Ciencia de Datos en Salud: Aplicaciones en CDC Perú

Andree Valle Campos, Candidato a Magíster en Ciencias de la Investigación Epidemiológica. Grupo de '

2020-01-08

La Ciencia de Datos en Salud es el trabajo científico realizado para mejorar la colecta, manejo y análisis de datos de individuos o poblaciones, con el objetivo de mejorar su salud[1]. Este integra tres componentes: tecnologías de la información, métodos bioestadísticos y epidemiológicos, y un marco conceptual basado en salud pública e inferencia causal. En CDC Perú, hay una creciente necesidad por tomar decisiones basadas en el análisis de grandes bases de datos existentes o la adquisición rápida de nueva información de campo. Esto ha demandado que los epidemiólogos empecemos a usar activamente estas nuevas formas de trabajo. En este sentido, aquí presentamos las características de tres herramientas de Software Libre[2] recientemente implementadas en la institución: KoBo Toolbox para la colecta de datos, R para la limpieza de bases, y paquetes del R Epidemics Consortium (RECON) para el análisis cuantitativo en respuesta a brotes y epidemias.

Primero, en situaciones de emergencia, el tiempo entre el diseño y la aplicación de una encuesta es crítico para evaluar hipótesis e implementar medidas de control. KoBo Toolbox[3] es una plataforma web que reduce este tiempo mediante el acceso a plantillas y medios rápidos de distribución. En comparación a otros softwares, su principal ventaja radica en el acceso a colecciones públicas de instrumentos compartidos por otras instituciones, acelerando la actividad científica bajo principios de Ciencia Abierta[2]. También permite importar encuestas diseñadas en archivos Excel con formato XLSform[4], modificarlas en línea, usarlas en web o aplicativo móvil (KoBoCollect)[5] para colectar información sin conexión a internet. En nuestra primera experiencia solo la empleamos como herramienta de digitación de encuestas llenadas a lápiz y papel. Sin embargo, el formato XLSform permitió la edición colaborativa vía Google Sheets reduciendo el tiempo de diseño y distribución. Del mismo modo, la base de datos exportada conservó tipos de variable y etiquetas facilitando su análisis.

Posterior a ello, todo proceso realizado debe ser reproducible, es decir estar escrito en código informático capaz de recrear todos los resultados a partir de los datos originales en cualquier momento[6]. Empleamos el software de programación estadística R[7] para escribir nuestros procedimientos. En una sola plataforma permite importar, limpiar, analizar y generar reportes en formato de artículos[8], presentaciones[9] o pósters[10]. Su disponibilidad de herramientas para la limpieza y resumen de bases[11] ha permitido integrar múltiples fuentes de información de diferentes instituciones con estructuras totalmente heterogéneas. Dos ejemplos son la generación de la lista de Enfermedades Raras o Huérfanas, y la lista de enfermedades priorizadas para el Plan Esencial de Aseguramiento en Salud. Ambos podrán reproducirse en cualquier momento para evaluar sus metodologías o actualizarlos en los próximos años ante la disponibilidad de nuevas bases de datos.

Por otra parte, los análisis más especializados requieren de paquetes de R, los cuales son un conjunto de funciones que extienden la caja de herramientas que el software posee por defecto. RECON[12] es la agrupación que crea, integra y evalúa paquetes específicos para el análisis de brotes, emergencias en salud y crisis humanitarias. Por ejemplo, los paquetes *incidence* y *projections* modelan y proyectan incidencias en periodos cortos de tiempo, respectivamente[13]. Su amplia disponibilidad de tutoriales[14] facilitó su rápida implementación en respuesta a la epidemia del Síndrome Guillain Barré del presente año[15]. Además, recientes métodos de análisis estadísticos a escala espacial[16] permitieron generar mapas con riesgos relativos estimados a nivel distrital y evaluar su asociación con enfermedad diarreica aguda[17]. De la misma forma, se podrán adaptar otros recursos como la generación estandarizada y automatizada de Salas Situacionales empleando el paquete *sitrep* de la organización internacional R4epis[18].

En resumen, el CDC Perú ha implementado herramientas de Ciencias de Datos en Salud que han acelerado

la colecta, análisis y reporte en respuesta a brotes, epidemias, situaciones de emergencia e investigaciones epidemiológicas. Finalmente, experiencias similares en otros campos han demostrado que esta transición ha favorecido a la transparencia, el ahorro de esfuerzos y la estandarización de procesos al largo plazo, acortando el tiempo para la toma de decisiones[2]. Por ello, los próximos pasos son capacitar y estimular el uso de estas herramientas en la institución y en las distintas Direcciones Regionales.

