COLABORACIÓN ESPECIAL

Recibido: 20 de junio de 2017
Aceptado: 7 de agosto de 2017
Publicado: 20 de octubre de 2017

MÉTODOS INDIRECTOS PARA LA ESTIMACIÓN DE POBLACIONES OCULTAS (*)

Rocío Lorenzo Ortega (1,2), Michela Sonego (3,4), José Pulido (3,5,6), Almudena González Crespo (4), Eladio Jiménez-Mejías(3,7), Luis Sordo(3,4,6).

(1) Servicio de Medicina Preventiva. Hospital Virgen de la Victoria. Málaga. España.

(2) Departamento de Medicina Preventiva y Salud Pública e Historia de la Ciencia. Universidad de Málaga. Málaga. España.

(3) Centros de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP). Madrid. España.

(4) Centro Nacional de Epidemiología. Instituto de Salud Carlos III. Madrid. España.

(5) Escuela Nacional de Sanidad. Instituto de Salud Carlos III. Madrid. España.

(6) Departamento de Medicina Preventiva y Salud Pública. Facultad de Medicina. Universidad Complutense de Madrid. España.

(7) Departamento de Medicina Preventiva y Salud Pública, Universidad de Granada. España.

(*) Financiación: parcialmente financiado por "Ministerio de Economía y Competitividad, Actuación de Formación Posdoctoral" (FPDI-2013-15827).

RESUMEN

En determinadas situaciones podemos encontrar dificultades a la hora de calcular prevalencias en algunas poblaciones. Es el caso de personas que tienen comportamientos que son difíciles de identificar debido a que pueden estar sancionados socialmente o ser ilegales. Es lo que llamamos poblaciones ocultas. Este artículo proporciona una revisión crítica de los métodos más utilizados para calcular el tamaño de una población de difícil acceso. Se trata de métodos indirectos, que estiman la prevalencia de una población oculta basándose en fuentes de datos incompletas. Se exponen diferentes métodos, cada uno de ellos tiene diferentes indicaciones para ser utilizado, dependiendo de los datos de los que dispongamos. Además, precisan de una serie de requisitos para poder ser aplicados y cada uno está expuesto a diferentes tipos de sesgos. Por estos motivos, hay que valorar correctamente los datos disponibles para aplicar el método más preciso, y si fuese posible, utilizar simultáneamente varios métodos para poder comparar los resultados obtenidos, además de valorar críticamente los resultados y comprobar si se ajustan a la realidad.

Palabras clave: Poblaciones Vulnerables, Métodos Epidemiológicos, Vigilancia Epidemiológica, Recogida de Datos.

ABSTRACT
Indirect methods to estimate hidden population

Estimating the prevalence of the so-called "hidden populations" can be challenging, because the identification of its members is difficult due to their socially sanctionable or illegal behaviors. This article provides a critical review of the most widely used methods for estimating the size of a hard-to-reach population. All are indirect methods, based on incomplete data sources. Depending on the available data, one method can be more appropriate than another. Besides, each method must fulfill a number of requirements, and each one may be subject to specific risk of bias. To choose the most suitable method, an accurate evaluation of the available data is necessary, and. if possible several methods should be used simultaneously to be able to compare the results and to critically evaluate if these results fit with the reality.

Keywords: Vulnerable Populations, Epidemiologic Study Characteristics as Topic, Epidemiological Monitoring, Data Collection.

Correspondencia:
José Pulido
Escuela Nacional de Sanidad. Instituto de Salud Carlos III.
Monforte de Lemos 3-5, 28029, Madrid.
jpulido@isciii.es

Cita sugerida: Lorenzo Ortega R, Sonego M, Pulido J, González Crespo A, Jiménez-Mejías E, Sordo L. Métodos indirectos para la estimación de poblaciones ocultas. Rev Esp Salud Pública, 2017, vol 91: 20 de octubre e201710039.

