OBTENCIÓN DE UNA COHORTE DE ADICTOS A OPIÁCEOS A PARTIR DE LA CONEXIÓN DE REGISTROS CONFIDENCIALES

Rafael M. Ortí Lucas / Dave Macfarlane / Antònia Domingo Salvany
Departament d'Epidemiologia i Salut Pública. Institut Municipal d'Investigacio Médica (IMIM).
Universitat Autònoma de Barcelona (UAB)

Resumen

La necesidad de unir varios archivos para crear una cohorte de adictos a los opiáceos, de tamaño suficiente para el análisis de su mortalidad, requiere la utilización del método probabilístico de conexión de registros. El presente estudio es una primera fase en la que se conectan dos subarchivos del Registro de urgencias toxicológicas del Hospital del Mar de Barcelona. Esta fase sirvió para adaptar la conexión probabilística a nuestros archivos, elaborar programas informáticos, definir criterios de concordancia, y evaluar la validez y rendimiento del método. Para salvaguardar la confidencialidad se limitaron las variables de identificación al sexo, fecha de nacimiento y tres iniciales de cada apellido. La conexión automatizada probabilística resultó factible, válida y eficiente; observándose que, a diferencia de las aproximaciones determinísticas, con una revisión visual inferior al 5% de los registros, se obtenían sensibilidades y especificidades superiores al 95%. Se discuten las dificultades planteadas a lo largo del proceso.

Palabras clave: Conexión de registros médicos. Conexión probabilística. Seguimiento automático. Estudios retrospectivos. Sistemas de información. Confidencialidad.

FORMATION OF A COHORT OF OPIATE ADDICTS THROUGH LINKAGE OF CONFIDENTIAL RECORDS

Summary

The need to combine several files in order to create a cohort of opiate addicts, sufficiently large for the analysis of its mortality, requires use of the probabilistic method of record linkage. This study is a preliminary phase in which two sub-files of the Hospital del Mar (Barcelona) Register of toxicological emergencies are linked. This phase served to adapt probabilistic record linkage to our files, develop computer programs, define agreement criteria, and evaluate the validity and performance of the method. In order to safeguard confidentiality, identification variables were limited to sex, birth date and three initial letters fron each surname. The automated probabilistc linkage was seen to be feasible, valid and efficient; in contrast to deterministic approaches, sensitivities and specificities above 95% were obtained with visual reviewing of under 5% of the records. Difficulties encountered during the process are discussed.

Keywords: Medical record linkage. Probabilistic linkage. Automated follow-up. Retroospective studies. Information systems. Confidentiality.

Introducción

La conexión de archivos médicos *(record linkage)* es de gran interés para la epidemiología, como sugiere su creciente utilización en estudios de la morbilidad y mortalidad de poblaciones de difícil seguimiento prospectivo[1-3] y de la asociación entre enfermedades y riesgos como los laborales[4,5] o efectos secundarios de algunos fármacos[6,7] que por actuar con bajo impacto o sobre grupos de riesgo especiales requieren tamaños muestrales muy grandes. El *record linkage* también ha sido utilizado para mejorar la calidad y exhaustividad de registros médicos y administrativos[8,9] y para facilitar la planificación y evaluación de los servicios sanitarios[10].

El proceso de conexión (incorporación a un registro individual de datos procedentes de diferentes

Correspondencia: R. Ortí. (IMIM). C/ Doctor Aiguader, 80. 08003 Barcelona.
Este artículo fue *recibido* el 5 de abril de 1993 y fue *aceptado* tras revisión el 13 de junio de 1994.

fuentes) es especialmente útil para complementar la información de registros de enfermos crónicos con datos administrativos o demográficos, en especial sobre mortalidad; o para aumentar el tamaño de un archivo de modo que aumente la precisión de ciertos análisis.

La revisión visual directa es el método de conexión de archivos más sencillo y válido cuando hay pocos registros y se dispone de un mínimo de información. Sin embargo, al intentar la conexión de archivos de mayor volumen aumenta el número de errores y disminuye su rendimiento[11].

El método determinístico, que establece la concordancia en base a la coincidencia, más o menos estricta, de las variables de identificación, es tal vez el más frecuentemente usado; pero tampoco es suficientemente válido cuando se pretende una gran sensibilidad. En este sentido, un estudio previo para la conexión de dos registros de urgencias toxicológicas subestimaba el 20% de los pares concordantes, obligando a una costosa revisión visual[12].

Por último, el método probabilístico, propuesto por Newcombe[13] y desarrollado posteriormente por el mismo Canadá[14], y por Achesson, Baldwin y Graham en Gran Bretaña[15], ha sido utilizado en diversos estudios, tanto en el ámbito sanitario como en el administrativo. Este proceso de conexión de registros reduce el trabajo de revisión visual a la vez que simula el comportamiento de la mente humana en la elección de la concordancia. Grandes sistemas que aplican el método probabilístico son el Laredo Epidemiology Project[16] que reconstruye genealogías a partir de 350.000 individuos, localizados en los registros civil y eclesiástico, con el fin de estudiar la herencia de enfermedades crónicas como el cáncer; el California Automated Mortality Linkage System[17] (CAMLIS) establecido en 1981 para facilitar el seguimiento de diversas cohortes en las que se pretendía determinar la mortalidad; y el Oxford Record Linkage Study que conecta los ficheros de pacientes hospitalizados con los registros de nacimientos y defunciones del área de Oxford[18].

