Ciencia de datos con R con impacto en salud pública. Una experiencia de uso de *tidyverse* para la detección de embarazos.

Abstract
La implementación de una historia clínica electrónica en el sistema de salud pública de la Ciudad de Buenos Aires derivó en una base de un gran volumen de datos. No obstante, al ser una base transaccional cuyo objetivo es mejorar la atención de pacientes, el uso secundario de análisis de datos presenta un desafío por el tipo de datos registrados (semi-estructurados, incompletos, subjetivos). Al día de la fecha la historia clínica no cuenta con un módulo destinado a identificar y caracterizar embarazos. La detección de embarazos es de gran relevancia ya que facilita el acceso a derechos de la salud sexual y reproductiva, y permite destinar fondos para ello. Durante el 2021, integrantes del equipo de ciencia de datos de la Gerencia Operativa de Gestión de Información y Estadísticas en Salud del Ministerio de Salud de la Ciudad Autónoma de Buenos Aires ejecutaron un proyecto de mejora del proceso existente de detección de embarazos y de consolidación de una base específica con actualización periódica. Este proyecto fue llevado a cabo por un equipo interdisciplinario y desarrollado en el lenguaje R utilizando principalmente la librería *tidyverse* para la exploración, manipulación y visualización de datos masivos. El primer paso del proceso resultó novedoso ya que consistió en utilizar una nueva fuente de información, el sistema de admisión, pases y egresos sobre internaciones hospitalarias. Así, se integró esta información sobre eventos obstétricos a los registros obtenidos a partir de la historia clínica -mediante el uso de expresiones regulares de los campos de texto estructurado y texto libre-. Los siguientes pasos consistieron en la extracción y el cálculo de variables, la clasificación de cada registro, la discriminación entre embarazos de una misma persona y la caracterización de cada uno.

El uso de *tidyverse* fue central para el éxito del proyecto, que se encuentra en una etapa de cierre y evaluación.

Keywords: Expresiones regulares, Salud/Salud pública, Historia clínica electrónica, ADT, Tidyverse, Software R.

Este trabajo repone una experiencia de uso de R[1] en producción con impacto en políticas públicas de acceso a la salud por parte de un equipo de ciencia de datos del Gobierno de la Ciudad de Buenos Aires (GCBA).

A partir del año 2016 el GCBA comenzó a implementar una historia clínica electrónica (Historia Integral de Salud, HIS) en los centros de salud del primer nivel de atención. La HIS consiste en una base transaccional de las atenciones que se dan de manera ambulatoria y cuyo principal objetivo es mejorar la atención de los pacientes de manera de constituir una verdadera red de atención de salud. A pesar de la ventaja de contar con un gran volumen de información para su análisis, el uso secundario de los datos presenta ciertas dificultades ya que el registro es incompleto, subjetivo, poco sistemático y semi-estructurado[2].

Dentro de este gran volumen de información los embarazos tienen gran relevancia, ya que su detección temprana permite facilitar la accesibilidad a derechos de la salud sexual y reproductiva, como por ejemplo controles de embarazo oportunos. La detección de este tipo de eventos también reviste de importancia económica en tanto son objeto de políticas públicas de financiamiento.

Si bien está previsto su desarrollo a mediano plazo, al día de la fecha la HIS no cuenta con un módulo vinculado a los embarazos que permita facilitar la identificación y extracción masiva de información vinculada a estos eventos. Es por ello que durante el 2021, dentro de la Gerencia Operativa de Gestión de Información y Estadísticas en Salud (GOGIES) del Ministerio de Salud de la Ciudad Autónoma de Buenos Aires se destinaron recursos humanos a la mejora

del proceso existente de detección de embarazos y a la consolidación de una base específica con actualización periódica.

Este proyecto fue llevado a cabo por un equipo de ciencia de datos interdisciplinario y desarrollado en el lenguaje R mediante RStudio Server[3], que se volvió clave en el contexto de teletrabajo impuesto durante la pandemia.

Para este desarrollo se utilizaron múltiples paquetes R. Parte de la colección de paquetes tidyverse[4] (*lubridate*, *dplyr*, *tidyr*, *stringr* y *ggplot2*) para el proceso de *data wrangling* y visualización de datos; *properties*, *dbplyr* y *odbc* para gestionar las conexiones con bases de datos SQL; y *agiseR*, librería desarrollada internamente por el equipo de Ciencia de Datos de GOGIES, para extraer información del *Data Warehouse* de la gerencia.

La decisión de dar inicio al proyecto respondió a diversos factores técnicos y sanitarios.

Respecto de los primeros, se destacan la búsqueda de reproducibilidad y de optimización de recursos dado que esta temática es transversal a varios proyectos, la reciente disponibilidad de una nueva fuente de información (admisión, pases y egresos, *ADT*), y la necesidad de adecuación del circuito de procesamiento de datos debido a la implementación de una nueva versión de la HIS (Tabla 1).

