

Aplicación del software libre R en el análisis de problemas de salud de causalidad multivariada

José A. Betancourt-Bethencourt⁽¹⁾ José Félix García-Rodríguez,⁽²⁾ Gaspar Abreu Salgado⁽³⁾

jfgr55@hotmail.com

RESUMEN

Introducción. Existen problemas de salud pública donde es necesario evaluar las informaciones y los datos para apoyar la toma de decisiones. **Objetivo.** Exponer posibilidades del programa R para lidiar con problemas de salud pública de causalidad multivariada. **Material y Método.** Durante el mes de enero de 2014 se realizó un estudio observacional analítico transversal en el policlínico Tula Aguilera del municipio de Camagüey, Cuba, que consistió en realizar una encuesta para identificar diversos factores de riesgos en los lugares que en esta comunidad tuvieron enfermos de diarreas infecciosas agudas. **Resultados.** Se obtienen estadígrafos de las variables estudiadas así como las frecuencias de las variables cualitativas, se evaluó la asociación o efecto entre el número de personas de los núcleos familiares y se obtuvo el mapa con la geo localización precisa de los casos. No se encontraron asociaciones entre el número de personas de los núcleos familiares y la cantidad de casos. Se detalla el comportamiento de los riesgos ambientales encontrados. **Conclusión.** Se muestra la utilidad del programa libre R en el análisis epidemiológico de problemas de salud pública de causalidad multivariada.

Palabras clave: *Herramientas de software, estadística, epidemiología.*

SUMMARY

Introduction. There are problems of public health where it is necessary to evaluate the information and the data to support the decisions making. **Objective.** The goal of this article is to show the advantages of using R to analyze public health's multi-causality problems. **Material and method.** During the month of January of 2014, an analytic observational study

took place in the policlinic Tula Aguilera of the municipality of Camagüey, Cuba. It consisted on carrying out a survey to identify diverse risks factors of communities confirmed by persons who had been sick with infectious diarrheas. **Results.** some environmental risk associated were identified, it was evaluated the association or effect among the household number and sick people, a map was obtained with the precise geolocalización of cases. They were not associations between the household number and the quantity of cases. **Conclusion.** The utility of free software R3 epidemiological analysis of public health problems of multivariate causality appears. We need to increase transdisciplinary and intersectoral actions to the community, really change habits and lifestyles.

Keywords: *Software Tools, statistics, epidemiology*

INTRODUCCIÓN

En la etapa inicial de un brote epidemiológico existen imprecisiones y fragmentaciones de las informaciones disponibles. La información real está limitada hasta que se hagan confirmaciones mediante exámenes específicos y análisis serológicos. Existe además co-circulación de múltiples cadenas de entidades biológicas, casos asintomáticos o con síntomas leves que no se detectan ni reportan, todo lo que trae retrasos en el diagnóstico y planteamiento de solución, lo que complica aún más la situación. En dichas circunstancias epidemiológicas, el análisis de datos estadísticos e información diversa resulta fundamental.¹

En la práctica, es usual la aplicación de diversas herramientas estadísticas y matemáticas aplicables a la modelación epidemiológica. Para un mejor desarrollo de este tipo de estudios, está disponible en la red de internet el programa libre R,² en el que se encuentran diversos modelos estadísticos y

⁽¹⁾ Profesor auxiliar de Salud Pública. Miembro del Departamento de Investigaciones de la Universidad de las Ciencias Médicas de Camagüey, Cuba.

⁽²⁾ Doctor en Finanzas Públicas. Profesor Investigador. Universidad Juárez Autónoma de Tabasco, México.

⁽³⁾ Especialista de primer grado en Epidemiología. Universidad de las Ciencias Médicas de Camagüey, Cuba.

matemáticos con código abierto y modificables, con paquetes específicos para los estudios epidemiológicos como el *epicalc*³ y otros para los análisis de mapas como el *osmar*,⁴ los cuales se están usando ampliamente y evalúan la interacción de los individuos y su entorno, lo que incluye infinidad de variables como pueden ser ambiente, clima, entorno laboral, familiar, político y otros.

Christian Salas en su artículo ¿Por qué comprar un programa estadístico si existe R?⁵ realiza una comparación entre los diferentes paquetes estadísticos comerciales y el R, a modo de resumen a continuación reflejamos algunos de sus argumentos: “El R es totalmente gratuito, su uso es libre, y su distribución gratuita con código abierto y modificable, a diferencia de los usuales programas como el SPSS y el SAS, estos últimos se enfocan en aquellos mercados y usuarios que les proporcionan los mayores beneficios. El R tiene una amplia variedad de análisis estadísticos y está desarrollado como un gran proyecto colaborativo de estadísticos de diversos países y disciplinas. El R permite el diseño personalizado de gráficos y se pueden escribir funciones propias de manera sencilla. R es el único programa que funciona de manera estable e íntegra en los tres sistemas operativos de mayor uso.”