En España, el año 1984 se realizó un seminario científico sobre el estado de las estadísticas vitales, en el que se revisó la aplicabilidad del *record linkage*. Es una ponencia[19], se citaba que "la posibilidad de realizar conexiones entre registros en nuestro país puede parecer utópica si tomamos en consideración el precario estado de nuestros registros demográficos". Nueve años después la experiencia en la informatización de los archivos médicos ha agilizado la elaboración de la información registrada. A pesar de ello, la escasa documentación publicada sugiere que la conexión de archivos automatizada y con criterios probabilísticos continúa siendo una aplicación infrautilizada en nuestro ámbito[20]. El desconocimiento de la metodología y la necesidad de desarrollar el "software" adecuado serían los principales frenos a su utilización. Sin embargo, el aprovechamiento de sus ideas genéricas, ligeramente modificadas, podría facilitar su adaptación en beneficio de diversos estudios epidemiológicos.

En este marco, *el objetivo del presente trabajo es la preparación de un sistema de conexión de registros que permita evitar la duplicidad de sujetos en una cohorte construida a partir de la conexión de tres archivos de adictos a opiáceos (Registro de episodios urgentes del Hospital del Mar (RHM). Registro de urgencias de otros hospitales de Barcelona y Registro de la red de atención al drogodependiente de la Generalitat de Catalunya).* La elección del método probabilístico viene condicionada por la necesidad de unir varios archivos y por la intención de realizar al seguimiento prospectivo histórico de una población de difícil identificación. Los objetivos específicos son: 1) adaptar el método de *record linkage* automatizado probabilístico a las características de nuestros archivos (grado de cumplimentación, existencia de un segundo apellido, etc.), 2) validar la metodología utilizada, y 3) evaluar los problemas inherentes al proceso (índices de eficacia y rendimiento) mediante 4) la conexión de los archivos de adictos a opiáceos atendidos en el servicio de urgencias del Hospital del Mar durante los períodos 1983-89 y 1990-91.

Sujetos y métodos

El estudio se divide en dos fases: *la elaboración del sistema de conexión con la creación de programas informáticos específicos que permita aplicar el método a nuestros datos y la validación del proceso mediante la conexión de dos archivos de adictos a opiáceos.* En la primera se utilizaron muestras de pares de registros correspondientes a adictos registrados durante 1989, año en el que se recogió información duplicada de las mismas urgencias del Hospital del Mar en dos bases de datos distintas[21,22]. En concreto, tras excluir los registros con información incompleta (ambos apellidos o día y mes desconocidos) se dispuso de una muestra con 1.379 pares de registros concordantes, cuya concordancia se definió a partir del número de registro asignado a cada episodio de urgencias. La combinación aleatoria de los pares concordantes permitió obtener 1.379 pares discordantes.

Para la fase de validación se conectaron los subarchivos del Registro de Urgencias Toxicológicas del Hospital del Mar, diferenciados por el cambio del protocolo de recogida de datos realizado en 1989. El primero constaba de 15.966 episodios, que revisados para eliminar posibles duplicados (episodios pertenecientes a una misma persona) y tras la exclusión de los registros de individuos con variables de identificación incompletas correspondían a 4.882 individuos; y el segundo, con datos de 1990 y 1991, incluía 5.202 episodios pertenecientes a 2.408 individuos.

Diseño y aplicación del método

El desarrollo y aplicación del método presentaba cuatro puntos decisivos[23]: 1) la selección de las variables que servirían para realizar la identificación; 2) la determinación del sistema más adecuado para reducir el número de comparaciones entre registros a las que se aplicarían los pasos siguientes; 3) el cálculo de los pesos específicos que, considerando el nivel de acuerdo de las variables identificadoras (concordancia parcial o total entre los dígitos que componen cada variable), reflejarían la probabilidad de que dos registros perteneciesen al mismo individuo; y 4) la definición de los criterios de concordancia (determinación de los umbrales).

Como *variables identificadoras* se usaron el sexo, la fecha de nacimiento y las tres letras iniciales de cada apellido que eran, entre las comunes a los archivos que queríamos conectar, las que permitían identificar los registros individuales de cada archivo.

La *reducción de las combinaciones de registros* a las que se aplicarían los pesos era un proceso más difícil. Al no disponer de códigos fonéticos, tipo Soundex code[24], adaptados a los apellidos locales, se crearon bloques de comparación mediante un sistema de códigos fonético-gráficos que tenía en cuenta los errores lingüísticos de los archivos (registros mal codificados por su ortografía o pronunciación). Previamente a la formación de bloques, se establecían equivalencias entre las iniciales o sílabas con variantes erróneas más habituales, así por ejemplo, ASQ y AZQ, o HE y ME, pasarían a formar parte del mismo bloque aunque no correspondieran al mismo fonema). En concreto, tras agrupar las iniciales de los apellidos, se seleccionaban las cuatro primeras consonantes y se les asignaba el código correspondiente, de manera similar a la técnica del Soundez code[14]. De este modo sólo se comparaban los registros con ciertas similitudes. Si después de calcular los pesos quedaba más de un registro con pesos superiores al umbral de discriminación se elegía el de mayor puntuación.