INTRODUCCIÓN

Las poblaciones ocultas o de difícil acceso tienen dos características que las definen: en primer lugar, sus miembros son difíciles de identificar debido a que comparten características que pueden estar sancionadas socialmente, ser estigmatizadoras o ilegales. En segundo lugar, son poblaciones que carecen de marco muestral; no conocemos ni su tamaño ni su distribución. Como ejemplo podríamos citar las personas que usan drogas, las que ejercen la prostitución o las que padecen infecciones de transmisión sexual[1,2] . Se trata habitualmente de colectivos vulnerables desde el punto de vista socio-sanitario, de los cuales es importante conocer el número para determinar sus necesidades. Sin embargo los métodos convencionales de estimación de prevalencia, llamados directos, como las encuestas poblacionales, difícilmente muestran datos certeros del tamaño de estas poblaciones. Por un lado porque sus miembros tienen menos probabilidad de ser escogidos por los métodos de muestreo usuales (teléfono, domicilio fijo, tarjeta sanitaria, etc). Por otro, sus características definitorias son aspectos de la vida que no son fáciles de revelar en una encuesta [3], afrontándose un importante sesgo de información.

Como punto de partida para poder calcular prevalencias en estas poblaciones muchas veces lo único que tenemos de ellas son fuentes de datos incompletas. Por eso es necesario recurrir a los métodos indirectos, que partiendo de la premisa de que las fuentes de datos disponibles no son completas, realizan diferentes cálculos para conseguir estimar la prevalencia real[4].

El propósito de este artículo es realizar una revisión crítica de los métodos indirectos más utilizados (tabla 1) para obtener estimaciones de la prevalencia de poblaciones de difícil acceso.

PRINCIPALES MÉTODOS INDIRECTOS PARA LA ESTIMACIÓN DE POBLACIONES OCULTAS

A. Método de Benchmark-Multiplier

Para la aplicación de este método tan solo necesitamos una cifra absoluta como punto de referencia, –llamada Benchmark,– y una tasa de incidencia relacionada (multiplicador). A partir de estos dos datos, se estima la prevalencia multiplicando el Benchmark por la inversa del multiplicador.

Imaginemos que la población oculta que queremos estimar es el número de personas usuarias de drogas en una ciudad determinada en un determinado año. Para su cálculo por este método podríamos partir del número de muertes (Benchmark) relacionadas con drogas ese año en dicha ciudad (dato accesible y fiable en registros de mortalidad). Adicionalmente necesitaríamos un multiplicador, que en este caso podría ser la tasa de mortalidad de los usuarios de drogas obtenida en estudios sobre cohortes de estas poblaciones. Así pues, si tuviéramos 405 fallecidos en un año (Benchmark), y una tasa de mortalidad de 2.3% (2.3 por 100 personas-año), tendríamos una prevalencia estimada de 17609 personas usuarias de drogas en esa ciudad en un año [5] (figura 1).

Sin embargo, hay una serie de requisitos que debemos tener en cuenta a la hora de llevar a cabo este método y que determinan sus limitaciones: 1) El punto de referencia o Benchmark debe ser exhaustivo, completamente fiable [3,4]; 2) El muestreo usado para estimar el multiplicador debe ser representativo de la población a la que va dirigido y obtenido de forma independiente del punto de referencia[3,4]; 3) la definición de caso usada para el punto de referencia debe coincidir (en lugar y tiempo) con aquella usada para obtener el multiplicador[3,4].

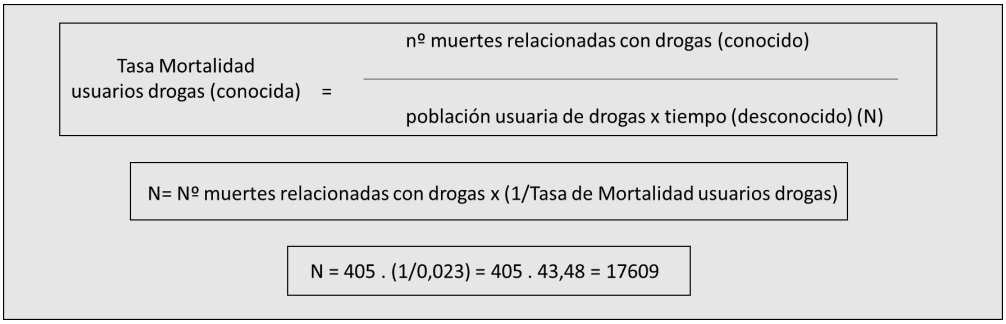
Cuando hablamos de estimaciones de parámetros hay que tener el nivel de incertidumbre de estas estimaciones, debiendo acompa-

