Análisis del impacto del proceso de data cleaning sobre indicadores de malnutrición

Agustín Dramis[1,2], María Soledad Fernández[1,2], Adriana Pérez[1], and Pablo Turjanski[1,2]

[1] Grupo de Bioestadística Aplicada, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina (GBA, FCEyN-UBA)
[2] Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina(CONICET).
*agustin.dramis@gmail.com

Resumen El registro sistemático de medidas antropométricas permite evaluar el estado nutricional de poblaciones, siendo un insumo fundamental para diseñar, dirigir y evaluar políticas públicas. Las medidas antropométricas suelen ser recolectadas en un proceso de registro manual por parte de los profesionales de la salud. Este proceso acarrea la aparición de errores de carga pudiendo impactar en la evaluación del estado nutricional de la población. Para remediarlo la OMS introdujo pautas de remoción de datos individualmente no plausibles. Sin embargo, no son consideradas suficientes para la detección de la totalidad de los errores. Existen métodos que detectan inconsistencias longitudinales en registros de un mismo individuo. En este trabajo simulamos una base de datos antropométrica (basados en una real), a la que aplicamos aleatoriamente cuatro tipos de errores descritos en la literatura. Observamos el impacto de los mismos y el del proceso de limpieza (transversal y longitudinal), sobre la prevalencia de un indicador de malnutrición. Se encontró un aumento de la prevalencia luego de introducir cada tipo de error, y un acercamiento a los valores originales de prevalencia luego de los procesos de limpieza, evidenciando la importancia de aplicar estos procesos de data cleaning previo a analizar los indicadores nutricionales.

Keywords: Calidad de datos · Simulación · Datos antropométricos.

1. Introducción

La baja talla para la edad es una de las formas en que más frecuentemente se manifiesta la malnutrición, y se considera como el mejor indicador de desigualdades en la salud de niños y niñas [25]. Afecta el desarrollo físico, cognitivo, social y emocional del niño y constituye uno de los principales mecanismos de transmisión intergeneracional de la pobreza y la desigualdad [8]. Este indicador se determina en base al puntaje-z de la talla para la edad (haz, por sus siglas en inglés "height for age z-score"), que indica cuánto se aleja un individiduo (en unidades de desvío estándar) de la talla esperada para su edad y sexo y que permite determinar la presencia de déficit de crecimiento. Un individuo se considera

con baja talla si su valor de haz es menor a -2 [6]). Las grandes bases de datos de talla generadas por los sistemas de salud representan un recurso invaluable para el análisis de datos a gran escala, permitiendo determinar la existencia de malnutrición a nivel individual y su prevalencia a distintos niveles jurisdiccionales [1,14,15]. Los análisis de estas bases de datos permiten realizar monitoreos y definir áreas o grupos prioritarios de políticas públicas, orientando decisiones en base a evidencia. Considerando que el proceso de toma de datos de medidas antropométricas habitualmente es manual, y que los datos siguen un largo camino hasta ser centralizados, es inevitable la aparición de errores e inconsistencias debido a la transcripción, codificación y malos entendidos [22]. Ignorar estos errores puede generar efectos adversos en reportes de datos [12], análisis estadísticos [19] e incluso invalidar los resultados [20]. Así, surge la necesidad de implementar, previo al análisis de datos, un proceso de limpieza, entendido como el "proceso de detección, diagnóstico y edición de datos defectuosos"[3]. Para detectar datos extremos, la Organización Mundial de la Salud (OMS) sugiere la remoción de aquellos puntajes-z extremos (por fuera de límites pre-establecidos) que representan valores biológicamente no plausibles [7,16,18]. Sin embargo, cuando se lleva a cabo el seguimiento en el tiempo de un mismo individuo surge el problema de la existencia de puntajes-z comprendidos en el rango de lo plausible, pero inconsistentes en el tiempo si se considera su pertenencia a un mismo individuo. Esto se evidencia, por ejemplo, en decrecimientos o crecimientos inusualmente grandes. En estos casos existe menos consenso sobre la metodología a emplear para su limpieza [13], Nuestro equipo ha desarrollado un método a tal efecto [9], aplicado sobre registros de talla provenientes de controles de salud de individuos de 0 a 18 años con cobertura pública exclusiva de todo el país (Fuente: Programa SUMAR [23]). Sin embargo, al desconocer cuáles de los registros eran realmente erróneos en la base de datos real utilizada, no fue posible evaluar cuantitativamente su desempeño. La gran actividad de la comunidad científica en el área de limpieza de bases de datos de registros longitudinales deja en evidencia que el problema no es exclusivo de una base de datos o una región particular, sino que es de alcance global [27,21,28,24,5,4]. Menos frecuente en la literatura son los estudios que documenten el impacto de la limpieza en indicadores nutricionales, aunque existen algunos ejemplos [11,2,10,13]. El objetivo general de este trabajo es cuantificar el impacto de cuatro tipos de errores habituales durante la carga manual de datos [26] sobre la proporción de individuos con baja talla para la edad y el desempeño de dos herramientas de limpieza complementarias, aplicadas secuencialmente para identificar dichos errores y preservar datos correctos. Para ello se utilizará una base de datos simulada de registros de talla provenientes de controles de salud.