La etapa principal, sin embargo, era el *cálculo de los pesos*[13,14]. El primer paso para su obtención consistía en el cálculo de las frecuencias (probabilidades) de cada nivel de acuerdo de las variables de identificación en las muestras de pares concordantes y discordantes. Las razones entre estas frecuencias (RF) debían presentar diferencias suficientes entre los diversos niveles de acuerdo (cambio próximo al doble al pasar de un nivel al siguiente) para establecer un gradiente de discriminación significativo, lo que llevó a agrupar algunas categorías. Hecho esto, la *puntuación total* de un par de registros resultaba del producto de las razones de frecuencias correspondientes a sus variables de identificación. Siguiendo a Newcombe, para facilitar los cálculos de obtuvo un *peso global* con la suma de los logaritmos de base dos de las RF.

Los pesos globales son iguales para cualquier par de registros con niveles de acuerdo idénticos. Sin embargo, la probabilidad de que dos pares de registros concuerden también depende de la frecuencia de cada categoría de las variables en los archivos que intentamos conectar, por ejemplo, si el apellido coincidente es raro aumenta la probabilidad de que los registros pertenezcan a la misma persona. Por ello, para aumentar el poder discriminador, se ajustaron los pesos globales según la frecuencia específica de las categorías concordantes. Este refinamiento del proceso se basa en la obtención de *factores de ajuste* (cociente entre la frecuencia general y las frecuencias específicas. Las *frecuencias específicas* (FE) o proporción de casos de cada categoría, pueden calcularse a partir de cualquiera de los archivos que se quieren conectar, generalmente el más voluminoso, o a partir de la unión de ambos. Las *frecuencias generales* (FG) resultan del sumatorio de los productos de las frecuencias específicas de cada categoría en un archivo por las frecuencias existentes en el otro (Tabla 4). Cuando un archivo es mucho mayor, las frecuencias generales pueden aproximarse por el sumatorio de las respectivas frecuencias específicas elevadas al cuadrado) que multiplican las razones de frecuencias para obtener un valor específico. La suma de los logaritmos de base dos de estos valores es el *peso ajustado o específico* de cada par de registros (Anexo).

Las distribuciones de frecuencias de los pares de registros concordantes y discordantes según sus pesos específicos permitían definir unos *umbrales que discriminarían la concordancia*. Estos límites, definidos inicialmente por los puntos donde se solapan ambas distribuciones, diferenciaban entre

Tabla 1. Razones de frecuencias para la variable sexo (n=2329)

Nivel de acuerdo	Concordantes/ Discordantes	Razón de frecuencias
A	2.291/1.412	1,623
M	4/4	1,000
D	34/913	0,037

A= Acuerdo en el sexo. D= Desacuerdo en el sexo. M= En un registro del par se desconoce el sexo.

pares concordantes (en principio los que tienen pesos mayores que el valor umbral superior) y discordantes (con pesos menores que el valor umbral inferior). Para decidir la concordancia de los pares con pesos intermedios se necesitaba más información.

El sistema de codificación fonético-gráfica y los pesos utilizados son aplicables para la conexión de registros con características similares. Los programas informáticos se elaboraron con el paquete SPSS-X[25].

Validación del proceso

Tras obtener los pesos específicos y definir los umbrales de discriminación se realizó y evaluó la conexión. Para ello, mediante la unión determinística de los *subarchivos del registro de urgencias toxicológicas del Hospital del Mar* y la revisión de los demás registros con información suplementaria (conexión determinístico-visual), se elaboró un estándar de concordancia *compuesto por 950 pares de registros concordantes (896 si se eliminaban variables con datos ausentes)* que serviríande patrón de referencia. Después se aplicaron los pesos específicos calculados y se completó la conexión probabilística de los subarchivos. La validez de la conexión se valoró comparando sus resultados con el estándar de concordancia. En concreto, se calcularon los índices de eficacia: sensibilidad (probabilidad de que un par concordante sea clasificado por el programa de conexión de registros como concordante), especificidad (probabilidad de que un par de registros pertenecientes a diferentes individuos aparezca como discordantes) y valores predictivos positivo y negativo. También se contabilizó el número de pares de registros que requerirían una revisión visual.

Resultados

Obtención de pesos específicos

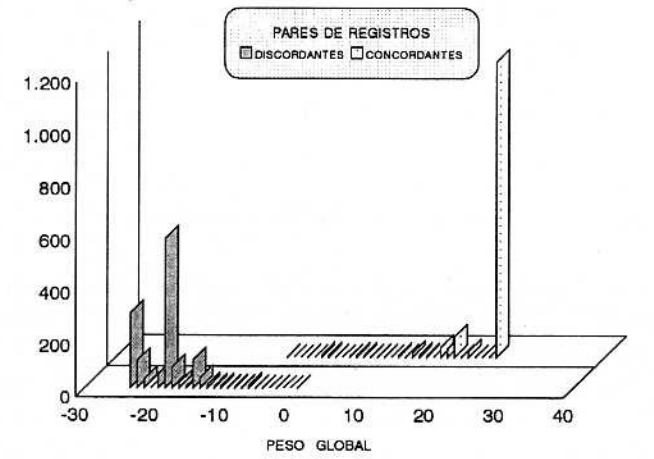
Para el cálculo de las razones de frecuencias (Tablas 1 y 2) se utilizaron las muestras de 1.379

