Buscando una aguja en un pajar: las técnicas de conexión de registros en los sistemas de información sanitaria

Pere Arribas, Eva Cirera y Manuel Tristán-Polo

Agència de Salut Pública de Barcelona (antes Institut Municipal de Salut Pública). Barcelona. España.

La conexión de registros de 2 o más bases de datos que tienen información complementaria de un tema permite identificar y conectar los registros que corresponden a un mismo individuo. Puede hacerse de forma manual, determinista o probabilística. La conexión probabilística se desarrolló para situaciones en que no existe un identificador único y el número de variables identificadoras es limitado y con poco poder discriminatorio. Considerando la posibilidad de que la conexión sea correcta o incorrecta y teniendo en cuenta el grado de acuerdo de las variables y sus frecuencias, este proceso pondera cada variable. En este manuscrito se presentan estas técnicas, se describe la operativa de la conexión probabilística paso a paso y se ilustra con ejemplos extraídos del funcionamiento de nuestros servicios.

Palabras clave: Conexión de registros. Conexión probabilística. Sistemas de información sanitaria.

Looking for a needle in a haystack: record linkage techniques in health information systems

Linking records from two or more data sets with information on an issue allows us to identify those belonging to the same individual. The linkage process can be manual, deterministic, or probabilistic. Probabilistic linkage was developed for those situations where there is no single identifier, the number of identifying variables is limited, and they have little discriminating power. Considering the possibilities for the linkage to be correct or incorrect and based on the degree of agreement among variables and their frequencies, this process weights each variable. This manuscript presents linkage techniques, describes probabilistic linkage step by step, and illustrates it with examples taken from the actual operation of our services.

Key words: Record linkage. Probabilistic linkage. Health information systems.

Introducción

La utilización de métodos de conexión de registros se está convirtiendo en una práctica creciente en el ámbito de la biomedicina, en el que se trabaja con grandes cantidades de información[1]. La conexión de registros de 2 o más bases de datos que tienen información complementaria de un cierto tema permite identificar y conectar los registros que corresponden a una misma entidad, generalmente un individuo. Cuando existe un identificador unívoco la conexión es trivial (es lo que hace la Agencia Tributaria con el Número de Identificación Fiscal cuando cruza datos fiscales de los contribuyentes), pero en ocasiones, debido a que los datos utilizados para la conexión no se han recogido para ese propósito, o por motivos de confidencialidad[2], no existe ese identificador y es necesario utilizar más de una variable para discriminar entre individuos. Aun así, en muchos casos no puede hacerse de manera unívoca[3]. Existen diferentes métodos para realizar este proceso[4]. Los criterios utilizados para elegir la más apropiada a cada situación son tres: la sensibilidad, la resolución y la rapidez, que se comentan en la tabla 1.

TABLA 1

Criterios de elección para la selección del mejor método de conexión de registros en cada situación

Sensibilidad (o precisión)	Capacidad de discriminar entre un registro perteneciente a un individuo y los que pertenecen a otros individuos
Resolución	Potencia para decidir si las discrepancias en los pares de registros se deben a errores en los datos o a que no pertenecen al mismo individuo
Rapidez	Para procesar grandes volúmenes de datos en un tiempo razonable

Métodos de conexión

El método más simple y válido cuando se tienen pocos registros y poca información es la conexión manual. Con este método un observador puede evaluar (con criterios básicamente heurísticos) la concordancia de cada registro de una de las bases de datos con todos los registros de la otra, para seleccionar el que presente mayor semejanza. El problema fundamental es que el número de errores en la evaluación de la concordancia se incrementa a medida que aumenta el volumen de datos a conectar. Además, puede suceder que se enlacen registros que concuerden con mayor probabilidad con alguno de los anteriormente enlazados[5]. Tampoco se garantiza que el observador mantenga un criterio sistemático (aunque algunos razonamientos poco formalizados de un observador «humano» pueden en ocasiones emparejar registros que difícilmente podrían relacionarse con los otros métodos).

Otro método es la conexión determinística, en la que se establece la concordancia mediante un identificador único o una combinación de variables[6]. Idealmente, las variables utilizadas como identificadores tendrían que ser únicas (para que las conexiones fueran precisas), universales (para que el mismo valor sea posible en cada registro del mismo individuo) y permanentes (para que su valor no cambie a lo largo del tiempo)[3]. El número de la Seguridad Social en los EE.UU. o el Documento Nacional de Identidad en España cumplirían, en principio, estos requerimientos. Este método valora la concordancia a partir de la coincidencia más o menos estricta de las variables de identificación disponibles. Pero cuando el volumen de datos es grande, la calidad de los datos es baja y/o existen divergencias entre los dominios de algunas variables utilizadas como identificadores, este método resulta poco eficiente.

