MANIPULACIÓN, ANÁLISIS Y VISUALIZACIÓN DE DATOS DE LA ENCUESTA DEMOGRÁFICA Y DE SALUD FAMILIAR CON EL PROGRAMA R

Akram Hernández-Vásquez[1,a,b], Horacio Chacón-Torrico[2,a]

RESUMEN

La Encuesta Demográfica de Salud Familiar (ENDES) es una encuesta nacional de base poblacional con representatividad a nivel departamental y área de residencia, constituyéndose en una fuente de información del estado de salud de la población peruana. Con el objetivo de estandarizar su procesamiento y posterior reutilización por parte de la comunidad académica y otros actores interesados; documentamos el código para la manipulación, análisis y visualización de datos del cuestionario de salud de la ENDES 2017, mediante un ejemplo sobre prevalencia de hipertensión arterial y obesidad, utilizando el entorno y lenguaje de programación estadístico R. Se presenta y detalla secuencialmente el código en R, así como, el sustento teórico de la estructura de la encuesta para la manipulación de las bases de datos, considerando que la compleja estructura de la ENDES podría ser una potencial barrera que enfrentan los investigadores. Finalmente, este ejemplo puede servir de base para que se generen mayores estudios basados en la ENDES que sean relevantes para la toma de decisiones en salud pública.

Palabras clave: Ciencia de Datos; Encuestas Epidemiológicas; Estadística como asunto; Hipertensión; Obesidad; Perú (fuente: DeCS BIREME).

MANIPULATION, ANALYSIS, AND VISUALIZATION OF DATA FROM THE DEMOGRAPHIC AND FAMILY HEALTH SURVEY WITH THE R PROGRAM

ABSTRACT

The Demographic and Family Health Survey (ENDES, in Spanish) is a national population-based survey with representation at the departmental level and area of residence, constituting a source of information on the health status of the Peruvian population. In order to standardize its processing and subsequent reuse by the academic community and other stakeholders, we documented the code for the manipulation, analysis, and visualization of data from the ENDES 2017 health questionnaire, through an example on the prevalence of hypertension and obesity, using the R statistical programming environment and language. The R code is presented and detailed sequentially, as well as the theoretical support of the survey structure for the manipulation of databases, considering that the complex structure of the ENDES could be a potential barrier faced by researchers. Finally, this example can serve as a basis for generating further studies based on the ENDES that are relevant to public health decision-making.

Keywords: Data Science; Epidemiological Surveys; Statistics as a matter; Hypertension; Obesity; Peru (source: MeSH NLM).

INTRODUCCIÓN

Las encuestas de base poblacional son una de las fuentes más importantes de información para el desarrollo, seguimiento y evaluación de políticas y programas de salud [1]. Estas encuestas recogen datos sobre la salud materna e infantil, uso de servicios de salud, enfermedades crónicas no transmisibles, nutrición y otros aspectos de salud de un país [2]. Un ejemplo de este tipo de encuestas son las Encuestas Demográficas y de Salud o *Demographic and Health Surveys* (DHS, por sus siglas en inglés).

Una clave para el desarrollo y ejecución de las DHS en el mundo ha sido el impulso proporcionado por el DHS Program (*Demographic and Health Surveys Program*), el cual proporciona asistencia a instituciones gubernamentales para la implementación de las encuestas en diversos países [3]. Es así, que mediante este programa, desde el 1984 hasta el 2010, se diseñaron y ejecutaron 284 encuestas en 84 países, que fueron la base para el desarrollo de 1117 publicaciones en revistas indexadas [4].

En Perú, el Instituto Nacional de Estadística e Informática (INEI) ejecuta desde 1986, la Encuesta Demográfica y de

[1] Universidad San Ignacio de Loyola, Vicerrectorado de Investigación, Centro de Excelencia en Investigaciones Económicas y Sociales en Salud. Lima, Perú.
[2] Universidad Peruana Cayetano Heredia, Facultad de Salud Pública y Administración. Lima, Perú.
[a] Médico cirujano; [b] magíster en Gestión y Políticas Públicas.

