN EPIDEMIOLOGICAL ANALYSES

SUMMARY

The article presents two methods of records linkage, deterministic and probabilistic, applied in epidemiological studies and medical registers. Sources of selection bias linkage was considered. The modified deterministic linkage by random selection of one from nonunique records was elaborated.

The quality of applied method of record linkage in survival analysis was empirically evaluated. Usefulness of record linkage in epidemiology was talked over.

PIŚMIENNICTWO

1. Automated data Collection in Cancer Registration Ed by R. J. Black, L. Simonato HH and E. Demaret IARC Technical Reports Lyon 1998;32:7–11.
2. Newcombr H. B, Smith M. E, Howe G. R, Mingay J., Strugnell A., Abbat J. D. Reliability of computerized versus manual death searchers in a study of the health of Eldorado uranium workers Comput Biol Med 1983;13:157–69.
3. Newcombe HB. Handbook of Record Linkage Oxford, Oxford University Press 1988.
4. Jaro M. A. Probabilistic linkage of large public health data files Statiscics in Medicine 1995;14: 491–8.
5. Shevchenko I, Lynch J T, Mattie A, Reed-Fourqet L. Verification of information in a large medical database using linkages with external databases Statistics in Medicine 1995;14:511–30.
6. Blakely T. Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991–94 mortality records The New Zealand Census-Mortality Study Department of Public Health Wellington School of Medicine; December 1999.
7. Borkowski W, Mielniczuk H. Childhood cancer survival in Poland – analysis of linked public health data files, Med Sci Monit – w druku.
8. Hawkins MM, Swerdlow AJ. Completeness of cancer and death follow-up obtained through the National Health Service Central Register for England and Wales 1992;66:408–13.
9. Smith ME, Newcombe HB. Accuracies of computer versus manual linkages of routine health records. Methods Infor Med 1979;18:89–97.
10. Roos LL, Wajda A, Sharp SM, Nicol JP. Software for heath care analyst a modular approach J Med Sys 1987;1:445–501.
11. Herrchen B, Gould J B, Nesbitt T S. Vital Statistics Linked Birth/infant Death and Hospital Discharge Record Linkage for Epidemiological Studies, Comp Biomed Research 1997;30:(4):290–305.
12. NCHS's Perinatal Mortality Data http://www.nber.org/data/perinatal.html

Adres autorów:

Włodzimierz Borkowski,
Centrum Medyczne Kształcenia Podyplomowego
ul. Marymoncka 99, 01-813 Warszawa
e-mail: bioinfo@cmkp.edu.pl