Correspondencia: P. Arribas.
Servei d'Informàtica. Agència de Salut Pública de Barcelona.
Pl. Lesseps, 1. 08023 Barcelona. España.
Correo electrónico: parribas@aspb.es

El tercer método es la conexión probabilística, que se desarrolló para las situaciones en que no existe un identificador único y el número de variables identificadoras es limitado y con poco poder discriminatorio[3]. En esencia, este método trata de unir registros con la máxima probabilidad de pertenecer al mismo individuo[7]. Se basa en 2 conceptos: la posibilidad de que la conexión sea correcta (en cuyo caso hay que tener en cuenta que dos registros del mismo individuo pueden tener valores diferentes en alguna de las variables identificadoras) y la posibilidad de que la conexión sea incorrecta (en este caso hay que considerar que dos individuos pueden tener valores coincidentes en algunas variables identificadoras)[3]. Considerando estos factores y teniendo en cuenta el grado de acuerdo de las variables y de sus frecuencias, el proceso utiliza un «peso» que determina la fiabilidad de la concordancia para cada variable.

En los tres métodos anteriores, el proceso de conexión se basa en la comparación más o menos condicionada de cada registro de una base de datos con todos los registros de la otra. Esto puede ser un problema para volúmenes grandes de información, debido a la cantidad de comparaciones que se han de realizar. Por ejemplo, si tenemos 1.000 registros de servicios de ambulancia y 1.000 registros de ingresos hospitalarios para conectar, la comparación de cada registro de una lista con cada registro de la otra nos da un total de 1.000.000 de comparaciones (el producto cartesiano). Esto sería inabordable con el método de conexión manual. El método de conexión determinista sería el más eficiente si tuviéramos identificadores únicos de buena calidad, como la fecha de nacimiento, el nombre o algún número de documento identificativo. Pero es fácil imaginar que esos datos pueden faltar, estar incompletos o ser parcialmente erróneos: pensemos que suelen introducirse de forma manual y sin validaciones o con validaciones incompletas. En este caso, el método de conexión determinista (en principio óptimo) obtiene resultados pobres y son más adecuados los métodos de conexión manual y probabilística (obviando el problema del tiempo de proceso) por su característica de «tolerancia a fallos».

Para reducir el número de comparaciones y, por ende, el tiempo de proceso, existen varias técnicas que consisten en limitar el número de registros sobre los que se realiza la comprobación. Una es la partición de los registros basándose en una variable informada y de buena calidad. En el ejemplo y bajo la suposición razonable de que el dato del mes está correctamente introducido, los registros se podrían separar por meses. Si suponemos una distribución uniforme en cada bloque de registros habrá 1.000/12 = 83,3 registros, lo que da un total de 6.944 comparaciones a realizar, que por 12 meses daría un total de 83.333 comparaciones. Este método de partición se puede llevar al extremo (semanas, incluso días), si bien puede ocurrir que algún registro se incluya en un bloque que no corresponda, sin que pueda compararse con el supuesto registro gemelo si éste se encuentra en otro bloque. Esto suele ocurrir cuando se utilizan variables continuas para segmentar, como pueden ser las relacionadas con el tiempo. En el ejemplo expuesto puede ocurrir que un registro de un servicio de ambulancia se realice poco antes de las 12 de la noche y el ingreso se produzca después de las 12, habiendo utilizado el día como condición de partición. Otro método de partición se basa en el principio de causalidad. En el ejemplo no es difícil suponer que la fecha/hora de ingreso en el hospital será mayor que la fecha/hora de evacuación en la ambulancia. Además, se puede suponer que la ambulancia no tardará más de unas horas en llegar al hospital. Con esto conseguimos dividir extraordinariamente el total de registros, lo que resulta en una reducción del tiempo y del esfuerzo en el proceso de comparación.

La combinación de estas técnicas de segmentación con una mayor tolerancia a imprecisiones en los datos a enlazar nos lleva a determinar que el método de conexión probabilística es el más adecuado para poder realizar el proceso de conexión. Incluso cuando existen identificadores únicos y/o variables de buena calidad, absorbe mejor las imprecisiones en los datos de las variables y no se ve penalizado en exceso por el volumen de datos.

