CARTA AL EDITOR / LETTER

DOI: https://doi.org/10.20453/rmh.v36i4.5990

Random forest en R: Aplicación práctica para la imputación de datos para la investigación clínica en Hemato-oncología

Random Forest in R: Practical application for data imputation for clinical research in Hemato-Oncology

Rafael Pichardo-Rodriguez[1], *Liz Córdova-Cueva*[1], *Jhony A. De La Cruz-Vargas*[2]

[1] Escuela de Medicina, Universidad César Vallejo, Piura, Perú.
[2] Instituto de Investigaciones en Ciencias Biomédicas (INICIB). Universidad Ricardo Palma, Lima-Perú.

Citar como:
Pichardo-Rodriguez R, Córdova-Cueva L, De La Cruz-Vargas JA. Random forest en R: Aplicación práctica para la imputación de datos para la investigación clínica en Hemato-oncología. Rev Méd Hered. 2025; 36(4): 383-386. DOI: 10.20453/rmh.v36i4.5990

Recibido: 14/11/2024
Aceptado: 21/08/2025

Declaración de conflictos de intereses:
No se declaran conflictos de intereses.

Correspondencia:
Nombre: Rafael Pichardo-Rodríguez
Dirección: Av Brasil 748, Jesus María
Celular: 989370702
✉ rafael_martin1352@hotmail.com

Sr. Editor:

En la investigación clínica en hemato-oncología, especialmente en nuestra región, los datos del mundo real ofrecen información valiosa para la toma de decisiones al aprovechar datos de pacientes en entornos clínicos reales, permitiendo identificar y cerrar brechas en la evidencia proporcionada por los ensayos clínicos tradicionales [1,2]. Sin embargo, la heterogeneidad de las fuentes de datos, la disponibilidad oportuna de datos de alta calidad, y en particular la pérdida de datos, limitan el uso completo de esta información. [1,2]

La ausencia de datos constituye un problema recurrente en investigación clínica, reduciendo el poder estadístico de los hallazgos [3]. En años recientes, el manejo de datos faltantes ha avanzado significativamente, con programas computacionales que ahora ofrecen opciones sofisticadas de imputación [3]. La imputación es una estrategia que permite reemplazar los valores ausentes de una variable por estimaciones plausibles, basadas en distribuciones de probabilidad condicionadas a otras variables observadas del mismo sujeto [4]. Últimamente, los métodos basados en algoritmos de Machine Learning, como los de bosques aleatorios o "Random Forest" (RF), están cobrando gran aceptación por su efectividad en la imputación múltiple de datos clínicos. [3,5]

Se presenta un ejemplo ilustrativo basado en un estudio ficticio que explora la efectividad de Polatuzumab + R-ICE, comparado con R-ICE, en la supervivencia global al primer año en pacientes con linfoma de células B grandes en recaída a partir de una cohorte retrospectiva multicéntrica nacional. La tabla 1 detalla la secuencia de código en R, desde la creación de la base de datos (en este caso simulada), imputación y exportación a formato ".xlsx", facilitando su uso para investigadores que emplean otros programas estadísticos y se anexa un instructivo en la referencia 5 para la importación de datos en el programa R [6]. Para una explicación detallada sobre el uso de R y RStudio se recomienda revisar la sexta referencia del presente manuscrito. [7]

Tabla 1. Codigos en R para la creacion, imputacion y exportación de la base de datos.

	Código	Resultado o salida en el software	Interpretación

Instalar y cargar las librerías necesarias	# Instalar librerias
	install.packages("mice")
	install.packages("missForest")
	install.packages("VIM")
	install.packages("naniar")
	install.packages("openxlsx")
	
	# Cargar librerias
	library(mice)
	library(missForest)
	library(VIM)
	library(naniar)
	library(openxlsx)		

Simulación de base de datos con datos perdidos*	##Base de datos simulada
	# Generar semilla
	set.seed(123) # Para reproducibilidad
	
	# Número de observaciones
	n <- 400
	
	# Generación de la base de datos simulada
	datos <- data.frame(
	Sexo = sample(c("Masculino", "Femenino"), n,
	replace = TRUE),
	Edad = sample(30:80, n, replace = TRUE),
	IPI = sample(c("Bajo", "Intermedio", "Alto"), n,
	replace = TRUE),
	Estadio_clinico = sample(c("I", "II", "III", "IV"), n,
	replace = TRUE),
	ECOG = sample(0:4, n, replace = TRUE),
	LDH = round(rnorm(n, mean = 300, sd = 100), 1),
	Tratamiento = sample(c("Polatuzumab + R-ICE", "RICE"),
	n, replace = TRUE),
	Tiempo_hasta_evento = round(runif(n, min = 1, max
	= 12), 1),
	Muerte = sample(0:1, n, replace = TRUE)
	