Tabla 1
Principales métodos indirectos de cálculo de poblaciones de difícil acceso

Métodos	Requisitos	Ventajas	Inconvenientes
Método Benchmark-Multiplier: Basado en multiplicar un número de personas conocido de una fuente de datos (ej.: muertes en usuarios drogas), por una tasa conocida (ej.:tasa mortalidad en usuarios drogas).	- Punto de referencia exhaustivo. - El muestreo usado para estimar el multiplicador debe ser representativo de la población a la que va dirigido. - La definición de caso usada para benchmark y multiplicador debe coincidir.	- Sencillez y facilidad.	- Posibilidad de sesgos si no se aplican bien los requisitos. - No aplicable si población heterogénea.
Captura-Recaptura: Basado en comparar varias fuentes de datos de la población de estudio, valorar el grado en que se repiten los individuos, e inferir el número total en la población.	- Población de referencia cerrada. - No existencia de falsos positivos. - Fuentes de datos independientes. - Cada caso igual probabilidad de ser capturado en cada lista. - Bases de datos representativas de la población.	- Sencillez. - Ampliamente utilizado, por lo que existe abundante experiencia.	Posibilidad de sesgos si no se aplican bien los requisitos: - Dependencia negativa: Supraestimación. - Dependencia positiva y Heterogeneidad: Infraestimación
Técnicas de Nominación: Bola de nieve vs Muestreo dirigido por los participantes (RDS): Técnicas basadas en el contacto directo con los participantes. Éstos proporcionan, a su vez, el acceso y/o información sobre otros usuarios.	- Muestra representativa. - Datos autorreportados fiables. - Población objetivo correctamente definida. - Población de estudio no segmentada, conectada por densas redes sociales.	- Proceso barato, simple y rentable. - Precisa poca planificación y mano de obra (excepto RSD). - RSD: Permite obtener estimadores generalizables a la población de referencia de donde se ha extraído la muestra.	- No permite obtener estimadores generalizables a la población de referencia (excepto RDS). - Sesgo de muestreo: Posibilidad de que los participantes pertenezcan a un subgrupo de la población a estudio (excepto RDS).
Método de Poisson Truncado: A través del número de individuos que han tenido contacto 1 vez, 2 veces, 3veces, etc. se estima el número. de individuos que no han tenido contacto en ninguna ocasión.	- Debe figurar el número de contactos con la fuente. - La probabilidad de contacto con el registro un número de veces determinado es constante para todos los individuos. - Esa probabilidad es independiente del número de veces que se haya contactado antes. - Precisa que la variable siga distribución de Poisson.		- Mayor complicación de los cálculos. - Precisa número de contactos con la fuente.

RDS: Responding Driving Sampling

Figura 1
Formulación y ejemplo del método Benchmark-Multiplier



ñarlas de su error estándar y/o intervalo de confianza. Sin embargo en el método de Benchmarck-multiplier, el intervalo de confianza podría proporcionar una falsa sensación de seguridad, ya que son muchos los posibles sesgos[5]. En este caso, lo que algunos autores recomiendan es comparar los resultados de este método con los de otros estudios de Benchmark-multiplier o de otras metodologías y valorar su grado de concordancia[5].

A modo de conclusión, se trata pues de un método muy fácil de aplicar por lo que es ampliamente utilizado. Sin embargo, pueden existir errores si los datos de referencia no son exactos, o no existe un multiplicador correctamente definido. Además, el estimador es poco fiable si la población objetivo es heterogénea.

B. Método Captura-Recaptura

Este método permite determinar la prevalencia de una población a partir de las coincidencias resultantes entre dos o más fuentes de datos incompletas.

Cuantas más fuentes de datos tengamos a nuestra disposición, mejor y más exhaustivo será el método. No obstante, para simplificar su explicación, en el presente artículo se presentará un ejemplo en el que solo se emplean dos.