El proceso de conexión probabilística

El proceso se inicia con la selección y preparación de las bases de datos a enlazar y un análisis de la distribución de frecuencias de las variables candidatas a utilizarse como identificadoras. Es importante que la distribución de frecuencias sea aproximadamente igual en los dos archivos. Se deben seleccionar las variables en común de los dos archivos que tengan un grado de fiabilidad óptimo. Dentro de este análisis se ha de valorar el grado de calidad, considerando la proporción de valores desconocidos y de posibles errores de codificación.

Valoración de la factibilidad del proceso

Una vez conocidas las variables candidatas, hay que valorar si el proceso será o no factible. Una manera de hacerlo es multiplicando el número de categorías diferentes de cada una de las variables que se utilizarán. Este producto ha de ser siempre superior al número de registros totales de las 2 bases de datos[1]. En el ejemplo, si utilizamos como identificadores el sexo, la fecha de accidente y la edad, el resultado sería 2*97*365 = 70.810, que es mayor que 1.000 + 1.000, por lo que el proceso es factible con estas variables.

Restricción de los pares de comparación

Ahora se tendría que determinar la condición que se debe aplicar para hacer los bloques de comparación, con el fin de acotar el número de comparaciones[1]. Algunos autores recomiendan que estos bloques no estén formados por más de 100 registros ni menos de 20[8]. Para establecer esta condición es importante basarse en las variables de mayor calidad en los datos y que posea el mayor poder discriminatorio a fin de identificar perfectamente esos bloques. Es aconsejable disponer de varias condiciones con objeto de emplearlas en caso de que la utilización de una de ellas deje registros sin enlazar[6].

Etapa de comparación

Consiste en confrontar las variables identificadoras de un registro con las variables identificadoras de los registros del bloque que serán los candidatos a ser conectados. El método de conexión probabilística combina información aportada por las diferentes variables de conexión para determinar si dos registros de ficheros diferentes corresponden o no al mismo caso, dependiendo de qué variables coinciden y cuáles no y con qué valores. Para evaluar este grado de parentesco entre pares de registros, la teoría de la conexión de registros probabilística ofrece un método que consiste en calcular un peso (W) = $f(R_1,R_2)$ en función de las variables identificativas utilizadas de cada registro (R_1 y R_2) que indica para cada par de registros la probabilidad de que sean del mismo caso[1]. En una situación ideal, si W supera un cierto valor umbral, se podría concluir que los dos registros pertenecen al mismo caso, y si W está por debajo de ese

valor, ambos registros no formarán pareja. En realidad, ese punto de corte no está tan definido y es posible que pares discordantes tengan un peso por encima del umbral de corte y que pares concordantes lo tengan por debajo. Es por esto que se debe trabajar con dos umbrales, que delimitarán tres categorías: casos de registros concordantes, casos de registros no coincidentes y casos dudosos.

Cálculo de los pesos específicos

Como se ha comentado, para determinar la conexión o no de dos registros, se pueden utilizar una o más variables, que a partir de ahora denominaremos conectoras o de conexión. Para cada una de estas variables calcularemos un peso parcial (W_p), la suma de los cuales determinará el peso total (W_t) para cada par de registros. Existen diferentes técnicas para calcular el peso (W_p), una de ellas se basa en la probabilidad condicional de un par dada la distribución de las variables conectoras[9].

Cada variable tiene dos probabilidades asociadas: la probabilidad que denominaremos (m) es la probabilidad de que las variables coincidan sabiendo que el par revisado es un par coincidente y la probabilidad (u) es la probabilidad de que un par coincida sabiendo que el par revisado no es un par coincidente. La probabilidad (m) tendrá el valor que resulte de restar a 1 la proporción de error asociada a aquella variable. Por ejemplo, si sabemos que una muestra de pares coincidentes para la variable sexo tiene un porcentaje de datos no especificados del 15%, la probabilidad (m) asociada a esta variable será de 1 – 0,15 = 0,85. En la mayoría de los casos, si no se conoce la proporción de error asociada, se puede operar arbitrariamente con el 10%. La probabilidad (u) corresponde con la probabilidad de que los registros coincidan al azar y se calcula con la distribución de la variable conectora. Además, (m) ha de ser siempre mayor que (u). De no ser así, la probabilidad de que dos registros coincidan en aquella variable por azar sería superior a la de que coincidan, sabiendo que se trata de un par coincidente y, por tanto, aquella variable no será útil en el proceso de conexión.