Tabla 2. Razones de frecuencias para las iniciales de los apellidos y la fecha de nacimiento (n=2329)

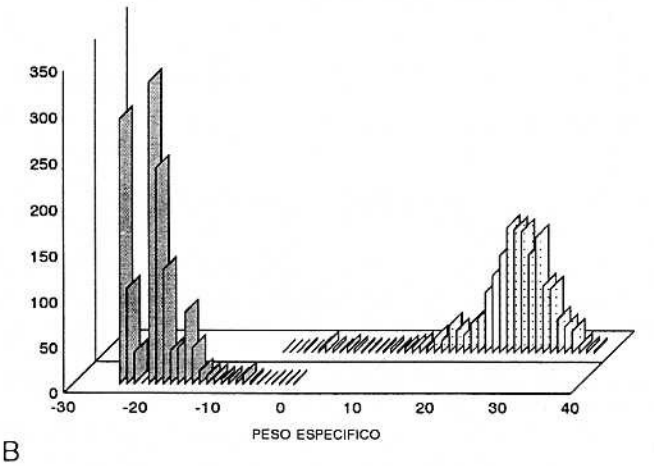
	Primer apellido		Segundo apellido		Fecha de nacimiento	
Nivel de acuerdo†	C/D	R.F.	C/D	R.F.	C/D	R.F.
AAA	2203/20	110,150	2172/16	135,750	2140/1	2140,000
DAA	27/59	0,458	20/56	0,357	44/6	7,333
ADA	7/16	0,437	7/11	0,636	26/4	6,500
AAD	48/28	1,715	44/35	1,257	35/10	3,500
ADD	6/111	0,054	6/94	0,064	11/91	0,121
DDA	6/165	0,036	5/165	0,030	3/68	0,044
DAD	6/335	0,018	7/279	0,025	3/158	0,019
DDD	25/1593	0,016	24/1528	0,016	7/1620	0,004
AAM			1/0	1,5*		
AMM	0/1	0,06*			47/51	0,922
MMM			43/143	0,301	8/12	0,667
DAM					0/1	0,02*
ADM			0/1	0,07*	0/1	0,15*
DMM					5/300	0,017
DDM	1/1	1,000	0/1	0,02*	0/5	0,005*
MDD					0/1	0,005*

† Los códigos de los diferentes niveles de acuerdo se refieren a la primera, segunda y tercera inicial en el caso de los apellidos; y al año, mes y día en el caso de la fecha de nacimiento.
R.F.= Razones de frecuencias. C/D = Número de pares de registros concordantes y discordantes.
A= Acuerdo. D= Desacuerdo. M= Desconocido.
* Pesos artificiales. Aplicados para los niveles de acuerdo sólo representados en una de las muestras de pares.

Figura 1. Distribución de pares de registros de las muestras según sus pesos. A. Pesos globales. B. Pesos específicos (ajustados por sexo, año de nacimiento e iniciales de los apellidos)



A



B

pares de registros concordantes y discordantes ampliadas, tras comprobar la ausencia de grandes variaciones en las frecuencias de las categorías con mayor número de casos, con los 950 pares concordantes obtenidos en la fase de validación y otros 950 pares discordantes procedentes de los mismos archivos. De este modo pudieron calcularse RF para algunos niveles de acuerdo poco representados, aunque todavía tuvieron que asignarse ocho pesos artificiales por la ausencia de casos en alguna de las muestras. La distribución de los *pesos globales* correspondientes a todas las posibles combinaciones de categorías de las variables oscilaban entre -5 y 25 para los casos concordantes y entre -24 y 0,6 para los discordantes (Fig. 1a). Los puntos de confluencia correspondían al par concordante con menor peso (-5) que presentaba fecha de nacimiento y segundo apellido iguales pero sexo y tercera inicial del primer apellido diferentes, y al par discordante con mayor peso (0,6) que correspondía a individuos con igual sexo y fecha de nacimiento pero con diferentes apellidos.

Después se calcularon las frecuencias específicas en los dos archivos, y tras confirmar la ausencia de grandes diferencias entre ellos, se consideraron representativas del total las del RHM, excepto para el sexo, caso en que se prefirieron las frecuencias del archivo codificado según el nuevo cuestionario en el que no aparecían datos incompletos (Tabla 3). Se obtuvieron factores de ajuste para el sexo, iniciales de los apellidos, y año de nacimiento (la distribución de meses y días se asumió aleatoria). Los varones eran el 73% del total por lo que el peso ajustado era mayor que el peso global cuando el sexo concordante era el femenino y menor cuando era el masculino. Los factores de ajuste para los apellidos eran más complicados dado el gran número de categorías resultantes de la combinación de tres letras. En el caso del año de nacimiento, los factores eran inferiores a la unidad cuando la proporción de nacimientos superaba el 5,5% del total (años 1961-1968).