Los códigos incluidos en el artículo están disponibles en el repositorio GitHub: https://github.com/horaciochacon/Analisis-Endes-Peru y OSF: https://osf.io/b36cs/

Recibido: 16/11/2018 Aprobado: 06/03/2019 En línea: 15/03/2019

Citar como: Hernández-Vásquez A, Chacón-Torrico H. Manipulación, análisis y visualización de datos de la Encuesta Demográfica y de Salud Familiar con el programa R. Rev Peru Med Exp Salud Publica. 2019;36(1):128-33.doi:10.17843/rpmesp.2019.361.4062.

Salud Familiar (ENDES) bajo el modelo y metodología del DHS Program. Desde el 2004, la ENDES se realiza anualmente [5] y provee una gran cantidad de información en materia de salud pública [6]. Es por esta razón que, en nuestro país el uso de la ENDES como fuente de datos secundaria para estudios específicos es cada vez más frecuente [7]. Más aún cuando la ENDES es actualmente utilizada para realizar un seguimiento sistemático del cumplimiento de los Objetivos de Desarrollo Sostenibles (ODS) [8] y los Indicadores de Resultados de los Programas Presupuestales (PPR) que el país utiliza desde el 2007.

La información y procesamiento de datos debe seguir ciertos criterios, sobre todo en el marco de la publicación científica. Así, los criterios FAIR definen que el tratamiento de los datos y su publicación deben incluir un conjunto de principios rectores para hacer que los datos sean localizables, accesibles, interoperables y reusables, tanto por los sistemas informáticos como por los investigadores [9]. Todos estos criterios se pueden lograr utilizando diversas herramientas, como repositorios de datos y programas de manejo de datos de gran versatilidad. Entre estos está R (https://www.r-project.org/), un entorno y lenguaje de programación estadístico libre, gratuito y con una amplia base de usuarios a nivel mundial en áreas tan diversas que van desde la meteorología hasta la salud pública. Durante los últimos años, su popularización en todos los campos de la ciencia ha aumentado, tanto así que ya se está convirtiendo en el programa de facto para la ciencia de datos [10].

Si bien las bases de datos (BD) de los módulos de la ENDES hasta el 2017 están alojadas en la página web del INEI (http://iinei.inei.gob.pe/microdatos/), usualmente su unión, transformación de variables y análisis suele ser complejo por la estructura propia de la encuesta, las diferentes unidades de análisis (hogares, niños, adultos, y adultos mayores), y «filtros» que deben emplearse para la selección de casos. Asimismo, la ENDES también puede ser descargada de la página web del DHS Program (https://dhsprogram.com/Data/) previo registro como usuario para fines de investigación [11]; sin embargo, las bases de datos más actuales corresponden al 2012. En tal sentido, Vanderelst y Speybroeck demostraron cómo el análisis de una DHS, a pesar de su aparente complejidad, podía ser realizado y replicado en R [12]. No obstante, con la implementación del cuestionario de salud en la ENDES, no se dispone de un manual o guía que permita su correcto análisis.

Siendo así, resulta indispensable documentar el proceso por el cual se obtienen los resultados oficiales de la ENDES. En tal sentido, con el objetivo de estandarizar su procesamiento y posterior reutilización por parte de la comunidad académica y otros actores interesados; documentamos el código para la manipulación, análisis y visualización de datos del cuestionario de salud de la ENDES 2017, mediante un ejemplo sobre prevalencia de hipertensión arterial y obesidad, utilizando el entorno y lenguaje de programación estadístico R.

OBTENCIÓN, TRANSFORMACIÓN Y UNIÓN DE LAS BASES DE DATOS

DESCARGA DE LAS BASES DE DATOS

Las BD de la ENDES son de libre acceso y pueden ser obtenidas de forma manual del portal web del INEI (http://iinei.inei.gob.pe/microdatos/) bajo el formato SPSS («*.sav») o dBase («*.dbf»). Adicionalmente, se ha creado un paquete alojado en la plataforma de desarrollo colaborativo GitHub denominado ENDES.PE (https://github.com/horaciochacon/ENDES.PE) que permite descargar y cargar automáticamente las BD mediante un solo comando desde R para los años que se especifiquen. Además, ya que este repositorio de código libre es colaborativo, cualquier usuario puede modificar y mejorar nuestro paquete para su uso específico. Independientemente del método para obtener las BD, para analizar las cifras de obesidad e hipertensión a partir del «Cuestionario de salud» (CSALUD01), es necesario contar además con las BD: RECH0, RECH1, RECH23, REC42 y RE223123. Los nombres de cada BD se presentan en la Tabla 1 y se utiliza la ENDES 2017 en formato SPSS para este ejemplo.