A partir de estas dos probabilidades calcularemos el peso (W_p) como: $\log_2(m/u)$ si las variables coinciden y como $\log_2([1–m]/[1–u])$ si no coinciden. Hay que destacar que el peso de las no coincidencias tendrá un valor negativo. Por ejemplo, veamos el cálculo del peso para una variable sexo considerando un porcentaje de error del 10% y que hay un 50% de cada sexo:

Si el sexo coincide: m = 1 – 0,1 = 0,9, u = 0,5 $W_p = \log_2(0,9/0,5) = 0,85$.
Si el sexo no coincide: 1-m = 0,1 1 – u = 0,5 5 $W_p = \log_2(0,1/0,5) = –2,3$.

Etapa de decisión. Enlazado

Una vez establecidos los bloques de comparación y decididos los pesos a aplicar para cada par de variables, se procede a realizar la conexión de los registros propiamente dicha. En ella, cada registro de la lista 1 se enlazará al registro de la lista 2 que tenga el peso más alto entre aquellos registros del bloque de comparación que no estuvieran ya enlazados o lo hubieran hecho con un peso inferior; en tal caso, se desharía el antiguo enlace para realizar el nuevo. Una vez procesados todos los registros, se ha de decidir la concordancia estableciendo los umbrales inferior y superior de valores de la variable peso (W_t), que delimitarán tres casos: los pesos por debajo el umbral inferior corresponderán a pares de registros no concordantes; los pesos por encima del umbral superior pertenecerán a pares de registros concordantes; los pesos entre estos dos valores harán referencia a casos dudosos que se tendrán que revisar manualmente o mediante otras técnicas para decidir si corresponden o no al mismo caso, probablemente valorando otras variables que ayuden a identificarlos.

Selección de los puntos de corte

Hay que tener en cuenta que cuanto más amplio sea el intervalo entre los dos umbrales más conservadora será la decisión (menos errores), pero esto llevará a que se tendrá que revisar más registros manualmente. Un análisis descriptivo de la distribución de pesos, número de coincidencias y de no coincidencias, la condición utilizada para la construcción de bloques y una estimación del tiempo empleado en la revisión manual permitirán decidir qué intervalo será el más apropiado[9]. Una simple regresión lineal puede bastar para realizar esta decisión[10]. Generalmente, si se representa el gráfico de la frecuencia de aparición de pesos, éste mostrará un perfil similar al de una distribución bimodal. Una regresión lineal doble con las variables peso y con diferentes umbrales clarificará qué intervalo es el más apropiado (fig. 1).

Validación del proceso

La metodología que se debe emplear para validar el proceso dependerá del tipo de fuente utilizada y de la factibilidad de obtener información complementaria. Un método válido sería establecer un estándar de concordancia conectando de manera determinista aquellos registros en que coincidan todas las variables utilizadas habitualmente como conectoras, juntamente con variables identificativas más específicas, como pueden ser nombre y apellidos en el caso de indivi-

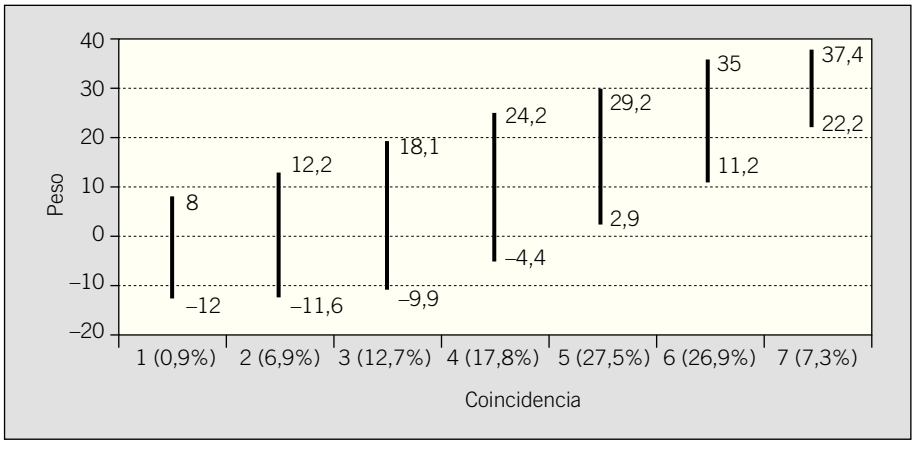

Fig. 1. Distribución del número de coincidencias según el intervalo de pesos seleccionado.