Elección de niveles umbrales de rechazo o aceptación

El ajuste permitía discriminar mejor entre los pares. La distribución de frecuencias de los pesos específicos (Fig. 1b) aumentó el espectro de valores, fundamentalmente a expensas de los pares concordantes. Los valores extremos oscilaban ahora entre -23,50 (pares discordantes) y + 36,50 para la máxima concordancia. Además el ajuste reducía el número de pares incluidos en la categoría de dudosos, pasando a esta categoría algunos registros en que las fechas de nacimiento (1961-1968) y las iniciales de los apellidos (GAR, GON, MAR...) eran las más frecuentes en los archivos. Los límites del rango de solapamiento entre las distribuciones deberían ser los umbrales más adecuados para discriminar la concordancia. Sin embargo, no contemplaban la posibilidad de que, por azar, apareciesen registros discordantes con idénticas iniciales en alguno de los apellidos, caso en que los pesos llegarían a 13. Conservadoramente se ampliaron los umbrales a [-3/13], de modo que serían concordantes los pares con peso superior a 13, discordantes los de peso inferior a -3 y de concordancia dudosa el resto.

Tabla 3. Cálculo de factores de ajuste para la variable sexo

Archivo	Sexo	N	Frecuencia específica	Frec. específica al cuadrado	Frecuencia general	Factor de ajuste
ROH	Varón	1010	0,732	0,5358	0,6071	0,8294
	Mujer	369	0,267	0,0713	0,6071	2,2738
RHM	Varón	1000	0,724	0,5242	0,5982	0,8262
	Mujer	375	0,272	0,0740	0,5982	2,2020
	Desconocido	4	0,003	0,00001	0,5982	199,4000

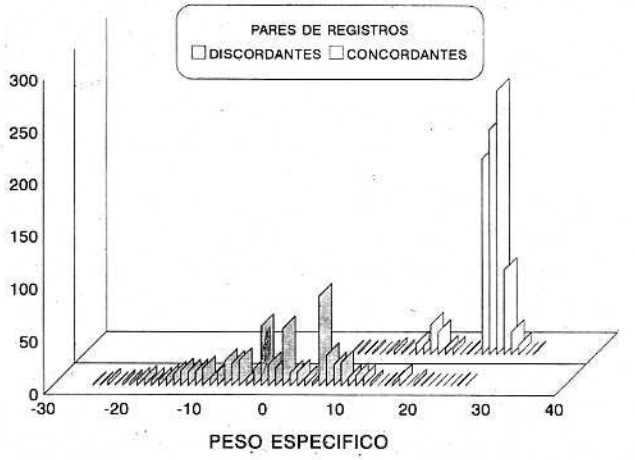
ROH. Datos sobre adictos ingresados en 1989 en el Hospital del Mar, que fueron obtenidos con el nuevo cuestionario.
RHM. Datos sobre adictos ingresados en 1989, obtenidos a partir del registro de urgencias toxicológicas del Hospital del Mar.

Conexión de los subarchivos del RHM (validación del proceso)

La comparación de los subarchivos del Hospital del Mar daría lugar a casi 12 millones (4.882 x 2.408) de pares de registros. Aunque la formación de bloques redujo esta cifra a 9.503, aún quedaban algunos registros con correspondencias múltiples. Por azar aparecían pares discordantes con pesos elevados que les hacían confundirse con pares concordantes. La selección del peso mayor, la mejor comparación para cada registro, redujo el número de pares de registros a 1.467.

La distribución de los pesos específicos correspondientes se resume en la figura 2. Con el umbral propuesto y, asumiendo la validez del estándar de concordancia *(896 pares concordantes)* se obtenían una sensibilidad y especificidad bastante buenas (Tabla 4); sin embargo, la aceptación de estos puntos de corte obligaba a revisar cerca del 27% de los registros. De este modo, considerando los criterios de validez y el coste, resultaba más eficiente el umbral [8/14]. Con este umbral se estableció directamente la concordancia de 844 pares de registros a los que se añadieron 10 más después de una revisión visual (4,2% de los pares de registros) en la que se disponía, en algunos casos, del nombre de pila. Así, además, al comparar los pares de registros conectados con los del estándar de concordancia, se observaba que los falsos positivos apenas presentaban mínimas modificaciones: una "N" por una "M" o una "H" por una "M" (seis casos), el sexo (tres casos), el año (seis casos), el día (tres casos), o el mes de nacimiento (dos casos); mientras que sólo aparecían cuatro falsos negativos, con pesos inferiores a 8, a los que habría que añadir cinco pares de registros más perdidos por la existencia de pares discordantes con peso superior, y 33 que no pudieron conectarse porque habían quedado ubicados en bloques de comparación diferentes. La sensibilidad y la especificidad resultantes correspondían a un 95,5% y a un 96,7% respectivamente.

Figura 2. Distribución de pares de registros según pesos específicos. Conexión de los subarchivos correspondientes a los períodos 1983-89 y 1990-91.



Incluye los mayores pesos específicos calculados para cada registro cuando se combina con el otro archivo.