IDENTIFICACIÓN DE LAS BASES DE DATOS

Las variables de análisis necesarias se encuentran consignadas principalmente dentro de la BD CSALUD01. Esta BD contiene los datos de personas de 15 o más años sobre diabetes, hipertensión, factores de riesgo de enfermedades no transmisibles, tuberculosis y salud mental, entre otras. No obstante, las BD de hogar, personas y mujeres son necesarias para unificar, completar, y filtrar la BD CSALUD01. Los nombres y códigos de las BD, el contenido y el uso que se le da a cada variable para el presente ejemplo se pueden observar en la Tabla 1.

CARGA Y UNIÓN DE LAS BASES DE DATOS

Los identificadores o llaves de cada una de las BD representan diversas estructuras relacionales y jerárquicas dependiendo de la base. Por ejemplo, la variable HHID es el identificador único del hogar, el cual está formado por el número del conglomerado y número de vivienda. Mediante dicho identificador, es posible relacionar los hogares o los individuos con sus hogares. Por otro lado, para relacionar los individuos se utiliza el CASEID que se forma por la concatenación del identificador único del hogar (HHID) y el número de orden del individuo en el hogar (QSNUMERO, HVIDX o HA0 según la BD). Es importante conocer estas precisiones para entender los pasos que se toman durante la unión de las bases, teniendo en cuenta que en algunas

Tabla 1. Variables incluidas en las bases de datos de la ENDES seleccionadas para procesamiento del cuestionario de salud

Utilización / Variable	Contenido	Base de datos (código de la base)
Variables de unión		
HHID	Identificador del hogar	Hogar (RECH0), Personas (RECH1), Vivienda (RECH23), Cuestionario Salud (CSALUD01), Mujer – Antropometría (RECH5)
HVIDX	Número de individuo por hogar	Personas (RECH1)
QSNUMERO	Número de individuo por hogar	Cuestionario Salud (CSALUD01)
HA0	Número de individuo por hogar	Mujer – Antropometría (RECH5)
CASEID	Identificador de persona	Mujer - Salud y Lactancia (REC42), Mujer - Historia Obstétrica (RE223123)
Variables de filtro		
V213	Actualmente embarazada	Mujer - Historia Obstétrica (RE223123)
QSRESINF	Resultado informante	Cuestionario Salud (CSALUD01)
Variables de diseño		
HV001	Número del conglomerado	Hogar (RECH0)
HV022	Estrato	Hogar (RECH0)
PESO15_AMAS*	Factor de ponderación	Hogar (RECH0)
Variables de análisis		
HV104	Sexo	Personas (RECH1)
HV270	Quintil de bienestar	Vivienda (RECH23)
SHREGION	Región natural	Vivienda (RECH23)
HV025	Urbano/Rural	Hogar (RECH0)
QS903 - QS905	Medición de presión arterial	Cuestionario Salud (CSALUD01)
HA2	Peso	Mujer – Antropometría (RECH5)
QS900	Peso	Cuestionario Salud (CSALUD01)
HA3	Talla	Mujer – Antropometría (RECH5)
QS901	Talla	Cuestionario Salud (CSALUD01)

*Para incorporarla en el análisis se debe dividir entre 1 000 000

BD sólo se dispone del HHID y el número de orden del individuo en el hogar (variables que en su conjunto se denominan «llave de unión»).

El proceso de preparación de las BD desde que cargamos los archivos hasta la obtención de la base final lista para el análisis puede observarse en el Script 1 (Anexo 1- visualizar en versión electrónica). El primer paso es la carga de los paquetes necesarios y luego la lectura de las BD que están en el formato «*.*sav*» (ver líneas 8 - 20 del Script 1). R por defecto en su configuración base no permite leer archivos «*.*sav*»; no obstante, la posibilidad de instalar paquetes adicionales (*haven, tidyverse, survey)* para extender sus capacidades permite ejecutar esta y otras tareas.

Una vez cargadas las bases en R, es necesario transformar algunas variables para poder realizar la posterior unión (ver líneas 22 - 34 del Script 1). Estas modificaciones se sustentan en la estructura de los identificadores mencionados en párrafos anteriores. Posteriormente, la unión de las BD sigue un procedimiento secuencial empezando por la unión de las bases con información de individuos y terminando la unión con las bases del hogar y vivienda (ver líneas 36 - 43 del Script 1). Las variables de unión utilizadas están descritas en la Tabla 1.