En cuanto a los factores sanitarios, se destaca la relevancia de estrategias de acompañamiento y búsqueda activa debido al impacto que tuvieron las medidas de aislamiento social en particular durante el comienzo del 2020 sobre el acceso de las personas a los centros de salud, acciones que dependen fuertemente de la identificación de los grupos de interés.

	HIS	**ADT**
Acto que registra	Atención ambulatoria longitudinal (consultas)	Internación: ingreso, pase y egreso
Ámbito	Atención primaria de la salud (APS) y consultorios externos y guardias de algunos hospitales.	Internación Hospitalaria
Tipo de información	Clínica y psicosocial	Administrativa
Tipo de datos recolectados	Semiestructurado (motivos de consulta) y campo de texto libre (evolución)	Estructurado (por ejemplo, tipo de evento y fecha)

Tabla 1. Fuentes de información. Comparación entre las dos fuentes de información utilizadas para el proyecto: módulo de Historia Integral de Salud (HIS) y módulo de admisión, pases y egresos (ADT).

El primer paso del proceso consistió en recopilar todos los registros por embarazo a partir de las dos fuentes de información (Figura 1). Mediante el uso de expresiones regulares[5] se detectaron consultas en la HIS relativas a embarazos que incluyeran información sobre edad gestacional. A partir de este dato se calcularon la Fecha de Última Menstruación (FUM) y la Fecha Probable de Parto (FPP) para todos los embarazos que se encontraran dentro del período de búsqueda. En el sistema ADT se registra la fecha de parto y la edad gestacional a ese momento: a partir de estos datos, calculamos la FUM.

Los siguientes pasos consistieron en clasificar cada registro (primera consulta o control) para poder discriminar cada embarazo (considerando la posibilidad de más de un embarazo por persona dentro del periodo de interés) y finalmente caracterizarlo. Las variables de interés refieren a la delimitación temporal del embarazo, a su finalización efectiva, la edad gestacional al primer registro y al momento de la finalización. Por último, conservamos los embarazos en curso durante el periodo de interés definido al inicio del proceso.

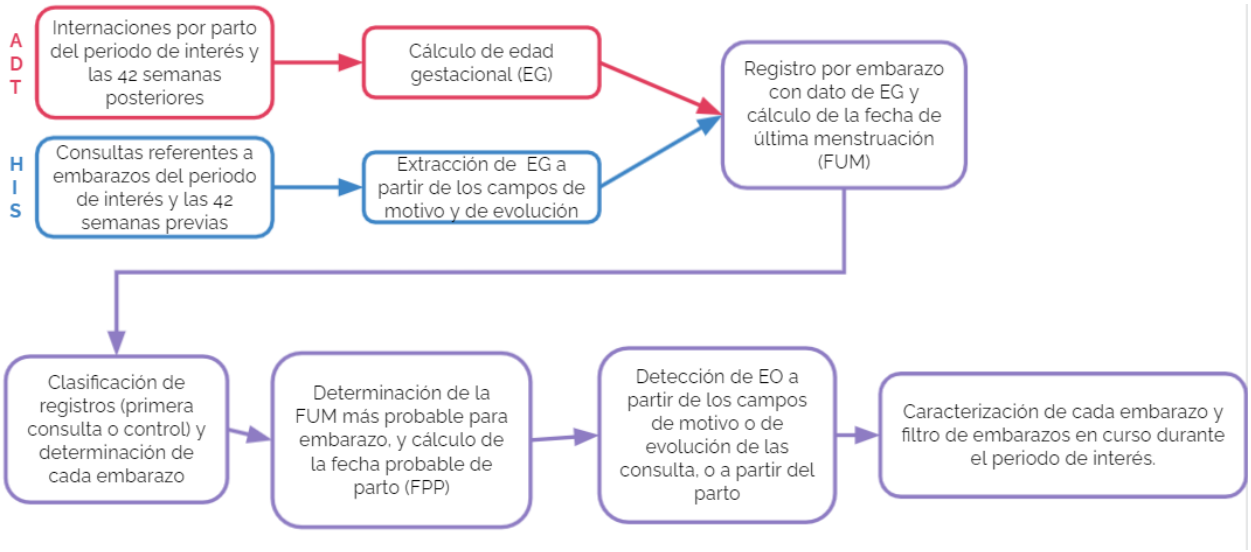

Figura 1. Esquema del proceso de detección de embarazos.

Como resultado se obtuvo un algoritmo que genera una tabla con los embarazos activos durante el período de interés y las variables comúnmente utilizadas en los pedidos de información relativos a la temática. Esto marca un avance importante dado que no es posible contar con esta información de manera directa a partir de los datos transaccionales.

Referencias

(1) R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

(2) Botsis, T., Hartvigsen, G., Chen, F., & Weng, C. (2010). Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010, 1.

(3) RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA www.rstudio.com.

(4) Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.

(5) Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2), 221-230.