	# Introducción de valores faltantes
	# Crear valores faltantes en Edad, LDH, Tiempo_hasta_evento, Estadio_clinico y Sexo
	datos$Edad[sample(1:n, size = 15)] <- NA
	datos$LDH[sample(1:n, size = 25)] <- NA
	datos$Tiempo_hasta_evento[sample(1:n, size = 5)] <-
	NA
	datos$Estadio_clinico[sample(1:n, size = 7)] <- NA
	datos$Sexo[sample(1:n, size = 10)] <- NA
	
	# Mostrar los primeros registros de la base de datos
	simulada
	head(datos)
	
	## Tambien se puede importar una base de datos en
	excel
	library(readxl)
	
	base<-
	read_excel("C:/Users/TU_USUARIO/Documents/mib
	ase.xlsx")
	View(base)		

*Para los que cuentan con una base de datos a imputar de su investigación, la pueden importar al software en lugar de realizar la simulación. Revisar la quinta referencia del presente artículo[6].
**Es importante convertir a las variables cualitativas a factores debido a que el algoritmo del código de Random Forest solo reconoce a las variables cualitativas como factores para proceder con la imputación.
***Se muestra como realizar la imputación de datos con Random Forest con dos paquetes estadísticos de R diferentes, la elección por cada uno de ellos dependerá del investigador.
****En la exp ortación de datos, deben colocar la dirección donde guardaran la base de datos, cuando cop ien la dirección tener en cuenta que viene con "backslash" o \ y p ara que p ueda ser ejecutado en R, deben cambiar
cada una a "slash" o / como en el ejemp lo.
Nota: En el software R solo debe copiarse los códigos de la segunda columna para que puedan ser ejecutados.

Tabla 1. (*Continuación*).

Convertir a factores las variables cualitativas**	#Convertir a factores datos$Sexo<-as.factor(datos$Sexo) datos$Estadio_clinico<- as.factor(datos$Estadio_clinico)		
Evaluación de datos perdidos	#Evaluación de datos perdidos	> mcar_test(datos)	Se genera un gráfico donde se puede visualizar el número y porcentaje de datos perdidos por variables (vis_miss).
	vis_miss(datos)#Composición	# A tibble: 1 × 4 statistic df p.value missing.patterns	La prueba de Little indica la presencia de una distribución completamente al azar de los datos perdidos en la base de datos (**p=0.603**)
	#Evaluación de distribución completamente al azar mcar_test(datos)	*<dbl> <dbl> <dbl> <int>* 1 57.5 61 p=**0.603** 9	
Imputar datos con Random Forest***	# Imputacion de datos	**MICE:** > dataimp <- mice(datos, method = "rf")	**MICE:** Las variables IPI y Tratamiento no fueron imputadas ya que no presentaban valores perdidos o mostraron comportamiento constante en los datos. Esto fue registrado como eventos sin imputación dentro del objeto mids del paquete mice.
	#### Con el Paquete MICE dataimp <- mice(datos, method = "rf") dataimp	iter imp variable 1 1 Sexo Edad Estadio_clinico LDH Tiempo_hasta_evento 1 2 Sexo Edad Estadio_clinico LDH Tiempo_hasta_evento 1 3 Sexo Edad Estadio_clinico LDH Tiempo_hasta_evento 1 4 Sexo Edad Estadio_clinico LDH Tiempo_hasta_evento	
		> dataimp Class: mids Number of multiple imputations: 5 Imputation methods: Sexo Edad IPI Estadio_clinico "rf" "rf" "" "rf"	
		ECOG LDH Tratamiento Tiempo_hasta_evento "" "rf" "" "rf" Muerte ""	
		Number of logged events: 2 it im dep meth out 1 0 0 constant IPI 2 0 0 constant Tratamiento	
Generación de base de datos imputada	# Generar base de datos imputada		Se genera una base de datos con los datos imputados.
	#### Con el Paquete MICE		Se genera un gráfico donde se puede visualizar ahora la base no cuenta con datos perdidos(vis_miss).
	dataimputada <- mice::complete(dataimp) vis_miss(dataimputada)#visualizar base de datos imputada		
Exportación de base de datos******	# Exportación		
	#### Con el Paquete openxlsx write.xlsx(dataimputada, file = "C:/Users/Documentos/dataimputada. xlsx")		

*Para los que cuentan con una base de datos a imputar de su investigación, la pueden importar al software en lugar de realizar la simulación. Revisar la quinta referencia del presente artículo[6].