Continuando con el ejemplo anterior, de nuevo se quiere estimar el número de usuarios de opiáceos en un país en un año. Para ello contamos con dos fuentes de datos que sabemos incompletas. Incluyen miembros de nuestra población de interés, pero sabemos que tan solo incluyen parte de la misma. Una podría ser el registro de un centro de dispensación de metadona (CDM) y otra, la del registro de la policía nacional de detenciones relacionadas con el consumo de opiáceos (RPN). Llamaríamos m a los usuarios presentes en ambas fuentes y N la población desconocida, la que queremos calcular. Así, asumiendo que ambas fuentes de datos son independientes, y que el hecho de aparecer en una no altera la posibilidad de aparecer en la otra, la Odds Ratio sería igual a 1. (figura 2). De ahí podemos extraer que d = (b*c/a), y a partir de este dato, calcular el número total de población que queremos conocer, que es N (N= a+b+c+d). Así, si tuviéramos 354 usuarios registrados en el CDM, y 221 en el RPN y sólo 11 coincidieran en ambas bases, la estimación de la población de usuarios de opiáceos sería de 7112 (figura 2).

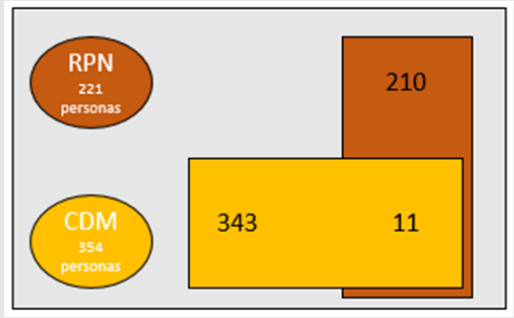
Para el cálculo del intervalo de confianza sugerimos usar la fórmula que ofrecen en el manual "*Estimating the Prevalence of Problem Drug Use in Europe*" de Hartnoll et al, la página 78, referente a la varianza de N (nuestra población a conocer)[3].

Figura 2
Formulación y ejemplo del método captura-recaptura

$$\text{Odds Ratio (OR)} = \frac{a/b}{c/d} = 1$$

$$d = \frac{c*b}{a}$$

$$N = a+b+c+d$$



		RPN (Recaptura)		
		Si	No	
CDM (Captura)	Si	a=11	b=343	354
	No	c=210	d	
		221		N

CDM: Centro dispensación metadona
RPN: Registro policía nacional

N: Población total que queremos conocer
m: nº de individuos que aparecen en ambas bases

Este método también precisa una serie de requisitos: 1) La población de referencia debe ser cerrada (un número fijo de personas), por lo tanto se recomienda usar tiempos de estudio cortos, de un año aproximadamente[3]; 2) Las fuentes de datos deben ser independientes (el aparecer en una de las fuentes no debe predisponer a aparecer o no en la otra) y representativas de la población a estudio[3]; 3) Cada caso en la población debe tener la misma probabilidad de ser capturado en cada fuente[3,5]. Estos requisitos son bastante exhaustivos y difíciles de conseguir por completo, por lo que dependiendo de su mayor o menor grado de cumplimiento, podremos estimar nuestra prevalencia con una mayor o menos precisión.

Los posibles sesgos a tener en cuenta si se aplica esta metodología son claros: el grado de independencia de las fuentes de datos y la falta de representatividad. En cuanto a la independencia de las fuentes de datos, es conveniente señalar que cuando trabajamos con tres o más fuentes, podemos evaluar este requisito a través de diferentes modelos de ajuste de los datos, pero este requisito no se puede valorar cuando trabajamos únicamente con dos fuentes. Una vez comprobado, si las fuentes de datos no son independientes, el hecho de aparecer en una puede implicar mayor posibilidad de aparecer en la otra: esto se denomina dependencia positiva, y supone un riesgo de infraestimación de la población. En el caso opuesto, denominado dependencia negativa, aparecer en una fuente de datos disminuye la posibilidad de aparecer en la otra, conllevando una sobreestimación de nuestra población[3]. En cuanto a la representativi-

dad de las fuentes de datos, ésta depende de la homogeneidad de la población a estudio. Si nuestra población fuese muy heterogénea podemos caer en un importante sesgo de infraestimación, pues en las fuentes de datos solo podría estar representado un subgrupo de esta población. Este sesgo se puede evitar realizando un análisis por dichos subgrupos[3].