Discusión

Validez del método

Las críticas sobre la validez epidemiológica de los estudios basados en datos obtenidos mediante la conexión de archivos médicos[26-28] no invalidan el método cuando la disponibilidad de información adecuada, relación coste-efectividad, etc. lo indican[29]; resultando incluso, en algunas ocasiones, más adecuado que el mismo seguimiento directo[30]. Los estudios de la población drogodependiente serían uno de los candidatos a beneficiarse de este método, especialmente cuando se pretende el seguimiento de un elevado número de individuos[31,32]. Sin embargo, conviene tomar las precauciones propias de los estudios basados en el uso de grandes bases de datos[33], y tener en cuenta la definición de la pobla-

Tabla 4. Tabla de validación para diferentes niveles umbrales

Valores umbrales	Índices de eficacia				Registros a revisar		
	S*	E	VP+	VP-	C (%)	D (%)	T (%)
Previsto [-3 / 13]	100	96,06	97,28	100	1,3	62,1	26,5
Sensible [7 / 14]	99,88	96,72	97,72	99,83	1,5	16,4	7,7
Intuitivo [8 / 14]	99,53	96,72	97,71	99,17	1,2	8,4	4,2
Específico [9 / 18]	99,53	98,85	99,19	99,34	6,5	10,5	8,2
Seguro [7 / 20]	99,88	99,51	99,65	99,84	19,2	10,0	13,8

* La sensibilidad debe reducirse en un 4 por cien por la pérdida de registros en el bloqueo.
S= Sensibilidad. E= Especificidad. VP +/- = Valores predictivos positivo y negativo.

ción estudiada, el nivel de análisis requerido, la calidad de los datos y la disponibilidad de un método adecuado para su conexión que permita complementar la información[34].

La finalidad del proceso de conexión es decidir qué pares de registros formados a partir de diferentes archivos corresponden a una misma persona. La ausencia de un número de identificación único para todos los archivos, considerado el mejor método[8], obliga a seleccionar entre las variables disponibles. En este sentido, la invariabilidad de los apellidos según el estado civil[8] y la existencia de un segundo apellido son características de los nombres españoles que aumentan su poder de discriminación.

Los resultados sugieren que el método probabilístico no sólo es factible para la conexión de registros de tamaño moderado sino que además resulta válido, fiable y eficiente (ahorra trabajo de revisión). La sensibilidad y el valor predictivo positivo, similares a los obtenidos en otros estudios[35], superaban los resultados de las técnicas determinísticas estrictas que, con nuestros datos, sólo hubiesen captado cerca del 86% de los pares concordantes. La especificidad y los valores predictivos negativos también eran adecuados. Sin embargo, su interés es menor dado que dependen del número de pares discordantes considerados y, a su vez, del método utilizado para reducir el número de comparaciones entre los registros, por lo que habitualmente no son calculados[33].

Una limitación de este estudio sería la definición de la concordancia del estándar a partir de una revisión visual y no a partir del número de episodio u otra técnica que ofreciese mayor validez. Así, las inexactitudes del estándar podrían modificar las distribuciones de los pares e impedir la elección de los umbrales adecuados, de modo que los resultados quedarían sesgados. No obstante, a pesar de la variación de los umbrales, los pesos seguían especificando los mismos niveles de acuerdo. En este sentido, más que mantener *a priori* un umbral constante, lo interesante es entender su significado. Además, la posibilidad de obtener una gama de puntos de corte que permita adaptarse a los objetivos de cada estudio sería una ventaja adicional. Por ejemplo, para formar una cohorte de individuos en la que destaque la ausencia de registros duplicados interesará una alta sensibilidad; mientras que si lo importante es no perder ninguna persona, aunque aumente artificialmente el tamaño de la cohorte, será mejor primar la especificidad.

Dificultades en la aplicación del método

Asumiendo la validez epidemiológica y metodológica, hay que tener en cuenta las dificultades inherentes al desarrollo del proceso. Abordando los problemas relacionados con su factibilidad técnica, destaca la *necesidad de una programación informática previa*, a veces demasiado elaborada para los objetivos de la conexión, lo que ha llevado a plantear métodos determinísticos menos estrictos, que aceptan ciertas diferencias en las variables identificadoras[36].

Newcombe generalizó una serie de *dificultades relacionadas con la aplicación del método*[14]. Algunas de las reproducidas en este estudio son: 1) La imposibilidad de calcular RF para todos los niveles de acuerdo posibles. La insuficiente precisión obtenida cuando el número de pares con ciertos niveles de acuerdo era muy reducido o, incluso nulo, obligó a utilizar muestras de mayor tamaño, extraídas de archivos con las mismas características (adictos

ingresados en urgencias). La comprobación de que el uso conjunto de los pares de las muestras y del estándar sólo producía ligeros aumentos de la sensibilidad y del valor predictivo positivo y pequeños descensos de la especificidad y valor predictivo negativo, apoyan la idea de que aunque es preferible disponer de una muestra de los archivos que se van a conectar, también es legítimo el cálculo a partir de archivos representativos. En esta línea, algunas conexiones probabilísticas han asignado pesos artificiales[35], práctica que, aunque puede alterar los pesos finales, facilita la elaboración del proceso; 2) La insuficiencia del poder discriminador de las variables identificadoras. Aun disponiendo de buenos indicadores, no se pudo discriminar en todos los casos, ya que la ausencia del nombre de pila impedía diferenciar entre posibles "gemelos" del mismo sexo; 3) La subjetividad en la elección del umbral más adecuado. Aunque el solapamiento de las distribuciones de frecuencias marca umbrales objetivos, la variación muestral, que a veces excluye casos que una revisión visual consideraría discordantes, obliga frecuentemente a ampliar los márgenes de seguridad; y 4) La desproporción entre el número de falsos positivos y falsos negativos. Con los niveles umbrales previstos [-3/13] se detectaban todos los casos corcondantes del estándar pero había un 2,7% de falsos positivos (pares que cumplen condiciones de los concordantes, p. ej. personas con apellidos y fecha de nacimiento iguales pero sexo diferente). Contrariamente, la aceptación de la concordancia en pares con un valor más negativo que el umbral inferior era en principio descartable ya que implicaba la existencia de diferencias notables entre los diversos identificadores.