Esta base preliminar obtenida de la unión de diferentes BD de ENDES es filtrada para eliminar los registros de gestantes y a los individuos que no respondieron completamente la encuesta (ver líneas 45 - 48 del Script 1). Es importante mencionar que el número de observaciones (n = 32 514) después de este paso es el mismo que el INEI ha publicado en su reporte de indicadores para la ENDES del año 2017 [13].

DEFINICIÓN DE LAS VARIABLES DE ANÁLISIS

Las variables de interés para el presente análisis fueron hipertensión arterial y obesidad. La definición de hipertensión arterial utilizada es el hallazgo de una presión arterial media sistólica mayor o igual a 140 mmHg o el hallazgo de una presión arterial media diastólica mayor a 90 mmHg según el *Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure* (JNC 7) [14]. El cálculo de las variables de hipertensión arterial y obesidad puede observarse entre las líneas 55 y 69 del Script 1. El resto de variables

utilizadas para el análisis están incluidas en las bases unificadas, no obstante, creamos una BD final donde se seleccionaron algunas variables de interés (ver líneas 104 - 109 del Script 1). Adicionalmente, en el Anexo 1 se presenta el Script 1 extendido donde se incluye el cálculo del consumo de frutas y verduras (ver líneas 71 - 102) según lo estimado por el INEI.

DISEÑO MUESTRAL DE LA ENDES

La BD final, que hemos llamado ENDES en el Script 1, contiene todas las variables necesarias para el análisis de la prevalencia de hipertensión arterial y obesidad. No obstante, realizar el análisis directamente sobre esta sería incorrecto. Cabe precisar que, la ENDES es una encuesta con diseño muestral complejo con inferencia hasta el nivel departamental y con factores de ponderación que permite recomponer la estructura de la población de referencia. Sin estas especificaciones dentro del proceso de análisis no se obtendrían una adecuada estimación de los indicadores. Por tal motivo, se debe incluir el diseño muestral de la ENDES con el comando *survey()* utilizando el conglomerado, el estrato y el factor de ponderación que previamente debe ser dividido entre un millón para un correcto análisis (ver líneas 111 a 113 del Script 1 en el Anexo 1).

ANÁLISIS Y VISUALIZACIÓN DE LOS DATOS

Al tener la base lista para el análisis y el diseño configurado para las ponderaciones respectivas, se puede realizar una descripción de las variables de interés (hipertensión y obesidad). Estas variables fueron estratificadas según sexo, quintil de bienestar, área residencial y región natural. Las tablas resultantes de las proporciones estimadas junto con sus intervalos de confianza al 95%, pueden generarse mediante los comandos de las líneas 7 a la 24 del Script 2 (Anexo 2 - visualizar en versión electrónica).

GRÁFICA DE LOS DATOS

En la Figura 1, utilizando el paquete *ggplot2,* se grafica en una cuadrícula la prevalencia de hipertensión arterial y obesidad según el quintil de bienestar y sexo junto con sus intervalos de confianza al 95%, siendo estos datos