Los programas estadísticos comerciales son excelentes pero de alto costo, por lo que en muchos casos se utilizan copias no autorizadas, lo cual limita la calidad de las publicaciones. Hay otros programas libres como el *Epinfo*, pero tiene limitaciones en análisis más complejos y en sus facilidades para generar gráficos. Frente a ellos, el R tiene poderosas herramientas para generar gráficos y su capacidad de análisis es ilimitada. El propósito central de este trabajo es exponer la utilidad del programa libre R para modelar problemas de salud pública de causalidad multivariada asociados a diversos determinantes y condicionantes, como son entre otros, los brotes de diarreas de etiología infecciosa que han tenido lugar en el municipio de Camagüey, Cuba.

MATERIAL Y MÉTODO

Población de Estudio. Durante el mes de enero de 2014, se llevó a cabo un estudio piloto de tipo observacional analítico transversal en el policlínico Tula Aguilera del municipio de Camagüey, Cuba, que partió de la aplicación de una encuesta exploratoria para identificar diversos factores de riesgo. En los hogares que en esta comunidad tuvieron casos de enfermos de diarreas de etiología infecciosa, las personas entrevistadas constituían el total de enfermos en el área seleccionada. Las variables investigadas son las usualmente reportadas por la literatura para esto casos: edad, sexo, tipo de abasto de agua, personas en el núcleo familiar, distancia entre fosa y pozo, letrinas sin tapas, pozos de aguas abajo de las letrinas, residuales de la casa tapados, y distancia al micro vertedero más cercano.

Objetivo del Proyecto. Exponer la utilidad del programa libre R para analizar problemas de salud pública de causalidad multivariada, asociados a diversos determinantes y condicionantes socioeconómicos y medioambientales que inciden en el panorama epidemiológico de una región.

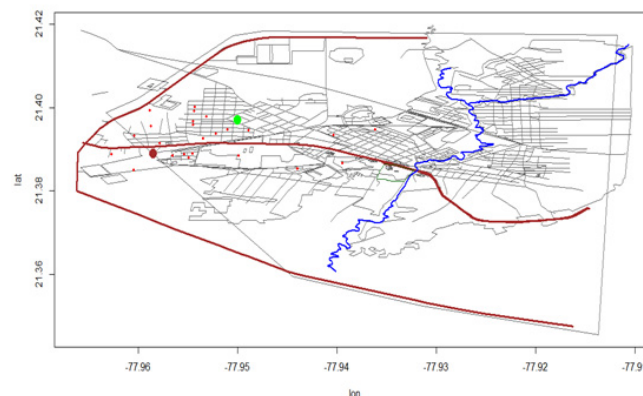
Método. La información se recolectó casa a casa, se realizó por personal entrenado de la propia área de salud y supervisado por un epidemiólogo. Se determinaron estadígrafos de tendencia central y dispersión de todas las variables estudiadas según procediera. Se recogió la geo localización de los pacientes, traduciendo el número de manzana a la latitud y longitud correspondiente, para confeccionar el mapa geográfico se definieron previamente fronteras del área a estudiar y se trabajó con el paquete *osmar*, en el cual se introdujeron las geo localizaciones de los casos registrados. Para indagar sobre posibles efectos del número de personas en los núcleos familiares sobre la cantidad de personas afectadas, se realizó una regresión Poisson. Todos los análisis estadísticos se realizaron con el programa libre R aplicando el paquete *epicalc* y el paquete *osmar*.

RESULTADOS

Se exponen posibilidades de aplicación del programa R para lidiar con problemas de salud pública de causalidad multivariada, como por ejemplo:

Con el paquete *osmar* fue posible ubicar casos y áreas de riesgos. En el mapa generado (figura 1), se pueden observar los casos de enfermos (puntos rojos), que en esta área formaron dos clústeres y se muestran además dos áreas de riesgos externos (círculos verde y marrón).

Figura 1. Casos de diarreas infecciosas identificados en un área del Tula Aguilera.

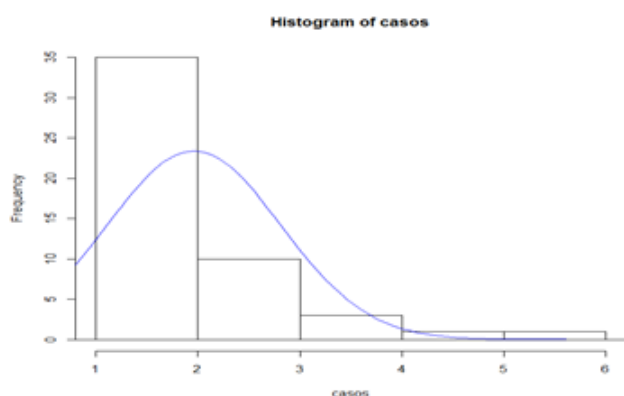


Fuente: Resultados del análisis aplicado con el paquete estadístico *osmar*.