duos. Si no se dispone de esta información, una aproximación válida sería duplicar la base de datos y establecer este estándar con cada par de registros. Los registros que cumplieran con ese estándar de concordancia se pasarían al proceso de conexión probabilística. La validación consistiría en calcular la sensibilidad (que si es alta maximizará los pares concordantes conectados mediante el proceso probabilístico) y la especificidad (que si es alta evitará que los pares discordantes se conecten mediante el proceso) teniendo en cuenta el estándar de concordancia fijado. Si se dispone de una base de datos en la que se garantiza la inexistencia de casos repetidos, puede comprobarse la especificidad partiendo la base de datos en dos: se crean así dos conjuntos disjuntos que pasarían al proceso de conexión probabilística.

Aproximación informática

Si bien existen diferentes aplicaciones informáticas que permiten realizar la mayor parte de este proceso[11], disponer de una aplicación que permita realizarlo todo sin intervención del observador resulta difícil debido a que hay que tomar algunas decisiones antes de iniciar el proceso:

– Establecer qué variables conectoras se van a utilizar en el proceso de conexión.
– Seleccionar la condición para la realización de bloques.
– Definir los criterios de revisión de casos dudosos.

En todo caso, la aplicación debe disponer de una interfaz que permita establecer todos estos parámetros previos y que realice todos los pasos del proceso de enlace de manera automática. Un algoritmo compatible con este proceso podría ser el que se presenta en la tabla 2.
Para el proceso de revisión de los pares dudosos, puede utilizarse la técnica de revisión dinámica propia de los planes de revisión secuenciales que se utiliza en los procesos de control de calidad, para no tener que revisar todos los pares. Este método permitiría ir ajustando los umbrales con los pares revisados. Los planes de revisión secuenciales calculan unas bandas de revisión. El sistema avisaría cuando la relación entre el número de pares no coincidentes y el de pares revisados dejase de estar dentro de estas bandas. Si se sale del límite superior de la banda de revisión, hay un número demasiado grande de pares no coincidentes y, por tanto, el sistema ha de proponer subir el umbral. Si se sale de la banda por la parte inferior es porque se encuentra un número demasiado grande de coincidentes y, por tanto, el sistema ha de proponer bajar el umbral (fig. 2).

Algunas aplicaciones

Desde hace años los servicios de la Agència de Salut Pública de Barcelona (anteriormente Institut Municipal de Salut Pública) han venido utilizando estas técnicas para mejorar la calidad de la información sanitaria que se gestiona y servir mejor a la salud pública de la población. Sus primeras aplicaciones se relacionaron con el conocimiento de las dimensiones y características de los usuarios de drogas por vía parenteral, una población difícil de acotar a la que hemos de proporcionar servicios apropiados[12]. A continuación se presentan 4 ejemplos de conexión probabilística de ficheros derivados de nuestra experiencia. Nuestros servicios gestionan diversas bases de datos necesarias para mejorar la salud pública, registradas con la Agencia de Protección de Datos y que tienen acceso a otras fuentes de información relevantes para sus fines.

TABLA 2

Algoritmo del proceso de conexión probabilística

1. Selección y homologación de los formatos de las variables identificadoras
2. Cálculo de frecuencias y pesos para cada una de estas variables
3. Selección de las variables/condiciones para el establecimiento de bloques
4. Revisión de todos los registros de la primera fuente y comparación con los registros que cumplan la condición del bloque, asignando el mejor par posible según el peso máximo obtenido
5. Análisis descriptivo de los pesos aplicados en el proceso de enlace. Establecimiento de los umbrales inferior y superior que delimiten los registros a revisar
6. Incorporación de información complementaria para la revisión de los pares dudosos
7. Revisión de los pares dudosos
8. Extracción de resultados en una base de datos unificada

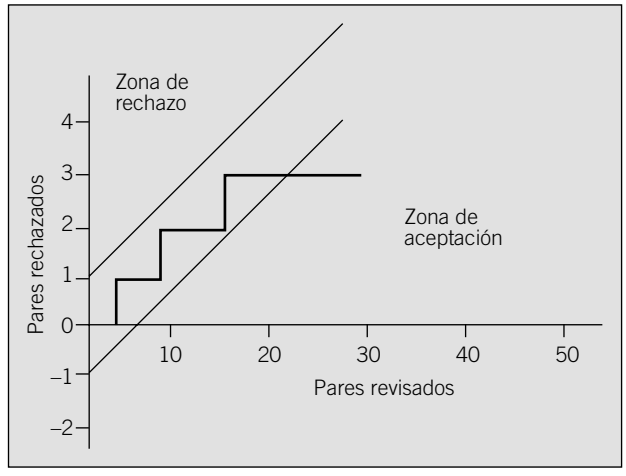

Fig. 2. Representación gráfica del proceso de revisión dinámica de pares dudosos.