2. Métodos

Se confeccionó una base de datos de registros de talla para 10.000 individuos de entre 0 y 18 años de edad (10 registros de talla longitudinales por individuo). La base simulada estuvo conformada por "ID del individuo (1:10.000)", "ID

del registro (1:100.000)", "fecha de control (dd-mm-aa)", "fecha de nacimiento (dd-mm-aa)", "sexo" y "talla (cm)". A cada individuo se le asignó un valor de puntaje-z (equivalente a asignarle un percentil de talla) y se mantuvo el mismo en sus 10 registros. Los puntajes-z fueron tomados de una distribución normal (0,1) para el 90 % de los individuos y al 10 % restante se le asignó un valor con una distribución uniforme en el intervalo [-2,-6], a fin de tener una prevalencia de aproximadamente 10 % de baja talla para la edad dentro de los valores plausibles según OMS. Se procuró que la base de datos simulada respete la estructura de la base real (SUMAR) en relación a estructura de edad, proporción de sexos, cantidad de registros por individuo y prevalencia de baja talla [15]. Para ello, al 90 % de los individuos se les asignó su puntaje-z en base a una distribución normal (0,1), mientras que al 10 % restante se le asignó un puntaje-z correspondiente a malnutrición, dentro de los límites plausibles, con una distribución uniforme entre -2 y -6. En cuanto a las edades, se procuró garantizar la presencia de registros a edades tempranas con mayor frecuencia, tal como se observa en los datos originales. Las bibliotecas provistas por la OMS (anthro y anthroplus) presentan funciones para determinar el puntaje de un individuo a partir de su sexo, edad y talla. Dado que no incluyen la funcionalidad de determinar la talla a partir del puntaje-z (necesario en este caso dada la necesidad de recrear los registros a distintas edades de un individuo de puntaje-z constante), se desarrolló un método que determina los puntajes-z de una talla extremadamente baja y una extremadamente alta, y luego, en forma similar a una búsqueda binaria, itera sobre distintas tallas intermedias hasta encontrar la correspondiente a un puntaje-z que se aleje menos de 0,1 del deseado, o hasta tener una diferencia de talla menor a 1cm entre dos iteraciones sucesivas.

Partiendo de esta base de datos inicial (BI) se generaron 4 bases modificadas incluyendo distintos tipos de errores en un 5 % de sus registros (Tabla 1)).

Cuadro 1. Tipos de errores en cada una de las cuatro bases modificadas y ejemplo numérico de valores originales y modificados.