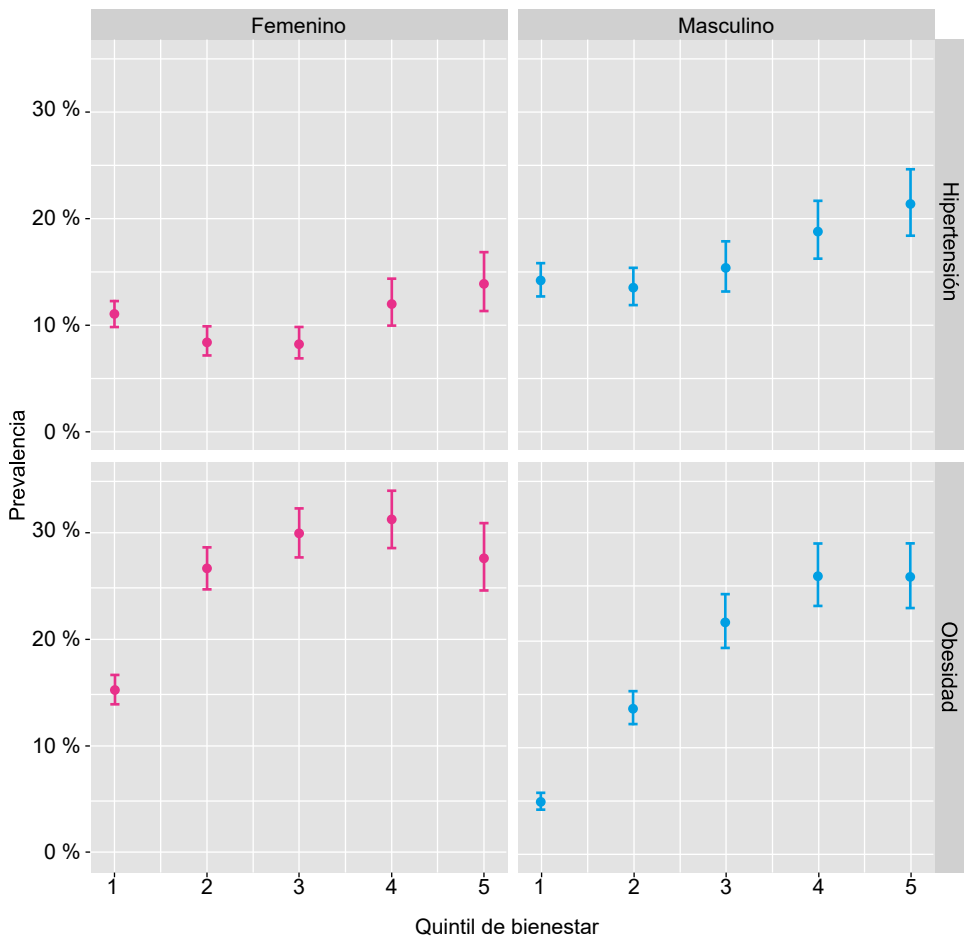

Figura 1. Hipertensión arterial y obesidad según sexo y quintil de bienestar, ENDES 2017

ponderados según el diseño muestral de la encuesta. Los resultados muestran una mayor prevalencia de obesidad e hipertensión en quintiles superiores en comparación a los quintiles inferiores, así como, mayores cifras de HTA en hombres y de obesidad en mujeres. Cabe precisar que para las presentes estimaciones todos los coeficientes de variación fueron menores a 10 (ver líneas 42 a 50 del Script 2 en el Anexo 2).

CONSIDERACIONES FINALES

La utilización de la ENDES como fuente secundaria de información para responder a ciertas preguntas de investigación ha sido realizada por varios autores [7]. No obstante, los procedimientos de tratamiento y análisis de los datos pocas veces están disponibles a los editores, revisores y lectores. Dadas las características únicas de diseño y estructura, el escrutinio del análisis de este tipo de encuestas debería ser un paso importante en las revisiones por pares para garantizar la transparencia, reproducibilidad e integridad científica [15], más aun tratándose de bases de datos de acceso libre.

La cantidad de información y datos disponibles es cada vez mayor. Teniendo en cuenta que el Perú ya se comprometió con una política de «Datos Abiertos» [16], no podemos sino esperar una creciente disponibilidad de fuentes de datos secundarios tanto en salud como en otros campos que cruzan transversalmente a la salud pública y que probablemente serán fuente de futuras investigaciones. Consideramos que una opción para garantizar la integridad de la investigación de fuentes de datos secundaría está en compartir los procedimientos de manejo y análisis de los datos. De tal forma, además de garantizar la calidad del análisis, se promueve un entorno de colaboración entre los investigadores y actores interesados.

Siendo ENDES una fuente muy importante de información en salud, creemos que su compleja estructura es una potencial barrera que enfrentan los investigadores y que evita que se generen más estudios con la información contenida en esta encuesta de base poblacional. Así, la manipulación, análisis y visualización de datos con la ENDES ejemplificada en el presente artículo, ha sido realizada con un entorno y lenguaje de programación estadístico libre bajo la licencia *GNU general public license*, con datos abiertos y una metodología reproducible en todos sus niveles. En tal sentido, mostramos cómo con unas pocas líneas de código se puede obtener información de la población peruana relevante para la toma de decisiones en salud pública y que puede facilitar el acceso de investigadores, funcionarios públicos y población en general libre de reproducir, modificar y mejorar el mismo procedimiento de análisis de la información contenida en la ENDES.