Además permitió caracterizar a la población afectada donde: se obtuvieron con el paquete epicalc los estadígrafos y frecuencia de datos, determinándose que la mediana de edad de las 50 personas encuestadas en hogares donde había casos de diarreas de etiología infecciosa fue de 33 años, de ellos 21 eran mujeres y 29 hombres. En general promediaban dos casos de enfermos por hogar con poca variabilidad. Asimismo, solo una casa recibe agua del acueducto, las 49 restantes dependen de pozos.

Por otro lado, la aplicación del paquete epicalc permitió realizar análisis de frecuencia de los datos. En este caso, la información de los enfermos existentes en el área analizada presentaba una distribución de frecuencia Poisson, por lo que se realizó una regresión adecuada a este comportamiento para buscar posibles efectos del tamaño del núcleo familiar sobre la cantidad de casos presentados. (Figura 2).

Figura 2. Distribución de frecuencia de casos.



Fuente: Resultados del análisis aplicado con el paquete estadístico epicalc.

En las salidas obtenidas durante el análisis, no se demostró asociación o relación entre la cantidad de miembros del núcleo familiar y la cantidad de casos ($p=0.115$, AIC 157,6), el valor de la bondad de ajuste del modelo demuestra que fue adecuado ($p=0.115$, $\chi^2=29.6578$, df. 48). No se encontraron asociaciones entre el consumo de agua de pozo y la transmisibilidad ($p=0.459$), la edad ($p=0.8604$) y los demás factores ambientales e higiénicos estudiados, aunque sí son identificados como factores de riesgos. (Tabla 1).

Tabla 1. Factores de riesgo detectados en el área.

Factor de riesgo	Presente	% casos
Letrinas sin tapa	30	60
Pozos aguas abajo	4	8
Residuales sin tapar	33	66
Consumo de agua sin clorar	40	80
Casas cercanas a microvertederos	21	42

Fuente: Resultados del análisis aplicado con el paquete estadístico epicalc.

DISCUSIÓN

Otros programas como el SAS o el SPSS pueden utilizarse sin dudas para el análisis de problemas de salud pública de causalidad multivariada, lo que sucede es que en el ámbito latinoamericano y en zonas muy pobres, pagar estos programas comerciales es algo prohibitivo y no sostenible. En los análisis, el uso del programa libre R demostró una amplia aplicabilidad. En la epidemiología moderna se está usando de manera muy amplia en la vigilancia y en los análisis espaciales.^{6,7}

En el estudio del panorama epidemiológico de una zona cercana al policlínico Tula Aguilera del municipio de Camagüey, Cuba, mediante la aplicación del paquete osmar se logró obtener una precisa y eficiente geo localización de casos por latitud y longitud; esta geo localización permitió ubicar los casos con una precisión de hasta dos metros. De esta manera, fue factible visualizar de manera adecuada la situación epidemiológica, lo que permitió apoyar la toma de decisiones y el trabajo de campo, tal como se reporta en varias latitudes,⁸ y preparan las condiciones para realizar posteriores análisis espaciales.

Mediante el programa epicalc se llevaron a cabo diversos análisis estadísticos. La regresión realizada no demostró asociaciones o efectos del número de personas en el núcleo familiar sobre la cantidad de casos, a pesar de que se reportan asociaciones en la literatura internacional sobre la cantidad de miembros del núcleo familiar sobre la dispersión de diarreas de etiología infecciosa.⁹ A juicio de los autores, esto se debe a la rápida y coordinada actuación del área de salud que realiza aislamiento de los enfermos y tratamiento quimio-profiláctico preventivo a los miembros aparentemente sanos del núcleo familiar, al tratamiento ambiental intra y extra domiciliario, y a una mayor información de los miembros de la familia sobre las vías de transmisión de estas enfermedades y las maneras de prevenirlas. Todo ello ha demostrado tener un impacto positivo en la disminución de la transmisibilidad de las enfermedades en el ámbito familiar. En otras latitudes, algunos investigadores han demostrado el efecto protector de

la quimioprofilaxis en los núcleos familiares, y argumentan la necesidad de profundizar en investigaciones para valorar daños y beneficios de la quimioprofilaxis.¹⁰