			Ejemplo numérico	
Nombre de la base	Nombre de error	Descripción	Valor original (cm)	Valor modificado (cm)
BMa	Ea	Anagrama. Se reemplazó el número por un anagrama, es decir un reordenamiento aleatorio de sus cifras	123	132
BMb	Eb	Variación de una cifra por uno. Se cambió el valor de una de sus cifras por el número anterior o el siguiente	145	155
BMc	Ec	Dato sin sentido. Se reemplazó el número por otro al azar, con la misma cantidad de cifras	87	94
BMd	Ed	Un dígito incorrecto. Se modificó el valor de una de sus cifras por otro número distinto entre 0 y 9	105	195

A cada una de las cuatro bases se les aplicaron de manera secuencial dos herramientas (H) de limpieza de datos: H1) Remoción de valores no plausibles (transversal) y H2) Remoción de valores asociados a cambios no plausibles (longitudinal). La H1 consistió en detectar valores considerados biológicamente no plausibles según los límites establecidos por la OMS, siendo estos de 6 desvíos de la media (haz < -6; haz > 6) [17]. La H2 implementa el método longitudinal desarrollado por el grupo [9]. Para ambas herramientas, a todos aquellos valores de haz que se detectaron como erróneos se les asignó el valor "NA".

2.1. Análisis de datos

Se calculó la prevalencia de baja talla (p.bt) como la cantidad de individuos que presentaron al menos un registro con haz < -2. La prevalencia fue calculada en la BI y en las cuatro BM previo y posterior a la aplicación de las herramientas de limpieza. Si las herramientas funcionasen correctamente se esperaría que eliminen correctamente los datos erróneos y que no afectasen (o lo hagan en baja proporción) a los datos correctos, esperando un acercamiento de la prevalencia de baja talla post limpieza a la de la BI. Para evaluar el desempeño de las herramientas de limpieza, en cada una de las cuatro BM se calcularon la sensibilidad y la especificidad de H1 (solo limpieza transversal) y H1 + H2 (transversal + longitudinal). La sensibilidad se calculó como # registros erróneos eliminados / # registros erróneos totales, y la especificidad se calculó como # registros sin errores retenidos/ # registros sin errores totales. Los procesos de limpieza y análisis se realizaron utilizando R versión 3.6.3 y RStudio versión 1.1.463, en una computadora de escritorio con procesador i5, memoria RAM de 64Gb, disco de 1Tb y Sistema Operativo Ubuntu 18.04

3. Resultados

En los datasets generados con fuentes de error introducidas (BM) la p.bt calculada aumentó entre 13,22 y 27,22 puntos porcentuales (Tabla 2) Al aplicar las herramientas de limpieza (H1 y H2, de manera secuencial) en todos los casos se obtuvo un valor de p.bt cercano al de la base BI, con una diferencia de entre 0,12 y 3,27 puntos porcentuales. Si bien el valor de la prevalencia fue mayoritariamente ajustado luego de implementar H1, luego de H2 se logró un acercamiento adicional de entre 0,19 y 0,53 puntos porcentuales. La cantidad de registros removidos por H1 fue ampliamente superior a la de la H2 para todos los tipos de error.

La sensibilidad de la limpieza aplicada fue de (44,49 - 83,56) % según tipo de error, siempre aumentando al aplicar adicionalmente la H2, siendo notoriamente menor para el Ed. La especificidad fue alta en todos los casos, entre 99,80 % y 99,89 % según tipo de error, siempre disminuyendo ligeramente al aplicar la H2 (Tabla 3).

Cuadro 2. Prevalencia de baja talla (p.bt) y cantidad de registros removidos para cada una de las bases luego de introducir errores y aplicar cada herramienta de limpieza.