El método de captura-recaptura es también un método fácil de aplicar y ampliamente utilizado en epidemiología, pero presenta unas suposiciones que no son siempre fáciles de conseguir, por lo que no está exento de sesgos.

C. Técnicas de Nominación

Engloban un conjunto de técnicas que tienen en común el contacto directo con los participantes, los cuales proporcionan el acceso a otros usuarios o a determinada información[6].

El más conocido, aunque no pueda ser usado para el cálculo de prevalencias, es el muestreo en bola de nieve. En él, se contacta con sujetos de la población de interés y estos a su vez ayudan a reclutar a otros sujetos de la misma población. Se va repitiendo el proceso hasta alcanzar un número de muestra predeterminado[3]. Es un proceso simple, barato y rentable muy utilizado en estudios etnográficos. Sin embargo, desde un punto de vista epidemiológico, este método produce estimadores sesgados y no generalizables.

La variante de este tipo de muestreo que sí permite determinar prevalencias es el denominado muestreo dirigido por los participantes (*respondent-driven sampling*). Este modelo trata de protocolizar el muestreo en cadena o en bola de nieve y obtener estimadores no sesgados que se puedan generalizar a la población general. El muestreo se inicia mediante la selección de informantes iniciales (semillas) de forma similar al método bola de nieve, los cuales no entrarán en el análisis posterior. Éstos, a su vez, seleccionan un número limitado de nuevos participantes y así sucesivamente. Cada grupo de participantes derivado de una semilla es una cadena, y el grupo reclutado en cada etapa, es una ola (figura 3). Se generará un número suficiente de olas para que las diferentes variables que puedan llevar a confusión queden estabilizadas y la muestra obtenida sea representativa de la población[1,2].

La principal ventaja de este método con respecto a otras técnicas de nominación es precisamente la sistematización del proceso y el equilibrio que alcanzan las posibles variables de confusión al finalizar el proceso.

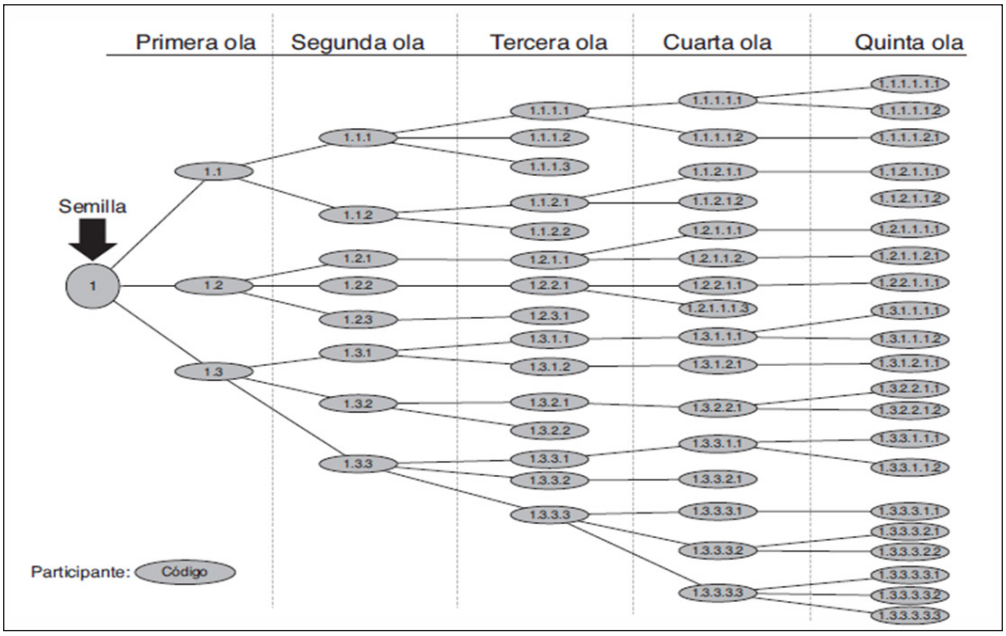
Como todos los métodos anteriores, las técnicas de nominación también deben cumplir unos requisitos: 1) Los datos aportados por los participantes deben ser muy fiables[3]; y 2) La población diana debe estar correctamente definida (no segmentada) y conectada por densas redes sociales[1].