Otro inconveniente deriva de la *claridad de los archivos.* El porcentaje de datos desconocidos dificultan la generalización del método al hacerlo dependiente de las características de cada archivo. En este sentido la exclusión de los registros con día y mes desconocidos reducía en gran medida el número de individuos de los archivos, mientras que su inclusión aumentaba el número de pares a revisar y rebajaba la fiabilidad del proceso (los casos duplicados en el mismo archivo distorsionaban la estimación de los pesos). Además del porcentaje de valores ausentes y de errores de codificación habría que tener en cuenta otras características de los archivos. Así, cuando se usen los pesos específicos para la conexión de nuevos archivos debería comprobarse que las frecuencias de las variables identificadoras no cambian (situación improbable cuando se utilizan variables tan generales como los apellidos o el sexo); y en caso contrario, cuando se utilicen archivos con características muy diferentes, corregir los factores de ajuste y hasta las razones de frecuencias.

Una dificultad añadida para la conexión de registros con nombres latinos viene dada por el *proceso de agrupación en bloques,* puesto que los más importantes códigos están basados en la fonética anglosajona (Soundex code, NYSIIS, name compression). Con nuestros códigos se perdieron un 4% de los pares concordantes. Además, la selección del peso mayor permitía que se aceptasen pares discordantes con pesos altos, superiores por azar a los verdaderos concordantes. A pesar de ello parece un buen sistema para reducir el número de pares de registros a revisar, complementario al uso de un método de bloqueo con códigos parcialmente adaptados a la fonética local.

Por último, la *necesidad de aceptar las normas de confidencialidad* de los diferentes registros[37], uno de los motivos que dificultan el consenso entre los responsables de los archivos puede generar problemas de accesibilidad y pérdida de tiempo. No obstante, esto no debiera ser un problema importante puesto que, una vez creados los programas informáticos, puede realizarse la conexión sin identificar los registros individuales. Para soslayar los problemas de confidencialidad bastaría con disponer de una muestra de pares concordantes y discordantes para el cálculo de los pesos específicos o utilizar éstos a partir de un cálculo previo estandarizado.

En conclusión, a pesar de las dificultades que lo condicionan, el método probabilístico sería adecuado para la conexión de archivos de personas drogodependientes y aplicable en nuestro medio con las variables de identificación habituales. Además, su eficiencia aumentaría cuando el volumen de la información registrada fuese muy grande, cuando se realizasen múltiples conexiones y en casos especiales como el uso de datos automatizados, recientemente regulado[38], que al ser "disociados" para guardar su confidencialidad disminuyen la información disponible.

Agradecimientos

Los autores agradecen a Esteve Fernández y Luis Prieto la revisión de anteriores versiones del manuscrito, y a José María Antó por sus comentarios acerca de los focos de interés del estudio.

Se presentaron resultados preliminares de este trabajo al I Congreso Iberoamericano de Epidemiología, Granada (España), en octubre de 1992.