Base	Nombre del error introducido	p.bt	n removidos H1	p.bt Post H1	n removidos H2	p.bt Post H2
BI	—	12,17 %	—	—	—	—
BMa	Ea	29,52 %	4.116	13,54 %	78	13.35 %
BMb	Eb	39,39 %	3.050	14,50 %	242	13,97 %
BMc	Ec	29,03 %	4.054	12,48 %	96	12,29 %
BMd	Ed	25,39 %	2.177	15,58 %	170	15,34 %

Cuadro 3. Prevalencia de baja talla (p.bt) y cantidad de registros removidos para cada una de las bases luego de introducir errores y aplicar cada herramienta de limpieza.

	BMa		BMb		BMc		BMd	
	H1	H1 + H2	H1	H1 + H2	H1	H1 + H2	H1	H1 + H2
Sensibilidad	82,45 %	83,56 %	60,45 %	63,70 %	81,27 %	82,50 %	42,40 %	44,49 %
Especificidad	99,89 %	99,87 %	99,98 %	99.80 %	99,89 %	99,85 %	99,89 %	99,82 %

4. Conclusiones

En este trabajo se muestra el impacto que tiene sobre la estimación del indicador de baja talla para la edad -indicador de malnutrición de gran importancia en salud-, la existencia de un 5 % de registros erróneos (un tipo de error a la vez) en una base de datos antropométricos. Se observó que en todos los casos de errores introducidos en este trabajo se genera un aumento de la p.bt, que puede hasta triplicarse. Esto es esperable ya que, al haber un 10 % de individuos inicialmente con baja talla, es más probable que los errores impacten sobre el 90 % restante. Además este incremento se encuentra magnificado por el hecho de calcular la prevalencia en base al menor registro de haz del individuo en el tiempo.

El error con mayor impacto sobre la p.bt fue el de variar una de las cifras de la talla en uno (Eb). Esto se puede explicar también por la manera de determinar la prevalencia en base al menor registro de haz del individuo en el tiempo, ya que es probable que en el 50 % de los errores introducidos el individuo disminuya su talla. En el caso del resto de los errores -anagrama (Ea), dato sin sentido (Ec) y un dígito incorrecto (Ed)-, como la mayoría de las tallas poseen un “1” en la primera cifra (los registros de tallas de la base de datos real se distribuyen mayoritariamente por encima de 100 cm), es más probable generar un valor más alto que el inicial. Por ejemplo, en el caso de anagrama (Ea) cifras distintas de 0 o 1 pueden finalmente ubicarse en la primera posición; en el caso de dato sin sentido (Ec) aumenta el rango de talla (puede ubicarse entre 0 y 999 cm); y por último el caso de un dígito incorrecto puede modificar el 1 de la primer

posición por valores entre 0 y 9. Como estos tres tipos de errores tienen una mayor probabilidad de aumentar el valor absoluto, se espera observar un mayor impacto relativo de éstos en el cálculo de indicadores de malnutrición por exceso, como puede ser la obesidad.

A su vez, se observa un acercamiento a la prevalencia de malnutrición original al aplicar las dos herramientas de limpieza en la base de datos. La mayoría de los errores se eliminaron con la herramienta H1. Sin embargo, esto no coincide con lo reportado para métodos longitudinales, como el propuesto por Phan y cols [21], quienes hallaron que el método de la OMS detectó menos del 1 % de los errores mientras que su método propio, que incorpora un análisis longitudinal, detectó un porcentaje mayor de errores. Sin embargo, debe tenerse en cuenta que en nuestro caso los protocolos de limpieza fueron aplicados secuencialmente, por lo que el impacto de la limpieza longitudinal se dio sobre una base con un 100 % de datos individualmente plausibles. Adicionalmente, puede estar sucediendo que nuestro trabajo no esté simulando los tipos de errores con los que se enfrentaron en su trabajo estos investigadores. Como trabajo futuro se evaluará la aplicación de la herramienta H2 de manera independiente.