El presente editorial ha sido generado usando Rmarkdown[19,20] y está disponible en GitHub[21].

Referencias

1. Dammann O, Smart B. Health Data Science. In: *Causation in Population Health Informatics and Data Science*. Springer; 2019:15-26. https://link.springer.com/chapter/10.1007/978-3-319-96307-5_2.

2. Lowndes JSS, Best BD, Scarborough C, et al. Our path to better science in less time using open data science tools. *Nature ecology & evolution*. 2017;1(6):0160. https://www.nature.com/articles/s41559-017-0160.

3. Pham P, Dorey A, Musaraj P, Patrick-Vinck FE, Kreutzer T, others. KoBoToolbox at the Harvard Humanitarian Initiative 14 story st. *Second floor, Cambridge, MA*. 2018;2138. https://www.kobotoolbox.org/.

4. XLSform community. XLSform. In: Online website. Accessed: 27-december-2019. https://xlsform.org/en/.

5. KoBo Toolkit. KoBoCollect. In: Online website. Accessed: 27-december-2019. https://play.google.com/store/apps/details?id=org.koboc.collect.android&hl=en.

6. Rodriguez-Sanchez F, Pérez-Luque AJ, Bartomeus I, Varela S. Ciencia reproducible: Qué, por qué, cómo. *Revista Ecosistemas*. 2016;25(2):83-92. https://www.revistaecosistemas.net/index.php/ecosistemas/article/view/1178.

7. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/.

8. Allaire J, Xie Y, R Foundation, et al. *Rticles: Article Formats for R Markdown*.; 2019. https://CRAN.R-project.org/package=rticles.

9. Xie Y. *Xaringan: Presentation Ninja*.; 2019. https://CRAN.R-project.org/package=xaringan.

10. Thorne WB. *Posterdown: An R Package Built to Generate Reproducible Conference Posters for the Academic and Professional World Where Powerpoint and Pages Just Won't Cut It*.; 2019. https://github.com/brentthorne/posterdown.

11. Grolemund G, Wickham H. R for Data Science. 2018. https://r4ds.had.co.nz/.

12. R Epidemic Consortium. R Epidemic Consortium. In: Online website. Accessed: 27-december-2019. https://www.repidemicsconsortium.org/.

13. Polonsky JA, Baidjoe A, Kamvar ZN, et al. Outbreak analytics: A developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B*. 2019;374(1776):20180276. https://royalsocietypublishing.org/doi/10.1098/rstb.2018.0276.

14. R Epidemic Consortium. Recon Learn. Free and open training resources to respond to outbreaks, health emergencies and humanitarian crises. In: Online website. Accessed: 27-december-2019. https://www.reconlearn.org/.

15. Munayco Escate C. Análisis de la etapa inicial de la epidemia de Síndrome de Guillain-Barré en el Perú. In: Boletín Epidemiológico del Perú. Volumen 28. Accessed: 27-december-2019. www.dge.gob.pe/portal/docs/vigilancia/boletines/2019/28.pdf.

16. Moraga P. *Geospatial Health Data: Modeling and Visualization with R-Inla and Shiny*. CRC Press; 2019. http://www.paulamoraga.com/book-geospatial/.

17. Reyes Vega M, Soto Cabezas G, Valle Campos A, et al. Análisis epidemiológico de la epidemia del Síndrome de Guillain-Barré en Perú, 2019. In: XIII Congreso Científico Internacional Del Instituto Nacional De Salud. Accessed: 27-december-2019. http://bit.ly/cdcperuins19.

18. R Epidemic Consortium and Doctors Without Borders. R4epis. In: Online website. Accessed: 27-december-2019. https://r4epis.netlify.com/.

19. Xie Y, Allaire J, Grolemund G. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC; 2018. https://bookdown.org/yihui/rmarkdown.

20. Allaire J, Xie Y, McPherson J, et al. *Rmarkdown: Dynamic Documents for R.*; 2019. https://github.com/rstudio/rmarkdown.

21. Valle Campos A. Ciencia de Datos en Salud: Aplicaciones en CDC Perú. In: Github. Accessed: 27-december-2019. http://bit.ly/cdcdatasci19.