Bibliografía

1. Allgulander C. Psychoactive drug use in a general population sample, Sweden: correlates with perceived health, psychiatric diagnoses, and mortality in an automated record-linkage study. *Am J Public Health* 1989; 79: 1006-10.
2. Newman SC, Bland RC. Mortality in a cohort of patients with schizophrenia: a record linkage study. *Can J Psychiatry* 1991; 36: 239-45.
3. Nienhuis H, Goldacre M, Seagroatt V, Gill L. Incidence of disease after vasectomy: a record linkage retrospective cohort study. *Br Med J* 1992; 304: 743-6.
4. Brand LP, Nielsen CV. Job stress and adverse outcome of pregnancy: a causal link or recall bias? *Am J Epidemiol* 1992; 135: 302-11.
5. Olsen JH. Occupational risks of sinonasal cancer in Denmark. *Br J Ind Med* 1988; 45: 329-35.
6. Guess HA, West R, Strand LM y cols. Fatal upper gastrointestinal hemorrhage or perforation among users and non users of nonesteroidal anti-inflammatory drugs in Saskatchewan, Canada 1983. *J Clin Epidemiol* 1988; 41:35-45.
7. West R, Sherman GJ, Downey W. A record linkage study of valproate and malformations in Saskatchewan. *Can J Public Health* 1985; 76: 226-8.
8. Storm HH. Completeness of cancer registration in Denmark 1943-1966 and efficacy of record linkage procedures. Danish Cancer Registry. *Int J Epidemiol* 1988; 1: 44-9.
9. Roos LL, Sharp SM, Wajda A. Assessing data quality: a computerized approach. *Soc Sci Med* 1989; 28: 175-82.
10. Goldacre MJ, Simmonds H, Henderson J, Gill LE. Trends in episode based and person based rates of readmission to hospital in the Oxford Record Linkage study area. *Br Med J* 1988; 296: 583-5.
11. 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.
12. Pérez C, Domingo A, Mcfarlane D, Hartnoll R. Conexión entre dos registros de urgencias toxicológicas. *X Reunión Sociedad Española de Epidemiología*. Madrid, 1991.
13. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959; 130: 954-9.
14. Newcombe HB. *Handbook of record linkage. Methods for Health and statistical studies, administration and business.* Oxford: Oxford medical publications, 1988.
15. Baldwin JA, Acheson ED, Graham WJ. *Textbook of Medical Record Linkage*. Oxford: Oxford University Press, 1987.
16. Buchanan AV, Weiss KM, Schwartz RJ y cols. Reconstruction of genealogies from vital record: The Laredo Epidemiology Project. *Comput Biomed Res* 1984; 17: 326-51.
17. Arellano MG, Petersen GR, Petitti DB, Smith RE. The California Automated Mortality Linkage System (CAMLIS). *Am J Public Health* 1984; 74: 1324-30.
18. Gill LE, Baldwin JA. Methods and technology of record linkage: some practical considerations. En: Baldwin JA, Acheson ED, Graham WJ. *Textbook of Medical Record Linkage*. Oxford: Oxford University Press, 1987.
19. Ramis-Juan O. Las técnicas del "Record Linkage". *Monografías de Salud Pública. Aplicaciones sanitarias de las estadísticas vitales*. Granada: II Seminario científico de la Sociedad Española de Epidemiología. IV Reunión Anual. 1985; 157-81.
20. Navarro C, Lizán M, Tormo MJ. Usos del certificado de defunción en un registro de cáncer de población. *Gac Sant* 1988; 2: 197-202.
21. Domingo A, Antó JM, Camí J. Epidemiological surveillance of opioid-related episodes in an emergency room of Barcelona, Spain (1979-1988). *Br J A* 1991; 86: 1459-66.
22. Domingo-Salvany A, Hartnoll RL, Antó JM. Opiate and cocaine consumers attending Barcelona Emergency Rooms: A one year survey. *Addiction* (en prensa).
23. Scheuren F. Methodologic issues in linkage of multiple data bases. Vital & Health Statisitics, *NCHS* 1985; 25: 75-87.
24. Knuth DE. Sorting and searching. En: *The art of computer programming*. Vol 3. Addison-Wesley Publishing Co. Inc, 1973: 391.
25. SPSS inc. *SPSS-X User's guide*. Chicago: SPSS inc, 1988.
26. Shapiro S. The role of automated record linkage in the postmarketing surveillance of drug safety: a critique. *Clin Pharmacol Ther* 1989; 46: 371-86.
27. Connell FA, Diehr P, Hart LG. The use of large data bases in health care studies. *Ann Rev Public Health* 1987; 8: 51-74.
28. Leibson CL, Ballard DJ, Whisnant JP, Melton LJ. The compression of morbidity hypothesis: promise and pitfalls of using record-linked data bases to assess secular trends in morbidity and mortality. *Milbank Q* 1992; 1: 127-54.
29. Stergachis AS. Record linkage studies for postmarketing drug surveillance: data quality and validity considerations. *Drug Intell Clin Pharm* 1988; 2: 157-61.
30. Shannon HS, Jamieson E, Walsh C, Julian JA, Fair ME, Buffet A. Comparison of Individual Follow-up and Computerized Record Linkage using the Canadian Mortality Data Base. *Can J Public Health* 1989; 80: 54-7.
31. Stoneburner RL, Des Jarlais DC, Benezra D y cols. A larger spectrum of severe HIV. 1- Related disease in intravenous drug users in New York City. *Science* 1988; 242: 916-9.
32. Perucci CA, Davoli M, Rapiti E y cols. Mortality of Intravenous Drug Users in Rome: A Cohort Study. *Am J Public Health* 1991; 81: 1307-10.
33. Strom BL, Carson JL. Use of automated databases for pharmacoepidemiology research. *Epidemiol Rev* 1990; 12: 87-107.
34. Pineault R, Champagne F, Fournier P. L'exploitation de grandes bases de données sur la morbidité pour l'évaluation des services de santé. *Rev Epidem Sante Publique* 1988; 36: 267-72.
35. Van Den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Humen PMH. Development of a Record Linkage Protocol for Use in the Dutch Cancer Registry for Epidemiological Research. *Int J Epidemiol* 1990; 19: 553-8.
36. FETT MJ. The development of matching criteria for epidemiological studies using record linkage techniques. *Int J Epidemiol* 1984; 13: 351-5.
37. Stern RS. Record linkage. A powerful tool for epidemiologic analysis [editorial]. *Arch Dermatol* 1986; 122: 1383-4.
38. Ley orgánica 5/1992 de 29 de octubre de regulación del tratamiento de los datos de carácter personal. *BOE* núm. 262 de 31 de octubre de 1992.