D. Método de Poisson Truncado

Este método permite realizar una estimación a partir de una única fuente de datos que sabemos incompleta. En esta fuente tendremos los individuos que han tenido contacto con un determinado servicio (como pudieran ser un programa de intercambio de jeringas o una consulta de enfermedades de transmisión sexual), y el número de veces que lo han hecho. El modelo de Poisson truncado permite inferir la población que no ha entrado en contacto con dicho servicio ninguna vez a partir de los datos de los individuos que sí han contactado. Es decir, a través del número de individuos que han tenido contacto una vez, de los que han tenido contacto dos veces, de los que han tenido contacto tres veces y así sucesivamente, se puede estimar el número de individuos que no han tenido contacto en ninguna ocasión[7,8].

El método de Poisson truncado tiene diferentes variantes, aunque la más utilizada es la ecuación de Zelterman[9]. (figura 4). Tomemos como ejemplo una base de datos de una unidad de dispensación de jeringas y agujas estériles compuesta por 403 usuarios de he-

Figura 3
Presentación esquematizada del método dirigido por los participantes



Fuente: Sordo L, Pérez-Vicente S, Rodríguez del Águila MM, Bravo MJ. Muestreo dirigido por los participantes para el estudio de poblaciones de difícil acceso. Med Clin (Barc) [Internet]. 2013;140(2):83–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0025775312007087

Figura 4
Formulación del método de Poisson truncado

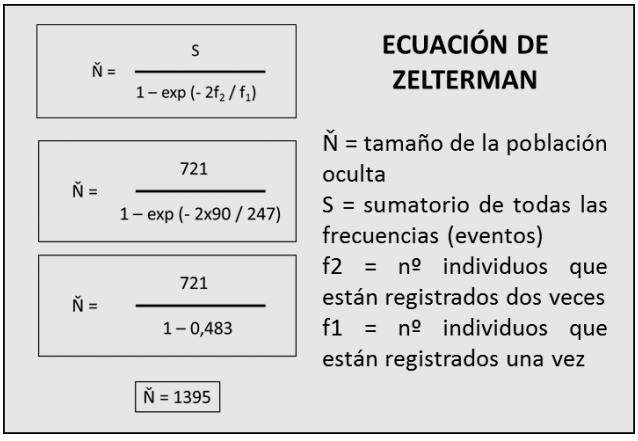
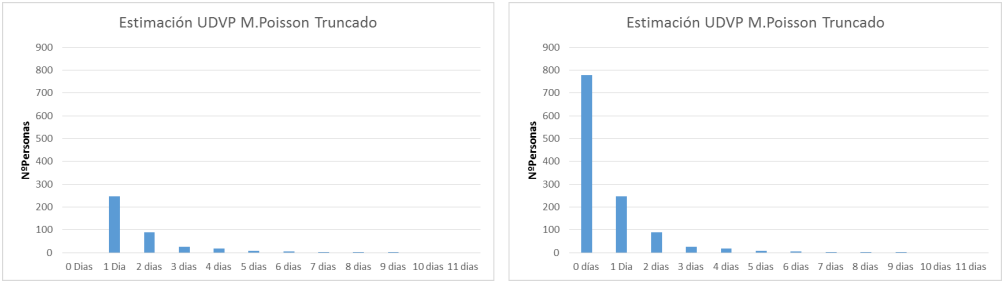


Figura 5
Ejemplo de método de Poisson truncado



roína y un total de 721 contactos durante el transcurso de 1 año. De los 403 individuos, 247 habían acudido sólo una vez, 90 dos veces, y el resto tres veces o más. Sustituyendo los parámetros en la ecuación principal (figura 4), se puede estimar una población total de usuarios de heroína inyectada de 779 personas. (figura 5). Para el cálculo del intervalo de confianza para este método, remitimos a otros trabajos [10,11].

Los requisitos que exige este método son: 1) Que la probabilidad de contactar con la fuente de datos sea la misma para todos los individuos; 2) Que esa probabilidad sea independiente del número de veces que se haya contactado previamente y 3) Que la población sea homogénea[7,8].

E. Otros Métodos

Además de los métodos descritos existen otras técnicas de estimación de poblaciones más específicas y que aquí simplemente vamos a mencionar porque al ser menos utilizadas no son objeto de este estudio.