**Es importante convertir a las variables cualitativas a factores debido a que el algoritmo del código de Random Forest solo reconoce a las variables cualitativas como factores para proceder con la imputación.

***Se muestra como realizar la imputación de datos con Random Forest con dos paquetes estadísticos de R diferentes, la elección por cada uno de ellos dependerá del investigador.

****En la exp ortación de datos, deben colocar la dirección donde guardaran la base de datos, cuando cop ien la dirección tener en cuenta que viene con "backslash" o \ y p ara que p ueda ser ejecutado en R, deben cambiar
cada una a "slash" o / como en el ejemp lo.

Nota: En el software R solo debe copiarse los códigos de la segunda columna para que puedan ser ejecutados.

Antes de la imputación, evaluamos la composición de los datos faltantes y lo realizamos con la función "vis_miss" del paquete 'naniar'. Observamos valores ausentes en algunas variables clave, representando aproximadamente el 1,7% de la base de datos. Tras examinar los patrones de ausencia de datos y analizar si los valores faltantes se presentan de forma completamente al azar, al azar o no aleatoriamente, confirmamos, mediante la prueba de Little (p=0,6) implementada con la función "mcar_test" del paquete 'naniar', que los datos perdidos eran completamente aleatorios, lo cual es un requisito para realizar una imputación válida. [3]

Para llevar a cabo la imputación con RF, aplicamos el enfoque utilizado en el paquetes 'mice'[8]000 random samples of 2,000 persons drawn from the 10,128 stable angina patients in the CALIBER database (Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Records; 2001–2010. En el método con mice, empleamos la función "mice(datos, method = "rf")", donde "rf" permite usar RF como método de imputación. Se observa directamente en la salida de la imputación en 'mice', que las variables IPI y Tratamiento no fueron imputadas ya que no presentaban valores perdidos o mostraron comportamiento constante en los datos. Para facilitar su uso en otros software, los datos finales fueron exportados en formato ".xlsx" mediante la función "write.xlsx" del paquete 'openxlsx'. Es importante señalar que no hay un número mágico (puntos de corte como 10% o 20% de pérdidas) de porcentaje aceptable para los datos perdido, debido a que [9] este porcentaje debe mantenerse por debajo de un umbral aceptable el cual puede variar según el objetivo del estudio y la patología en evaluación [9,10]. Además, no recomendamos imputar variables de desenlace ni aquellas relacionadas con el tiempo de seguimiento.

Es esencial reconocer que, aunque RF es una herramienta robusta para la imputación de datos, no debe considerarse una solución universal. Evaluar la aplicabilidad de estos nuevos métodos en la hemato-oncología es crucial para maximizar la precisión y relevancia de los resultados en futuras investigaciones nacionales.

REFERENCIAS BIBLIOGRÁFICAS

1. Tang M, Pearson S-A, Simes RJ, Chua BH. Harnessing Real-World Evidence to Advance Cancer Research. Curr Oncol. 2023;30(2):1844. doi:10.3390/curroncol30020143
2. Derman BA, Belli AJ, Battiwalla M, Hamadani M, Kansagra A, Lazarus HM, et al. Reality check: Real-world evidence to support therapeutic development in hematologic malignancies. Blood Rev. 2022;53:100913. doi:10.1016/j.blre.2021.100913
3. Montenegro-Montenegro E, Oh Y, Chesnut S. No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos. Actual En Psicol. 2015;29(119):29–42. doi:10.15517/ap.v29i119.18812
4. Austin PC, White IR, Lee DS, van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Can J Cardiol. 2021;37(9):1322–31. doi:10.1016/j.cjca.2020.11.010
5. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20(1):199. doi:10.1186/s12874-020-01080-1
6. Cargar base de datos a RStudio [Internet]. 2020 [citado el 22 de junio de 2025]. Disponible en: https://www.youtube.com/watch?v=DLfRH-CD6jU
7. Vicente Coll Serrano. Introducción al Análisis Exploratorio de Datos. Aplicaciones con R y datos reales. [Internet]. 2020 [citado el 22 de junio de 2025]. 389 p. Disponible en: https://www.researchgate.net/publication/376184397_Introduccion_al_Analisis_Exploratorio_de_Datos_Aplicaciones_con_R_y_datos_reales
8. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. Am J Epidemiol. 2014;179(6):764–74. doi:10.1093/aje/kwt312
9. Rodríguez MD, Díaz JL, Massons JMD. Estudios para pruebas diagnósticas y factores pronósticos. Graunt21; 2022. 220 p.
10. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019;110:63–73. doi:10.1016/j.jclinepi.2019.02.016