Estudio de las lesiones por accidentes de tráfico

Este proyecto permitió mejorar la información relativa a las lesiones por accidentes de tráfico, combinando las variables relacionadas con el accidente (que recoge la Guardia Urbana) con las relativas a las lesiones (que recogen los hospitales)[13].
– Las bases de datos que intervinieron en este estudio son: los accidentes de tráfico en la ciudad de Barcelona recogidos por la Guardia Urbana en 1997 (12.481 registros) y los registros de urgencias por accidentes de tráfico de 7 centros hospitalarios de la ciudad participantes en el proyecto DUHAT (16.733 registros).
– Las variables conectoras utilizadas fueron: código postal, edad, tipo de vehículo, hospital, posición del individuo, sexo y fecha de accidente.
– La condición que se utilizó para crear los bloques fue: fecha de atención en urgencias mayor que fecha de accidente pero inferior a tres días desde esa fecha.
– Pesos encontrados: mínimo, –12,0; máximo, 37,4.
– Resultados: correctos, 55%; incorrectos, 28,2%; dudosos, 16,8%; con umbrales de revisión mínimo de –8 y máximo de 12.

Estudio de la mortalidad de los enfermos de sida

Este proyecto tuvo como antecedente la conexión manual entre los ficheros de mortalidad y de tuberculosis utilizada desde 1986 para mejorar la calidad de los datos de ambos registros e identificar posibles fallos del tratamiento farmacológico[14]. La

conexión manual era factible con la tuberculosis por el escaso número de defunciones atribuibles a esta enfermedad, pero imposible con el sida por los efectivos implicados, por lo que se optó por la conexión probabilística en este proyecto[15].
– Las bases de datos que intervinieron en el estudio de mortalidad de los enfermos de sida son: Registro de casos de sida de Barcelona 1991-2000 (2.991 registros) y defunciones por sida de residentes en la ciudad en el Registro de mortalidad de Barcelona 1991-2000 (2.413 registros).
– Las variables conectoras utilizadas fueron: sexo, nombre, apellidos, fecha de nacimiento y fecha de defunción.
– La condición utilizada para crear los bloques fue: mes + año de defunción coincidentes.
– Pesos encontrados: mínimo, –99,0; máximo, 70,21.
– Resultados: correctos, 77,1%; incorrectos, 13,6%; dudosos, 9,3%; con umbrales de revisión mínimo de –16 y máximo de 32.

Enfermedades infecciosas en usuarios de drogas no institucionalizadas

Como parte de la evaluación de las políticas e intervenciones preventivas en los usuarios de drogas se analizó la incidencia de diversas enfermedades infecciosas que se desea evitar en estas personas.
– Las bases de datos que intervinieron en este estudio son: Registro de enfermedades de declaración obligatoria de Barcelona 1987-2002 (30.598 registros) y Sistema de información de drogas de Barcelona (SIDB) 1992-2000 (35.550 registros).
– Las variables conectoras utilizadas fueron: sexo, nombre, apellidos y fecha de nacimiento.
– La condición utilizada para crear los bloques fue: iniciales de apellidos + mes y año de nacimiento coincidentes.
– Pesos encontrados: mínimo –9,45, máximo 47,9.
– Resultados: correctos, 10,6%; incorrectos, 88,29%; dudosos, 1,11%; con umbrales de revisión mínimo de 13 y máximo de 20.