Son diferentes variantes de los métodos vistos anteriormente, como el modelo de covariantes en el método de captura-recaptura[4] que permite comprobar la heterogeneidad de los individuos de las fuentes de datos, y ajustar los datos mediante estratificación en subgrupos. Otro método usado en ocasiones es el llamado de cálculo retrospectivo [4], en el cual, conociendo la incidencia y el punto final de un determinado proceso (uso de drogas), podemos estimar el punto inicial del mismo (inicio del consumo) y calcular así la prevalencia de dicho proceso.

CONCLUSIONES

Disponemos de diferentes formas para calcular de forma indirecta el tamaño de poblaciones ocultas. Todas ellas se basan en el empleo de fuentes de datos incompletas.

No existe un método ideal a utilizar, por lo que nuestra elección variará en función de los datos que dispongamos y de la evaluación epidemiológica de los mismos. En referencia a los datos, si sólo disponemos de una fuente, por ejemplo, registro de sujetos que acuden a un servicio o programa de dispensación de material estéril (agujas y jeringuillas), se recomienda el método de Poisson truncado. Si además del acceso a una fuente de datos (p. e., el registro de mortalidad por consumo de drogas durante un año) conocemos la existencia de otros datos complementarios (como la tasa de mortalidad entre los usuarios de drogas)

podemos aplicar los métodos multiplicativos; y si disponemos de información de dos o más fuentes de datos (independientes entre sí), es recomendable aplicar el método Captura-Recaptura. La evaluación epidemiológica determinará la evaluación de sesgos en las fuentes de datos o registros en función de la población de estudio, su fiabilidad y el grado de cumplimiento de los requisitos para el empleo de las diferentes metodologías.

Por todo ello, a la luz de la revisión de estos métodos indirectos, para la estimación de la prevalencia de las poblaciones ocultas se recomienda aplicar varios métodos sobre la misma población evaluando cuidadosamente las limitaciones de cada uno de ellos y la concordancia entre los resultados obtenidos. Finalmente se valorará la plausibilidad y coherencia de dichos resultados con la realidad.

BIBLIOGRAFÍA

1. Sordo L, Pérez-Vicente S, Rodríguez del Águila MM, Bravo MJ. Muestreo dirigido por los participantes para el estudio de poblaciones de difícil acceso. Med Clin (Barc) [Internet]. 2013;140(2):83-7.

2. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance. Aids. 2005;19(Suppl 2):S67-72.

3. Hartnoll R, Cohen PDA, Domingo-Salvany A, Simon R, Frischer M, Taylor C, et al. Estimating the Prevalence of Problem Drug Use in Europe. Strasbourg; 1996.

4. United Nations Office on Drugs and Crime. Estimating Prevalence : Indirect Methods for Estimating the Size of the Drug Problem Estimating Prevalence : Austria. 2003.

5. Hickman M, Taylor C, Chatterjee a., Degenhardt L, Frischer M, Hay G, et al. Estimating the prevalence of problematic drug use: a review of methods and their application. Bull Narcotics. 2002;65:15-32.

6. Hay G, Kraus L. Scientific Review of the Literature on Estimating the Prevalence of Drug Misuse on the Local Level. 1999.

7. van der Heijden PGM, Cruts G, Cruyff M. Methods for population size estimation of problem drug users using a single registration. Int J Drug Policy [Internet]. 2013;24(6):614-8.

8. Van der Heijden PGM, Cruyff M, Van Houwelingen HC. Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model. Stat Neerl. 2003;57(3):289-304.

9. Hay G. International Consultant on Estimation of the Prevalence of Problem Drug use in Lithuania. 2007.

10. Cruyff MJLF, van der Heijden PGM. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. Biom J. 2008;50(6):1035-50.

11. van Der Heijden PGM, Bustami R, Maarten JLF Cruyff MJLF, Engbersen G, van Houwelingen HC. Point and interval estimation of the population size using the truncated Poisson regression model. Stat Model. 2003;3(4):305-22.

12. Astrauskienė A, Dobrovolskij V, Stukas R. The prevalence of problem drug use in lithuania. Med Kaunas Lith [Internet]. 2011;47(6):340-6.