En cuanto a la limpieza longitudinal, la herramienta H2 removió una mayor cantidad de registros en aquellas bases en las que la H1 removió menos (BMb y BMd), por lo que la limpieza longitudinal resultó ser un buen complemento a las pautas de la OMS para la remoción de valores no plausibles longitudinalmente, sobre todo en bases de datos con una mayor proporción de errores que generaron tallas erróneas pero plausibles. En relación a la sensibilidad, esta aumentó mayormente en las bases BMb y BMd, si se las compara con las BMa y BMc, luego de incorporar la herramienta H2.

Se pudo apreciar una distinción entre dos grupos de errores. Los errores de las bases BMa y BMc, que afectan a todas las cifras de manera simultánea, mostraron una menor cantidad de registros removidos por parte de la herramienta H2, en contraste a las bases BMb y BMd cuyos errores afectan una única cifra.

La disminución (aunque leve) de la especificidad luego de aplicar H2 podría mejorar a futuro utilizando criterios más estrictos de cambios no plausibles de manera de refinar el método y evitar la remoción de registros correctos.

Si bien anteriormente se ha estudiado el impacto del aumento de la calidad de datos en indicadores de malnutrición [11], esto se hizo sobre bases de datos reales, utilizando métricas de calidad de datos. En el presente trabajo, sobre una base simulada, se evidenció el impacto de la introducción de un porcentaje de errores concretos y comunmente observados en procesos de carga manual [26] sobre la prevalencia de baja talla para la edad, un indicador de relevancia en salud pública. A su vez, el trabajo resalta la relevancia de realizar un proceso de limpieza de datos que incluya métricas de desempeño concretas de las herramientas utilizadas con el fin último de aumentar la calidad de la información reportada, facilitar la comparabilidad de resultados y orientar a tomadores de decisiones.

Referencias

1. Mapping child growth failure across low-and middle-income countries. Nature **577**(7789), 231–234 (2020)
2. Boone-Heinonen, J., Tillotson, C.J., O'Malley, J.P., Marino, M., Andrea, S.B., Brickman, A., DeVoe, J., Puro, J.: Not so implausible: impact of longitudinal assessment of implausible anthropometric measures on obesity prevalence and weight change in children and adolescents. Annals of epidemiology **31**, 69–74 (2019)
3. Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., Herbst, K.: Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine **2**(10), e267 (2005)
4. Chen, S., Banks, W.A., Sheffrin, M., Bryson, W., Black, M., Thielke, S.M.: Identifying and categorizing spurious weight data in electronic medical records. The American journal of clinical nutrition **107**(3), 420–426 (2018)
5. Daymont, C., Ross, M.E., Russell Localio, A., Fiks, A.G., Wasserman, R.C., Grundmeier, R.W.: Automated identification of implausible values in growth data from pediatric electronic health records. Journal of the American Medical Informatics Association **24**(6), 1080–1087 (2017)
6. De Onis, M., Onyango, A.W.: Who child growth standards. The Lancet **371**(9608), 204 (2008)
7. De Onis, M., Onyango, A.W., Borghi, E., Garza, C., Yang, H., Group, W.M.G.R.S., et al.: Comparison of the world health organization (who) child growth standards and the national center for health statistics/who international growth reference: implications for child health programmes. Public health nutrition **9**(7), 942–947 (2006)
8. Dewey, K.G., Begum, K.: Long-term consequences of stunting in early life. Maternal & child nutrition **7**, 5–18 (2011)
9. Fernández, M.S., Altszyler, E., Dramis, A., Cueto, G., Pérez, A., Núñez, P., Turjanski, P.: Método de remoción de medidas anómalas en datos de crecimiento infanto-juvenil: una aplicación para grandes bases de datos en salud. In: VII Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2021)-JAIIO 50 (Modalidad virtual) (2021)
10. Freedman, D.S., Lawman, H.G., Pan, L., Skinner, A.C., Allison, D.B., McGuire, L.C., Blanck, H.M.: The prevalence and validity of high, biologically implausible values of weight, height, and bmi among 8.8 million children. Obesity **24**(5), 1132–1139 (2016)
11. Harkare, H.V., Corsi, D.J., Kim, R., Vollmer, S., Subramanian, S.: The impact of improved data quality on the prevalence estimates of anthropometric measures using dhs datasets in india. Scientific Reports **11**(1), 10671 (2021)
12. Horn, P.S., Feng, L., Li, Y., Pesce, A.J.: Effect of outliers and nonhealthy individuals on reference interval estimation. Clinical chemistry **47**(12), 2137–2145 (2001)
13. Lawman, H.G., Ogden, C.L., Hassink, S., Mallya, G., Vander Veur, S., Foster, G.D.: Comparing methods for identifying biologically implausible values in height, weight, and body mass index among youth. American journal of epidemiology **182**(4), 359–365 (2015)
14. Nuñez, P.A., Fernández, M.S., Turjanski, P., Pérez, A., Rivero, M.R., De Angelo, C., Salomón, O.D., Cueto, G.: Substantial reduction in child stunting is differentially associated to geographical and socioeconomic disparities in Misiones Province, Argentina. Tropical Medicine & International Health **25**(7), 874–885 (2020)