Defectos congénitos en hijos de mujeres usuarias de drogas

Este proyecto se realizó con el objeto de validar la calidad de la declaración de los datos de consumo de drogas en la encuesta del Registro de defectos congénitos de Barcelona[16].
– Las bases de datos que intervinieron en este estudio son: Registro de defectos congénitos de Barcelona 1991-2000 (4.794 registros) y Sistema de información de drogas de Barcelona (SIDB) 1992-2000 (35.550 registros).
– Las variables conectoras utilizadas fueron: nombre, apellidos y fecha de nacimiento.
– La condición utilizada para crear los bloques fue: iniciales de apellidos + mes y año de nacimiento coincidentes.
– Pesos encontrados: mínimo, –17,79; máximo, 46,94.
– Resultados: correctos, 0,65%; incorrectos, 93,6%; dudosos, 5,75%; con umbrales de revisión mínimo de –5 y máximo de 15.

Otras consideraciones

Según nuestra experiencia, las ventajas del método de enlace probabilístico se manifiestan sobre todo en situaciones en las que los datos a enlazar presentan un cierto grado de borrosidad. En otras situaciones, los métodos estándar son más fácilmente aplicables. Hay que tener en cuenta que el método requiere un análisis previo relativamente exhaustivo de los datos, para verificar la viabilidad y fiabilidad de su aplicación, y suele precisar el apoyo de un profesional versado en el manejo de grandes bases de datos con instrumentos estadísticos e informáticos. En contrapartida, el método es aplicable a un amplia gama de situaciones y permite abordar análisis exploratorios innovadores. La creciente presencia en España de nuevos residentes cuyos nombres son de fonética y grafía ajenas a nuestra tradición obligará a usar técnicas de este tipo para evitar pérdidas de información en los archivos de datos.
En cualquier caso, quienes aborden el enlace probabilístico (como cualquier otro tipo de proceso que pueda relacionar fuentes de información individual recogidas con fines diferentes), deben tener en cuenta la confidencialidad y otras implicaciones asociadas a la Ley de Protección de Datos vigente en España.

Agradecimiento

Los autores agradecen las sugerencias y críticas aportadas por el Dr. J.R. Villalbí.

REFERENCIAS BIBLIOGRÁFICAS

1. Jaro A. Probabilistic linkage of large public health data files. Stat Med 1995;14:491-8.
2. Churches T. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. BMC Med Res Methodol 2003;3:1.
3. Neutel CI, Johansen HL, Walop W. New data from old: epidemiology and record-linkage. Progress Food Nutrition Sciences 1991;15:85-116.
4. Gill L. Methods for automatic record matching and linking and their use in National Statistics. National Statistics Methodological Series No. 25. London: National Statistics, 2001.
5. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Abbatt JD. Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. Comput Biol Med 1983; 13:157-69.
6. Ortí RM, Macfarlane D, Domingo A. Obtención de una cohorte de adictos a opiáceos a partir de la conexión de registros confidenciales. Gac Sanit 1994;8:229-38.
7. Newcombe HB. Handbook of record linkage: methods for health an statistical studies. Administration and business. Oxford: Oxford University Press, 1988.
8. National Association of Governors highway safety representatives So you want to link your data. National Highway Traffic Safety Administration, 1996.
9. National Center for Statistics and Analysis. Crash Outcome Data Evaluation System (CODES) Technical Report. National Highway Traffic Safety Administration, 1995.
10. Norusis MJ. SPSS for Windows Base system user's guide. Release 6.0 SPSS Inc. 1993.
11. MatchWare Technologies: AutoStan and AutoMatch User's Manuals. Kennebunk, Maine 1998. Disponible en: http://www.ascentialsoftware.com
12. Brugal MT, Domingo-Salvany A, Maguire A, Caylà JA, Villalbí JR, Hartnoll R. A small area analysis estimating the prevalence of addiction to opioids in Barcelona, 1993. J Epidemiol Community Health 1999; 53:488-94.
13. Cirera E, Plasencia A, Ferrando J, Arribas P. Probabilistic linkage of police and emergency department sources of information on motor-vehicle injury cases: a proposal for improvement. J Crash Prevention Injury Control 2001;2:229-37.
14. Villalbí JR, Galdós-Tangüis H, Caylà JA. El control de la tuberculosis basado en la evidencia: una aproximación desde la salud pública. Med Clin (Barc) 1999;112:111-6.
15. Rodríguez-Sanz M, Borrell C, García de Olalla P, Pasarín MI, Brugal MT, Caylà JA. Evolución de la mortalidad por sida ¿es diferencial según nivel socioeconómico? Gac Sanit 2002;16(Supl 1):67-71.
16. Salvador J, García-Miñaur S, Caballín MR, Mosquera C, Baena N, García E, et al. Registros poblacionales de defectos congénitos en España. An Esp Pediatr 1998;48:575-82.