Los análisis de frecuencia de datos mostraron que los factores de riesgos analizados como son las letrinas destapadas, los residuales sin tapar y la no cloración del agua de tomar, son factores asociados a los hábitos de vida y a insuficiente percepción de riesgo entre las familias, factores que pueden ser modificados a través de una efectiva aplicación de programas de promoción de la salud enfocados a modificar estilos de vida y no con un simple propósito de transmisión de información.¹¹

De esta manera, el programa R utilizado demostró las potencialidades de su aplicación en investigaciones relacionados con problemas de salud pública. En particular, el paquete osmar permitió un excelente trabajo de geo localización y mapeo de los problemas epidemiológicos. Por otra parte, el programa contiene además de los análisis estadísticos usuales, otras potencialidades analíticas como son las simulaciones con ecuaciones diferenciales disponibles en el paquete Pomp.¹² En síntesis, el programa libre R permite aplicar una diversidad de modelos analíticos imprescindibles en un estudio con abordaje transdisciplinario,¹³ y fundamental en el trabajo cooperativo entre equipos de investigadores.¹⁴

CONCLUSIÓN

El presente trabajo mostró bondades del programa libre R para lidiar con problemas de salud pública de causalidad multivariada asociados a diversos determinantes. En el presente caso se pudo establecer, gracias a los análisis realizados en R, que hubo insuficiente percepción de riesgos en estas áreas y que es necesario implementar novedosos programas de promoción de salud.

Por otra parte, el uso del software libre R, al ser de libre distribución contribuye a la democratización del conocimiento científico, pues permite a los investigadores llevar a cabo sus estudios sin las restricciones económicas que impone la adquisición de softwares con licencia.¹⁵

REFERENCIAS

1. Torvaldsen S, McIntyre P. Do pertussis notifications reflect incidence or surveillance practices? A comparison of infant notification rates and hospitalization data in NSW. *NSW Public Health Bull.* 2003;14(4):81-4.

2. Team RC. R: A language and environment for statistical

computing. R Foundation for Statistical Computing 2014. Available from: <http://www.R-project.org>.

3. Virasakdi C. epicalc: Epidemiological calculator 2012. Available from: <http://CRAN.R-project.org/package=epicalc>.

4. Eugster MJA, Schlesinger T. osmar: OpenStreetMap and R. *R Journal* [Internet]. 2012. Available from: <http://osmar.r-forge.r-project.org/RJpreprint.pdf>.

5. Salas, Ch. ¿Por qué comprar un programa estadístico si existe R? *Revista Electrónica Ecología Austral.* 2010; 18(2):5. Available from: <http://www.ecologiaaustral.com.ar/files/ea190c83d4.pdf>.

6. Jombarta T, Aanensen DM, Baguelin M, Birrell P, et al. e. OutbreakTools: A new platform for disease outbreak analysis using the R software. *Epidemics.* 2014; 7:28–34.

7. Meyer S, Held L, Hohle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance2014. Available from: <http://surveillance.r-forge.r-project.org/>.

8. Fernandez L. Elevation and cholera: an epidemiological spatial analysis of the cholera epidemic in Harare, Zimbabwe, 2008-2009. *BMC Public Health* [Internet]. 2012; (12). Available from: <http://www.biomedcentral.com/1471-2458/12/442>.

9. Ali M, Lopez AL, You YA, Kim YE, Sah B, Maskerya B, et al. The global burden of cholera. *Bull World Health Organ* 2012;90(2):209–18A.

10. Reveiz L, Chapman E, R-P, P, Koehlmoos T, Cuervo LG et al. Chemoprophylaxis in Contacts of Patients with Cholera: Systematic Review and Meta-Analysis. *PLoS ONE.* 2011;6(11): e27060 doi:10.1371/journal.pone.0027060.

11. Becoña Iglesias E, López Vázquez F, Oblitas Guadalupe L. “Promoción de los estilos de vida saludables”. 2004; (5). Available from: <http://www.alapsa.org/detalle/05/index.htm>.

12. King AA, Ionides EL, Bret'ò CM, S. P. Ellner, B. E. Kendall, Wearing H, et al. pomp: Statistical inference for partially observed Markov processes 2010. Available from: <http://pomp.r-forge.r-project.org>.

13. Betancourt Bethencourt JA. Modelo transdisciplinario para la investigación en Salud Pública. *Revista Panamericana de Salud Pública.* 2014;34(5):359–63.

14. Betancourt Bethencourt JA, Mirabal M, Acao L. Hacia la investigación transdisciplinaria mediante el aprendizaje cooperativo. *Revista Electrónica Educare, [Internet].* 2014; 18(1):[1-15 pp.]. Available from: <http://www.revistas.una.ac.cr/index.php/EDUCARE/issue/current>.

15. Bolaños Guerrero, F. R Project: Su aplicación como software libre para análisis en componentes principales. *Actualidades Investigativas en Educación.* (2011);Vol 11. Número especial.