15. Nuñez, P.A., Fernández-Slezak, D., Farall, A., Szretter, M.E., Salomón, O.D., Valeggia, C.R.: Impact of universal health coverage on child growth and nutrition in Argentina. American journal of public health **106**(4), 720–726 (2016)
16. Onis, M.d., Onyango, A.W., Borghi, E., Siyam, A., Nishida, C., Siekmann, J.: Development of a who growth reference for school-aged children and adolescents. Bulletin of the World health Organization **85**(9), 660–667 (2007)
17. Organization, W.H., et al.: Physical status: The use of and interpretation of anthropometry, Report of a WHO Expert Committee. World Health Organization (1995)
18. Organization, W.H., et al.: WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. World Health Organization (2006)
19. Osborne, J.W.: Data cleaning basics: Best practices in dealing with extreme scores. Newborn and Infant Nursing Reviews **10**(1), 37–43 (2010)
20. Osborne, J.W.: Is data cleaning and the testing of assumptions relevant in the 21st century? (2013)
21. Phan, H.T., Borca, F., Cable, D., Batchelor, J., Davies, J.H., Ennis, S.: Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. Scientific reports **10**(1), 10164 (2020)
22. Pritzker, L., Ogus, J., Hansen, M.H.: Computer editing methods-some applications and results. Bulletin of the International Statistical Institute **41**(1), 442–472 (1965)
23. Sabignoso, M., Zanazzi, L., Sparkes, S., Mathauer, I., Organization, W.H., et al.: Strengthening the purchasing function on through results-based financing in a federal setting: lessons from argentina's programa sumar (2020)
24. Shi, J., Korsiak, J., Roth, D.E.: New approach for the identification of implausible values and outliers in longitudinal childhood anthropometric data. Annals of epidemiology **28**(3), 204–211 (2018)
25. WHO, E.: Committee physical status: The use and interpretation of anthopometry: Report of a who expert committee. WHO Technical Report Series **854** (1995)
26. Wiseman, S., Cairns, P., Cox, A.: A taxonomy of number entry error. In: Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction 25. pp. 187–196 (2011)
27. Woolley, C.S., Handel, I.G., Bronsvoort, B.M., Schoenebeck, J.J., Clements, D.N.: Is it time to stop sweeping data cleaning under the carpet? a novel algorithm for outlier management in growth data. PloS one **15**(1), e0228154 (2020)
28. Wu, D.T., Meganathan, K., Newcomb, M., Ni, Y., Dexheimer, J.W., Kirkendall, E.S., Spooner, S.A.: A comparison of existing methods to detect weight data errors in a pediatric academic medical center. In: AMIA Annual Symposium Proceedings. vol. 2018, p. 1103. American Medical Informatics Association (2018)