Agradecimientos: A Mixsi Casas Bendezú, Licenciada en Estadística del Instituto Nacional de Estadística e Informática, por la revisión del manuscrito y validación de los scripts.

Contribuciones de autoría: AHV tuvo la idea del artículo, diseñó el estudio, recopiló y procesó los datos. HCT elaboró la primera versión de los scripts. AHV y HCT participaron en la revisión y validación de los scripts, análisis de los datos, interpretación de los resultados, redacción del manuscrito, y aprobación de la versión final.

Fuentes de financiamiento: autofinanciado.

Conflictos de interés: los autores declaran no tener conflictos de interés.

Referencias Bibliográficas

1. Velásquez Hurtado JE, Rivera Sivirichi RA. Encuestas en salud: instrumentos esenciales en el seguimiento y evaluación de los programas presupuestales. Rev Peru Med Exp Salud Publica. 2017;34(3):512-520. doi: 10.17843/rpmesp.2017.343.3031.
2. Dandona R, Pandey A, Dandona L. A review of national health surveys in India. Bull World Health Organ. 2016;94(4):286–96A. doi: 10.2471/BLT.15.158493.
3. The DHS Program. Measure DHS [Internet]. [citado el 11 de octubre de 2018]. Disponible en: https://dhsprogram.com/
4. Short Fabic M, Choi Y, Bird S. A systematic review of Demographic and Health Surveys: data availability and utilization for research. Bull World Health Organ. 2012;90(8):604–12. doi: 10.2471/BLT.11.095513.
5. Instituto Nacional de Estadística e Informática. ENDES - Historia [Internet]. INEI; 2018 [citado el 27 de octubre de 2018]. Disponible en: https://proyectos.inei.gob.pe/endes/
6. Instituto Nacional de Estadística e Informática. Encuesta Demográfica y de Salud Familiar [Internet]. Disponible en: https://proyectos.inei.gob.pe/endes/
7. Ruiz-Maza JC, Pezo-Pezo AM, Soto-Azpilcueta RA. Producción científica en base a cinco encuestas nacionales de Perú. Rev Peru Med Exp Salud Publica. 2018;35(1):166–7. doi: 10.17843/rpmesp.2018.351.3554.
8. Instituto Nacional de Estadística e Informática. Perú: Línea de base de los principales indicadores disponibles de los Objetivos de Desarrollo Sostenible (ODS) [Internet]. INEI; 2018 [citado el 02 de marzo de 2019]. Disponible en: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1578/libro.pdf
9. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: https://doi.org/10.1038/sdata.2016.18
10. Goztepe K. De Facto Language of Data Science: The R Project. Journal of Military and Information Science. 2017;104–107. doi: 10.17858/jmisci.288183.

11. Padilla TA. Relevancia y perspectiva para el desarrollo de los sistemas de información en población y salud sexual y reproductiva en el Perú. Rev Peru Med Exp Salud Publica. 2007;24(1):67–80.
12. Vanderelst D, Speybroeck N. Loading, merging and analysing demographic and health surveys using R. Int J Public Health. 2014;59(2):415–22. doi: 10.1007/s00038-013-0538-2.
13. Instituto Nacional de Estadística e Informática. Perú: Enfermedades no transmisibles y transmisibles, 2017 [Internet]. INEI; 2017 [citado el 27 de octubre de 2018]. Disponible en: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1526/libro.pdf
14. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, *et al.* The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. JAMA. 2003;289(19):2560–72. doi: 10.1001/jama.289.19.2560
15. Resnik DB, Shamoo AE. Reproducibility and Research Integrity. Account Res. 2017;24(2):116–23. doi: 10.1080/08989621.2016.1257387.
16. Presidencia del Consejo de Ministros. Estrategia nacional de datos abiertos gubernamentales del Perú (2017-2021) [Internet]. Lima: Secretaría de Gestión Pública; 2017 [citado el 27 de octubre de 2018]. Disponible en: https://www.peru.gob.pe/estrategia.pdf

Correspondencia: *Akram Abdul Hernández Vásquez*
Dirección: Universidad San Ignacio de Loyola, Av. La Fontana 550, La Molina, Lima, Perú
Teléfono: (00511) 317-1000
Correo electrónico: ahernandez@usil.edu.pe