

Método de Remoción de Medidas Anómalas en Datos de Crecimiento Infanto-Juvenil: una Aplicación para Grandes Bases de Datos en Salud

María Soledad Fernández^{1,2}, Edgar Altszyler^{1,2}, Agustín Dramis¹, Gerardo Cuetto^{1,2}, Adriana Pérez¹, Pablo Núñez^{1,2}, 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, Argentina (CONICET).

*pturjanski@dc.uba.ar

Resumen La malnutrición infantil y juvenil es una problemática de gran relevancia en salud pública. Dentro de los indicadores antropométricos para evaluar malnutrición infantil se encuentran la talla para la edad, el peso para la talla y el índice de masa corporal, que utilizan para su cómputo la talla del individuo. En la Argentina, el Programa SUMAR recolecta desde el año 2013 información de controles de salud de la población infantil y juvenil más vulnerable del país. Esto conforma una gran base de datos de registros antropométricos (incluida la talla) a partir de la cual se estiman las prevalencias de distintos tipos de malnutrición, como baja talla, bajo peso para la talla, obesidad y sobrepeso. Por su volumen y manera de generar los datos, dicha base posee inexorablemente errores, los cuales impactan sobre el cálculo de los indicadores a nivel individual y de las prevalencias. Existen pautas de la OMS para la identificación de valores de indicadores no plausibles, según edad y sexo. Sin embargo, estas pautas no resuelven el problema de la existencia de registros que, si bien son individualmente posibles, no lo son si se considera el crecimiento en el tiempo. El objetivo de este trabajo es presentar un método para detección de datos de talla inconsistentes en el tiempo, ya sea por decrecimiento o por crecimiento excesivo. Este método podrá ser incluido en un proceso de limpieza global de esta base así como de otras similares, de manera de aumentar la calidad de los datos.

Keywords: Talla · Datos Longitudinales · Malnutrición Infantil · Argentina.

1. Introducción

La malnutrición infantil y juvenil es una problemática de gran relevancia en salud pública. Permite evidenciar inequidades en salud y se asocia a la morbilidad y mortalidad infantil [21]. La malnutrición por defecto o desnutrición afecta

adversamente el desarrollo físico, social y cognitivo durante la niñez y genera consecuencias adversas en el desarrollo educativo y económico durante la vida adulta [3,10,22,26]. La malnutrición por exceso (sobrepeso u obesidad) se asocia a mayores probabilidades de padecer enfermedades crónicas no transmisibles durante la adultez [30]. Para el estudio de la malnutrición infantil se utilizan indicadores (puntajes-z) que se calculan a partir de la comparación de registros antropométricos (talla, peso, IMC) con valores de referencia para individuos del mismo sexo y edad [29]. En Argentina, existe una base de datos de dichos registros que cubre ampliamente a la población infanto-juvenil más vulnerable del país, recolectada por el Programa SUMAR (2013 - actualidad), dependiente del Ministerio de Salud de la Nación [14]. La base de datos del Programa abarca controles de salud de aproximadamente 4 millones de niños, niñas y adolescentes, distribuidos en más de 6.500 establecimientos de salud públicos de todo el país. La base se construye a partir de la carga manual de datos por parte de los profesionales de salud. Esta manera de registrar la información genera la aparición de datos erróneos que puede generar indicadores con valores biológicamente no plausibles o falta de consistencia entre distintos registros longitudinales de un mismo individuo. La existencia de errores en grandes volúmenes de datos debido a la transcripción, codificación y malentendidos, es frecuente [23]. Ignorar estos errores puede llegar a generar efectos adversos en reportes de datos, análisis estadísticos e incluso invalidar sus resultados [11,19,20]. En el caso de la base de datos bajo estudio, estos errores repercuten tanto en la valoración del estado nutricional de un individuo como en la estimación de la prevalencia de malnutrición a distintos niveles jurisdiccionales.

Es por esto que es necesario realizar una limpieza previa al cálculo de los indicadores de malnutrición, definiendo la limpieza como el "proceso de detección, diagnóstico y edición de datos defectuosos"[5]. Distinguir errores originados en anomalías genuinas en datos antropométricos como talla y peso es muy complejo, dado que dichos datos biológicos son heterogéneos y pueden contener valores inusuales pero plausibles. Las pautas de la Organización Mundial de la Salud (OMS) pueden ser un buen punto de partida para excluir valores extremos (ej. biológicamente no plausibles) para los indicadores calculados a partir de los datos de talla y peso [9,18,29]. Estas pautas se basan en la remoción de valores extremos de los puntajes-z de cada registro, utilizando límites preestablecidos (ej. exclusión del puntaje-z de "talla para la edad" (*haz*) si el mismo supera en 6 desvíos la media). Estos criterios permiten detectar y remover los puntajes-z considerados biológicamente no plausibles para los controles de salud para distintos indicadores que involucren a la talla, el peso o a ambas medidas en sus cálculos. Sin embargo, existe menos consenso sobre la metodología a emplear para la limpieza de datos antropométricos longitudinales [13], es decir, aquellos que implican un seguimiento en el tiempo de un mismo individuo. En estudios longitudinales surgen nuevas situaciones, tales como la presencia de puntajes-z comprendidos en el rango de lo plausible individualmente, pero inconsistentes en el tiempo si se considera su pertenencia a un mismo individuo (ej. decrecimientos o crecimientos excesivos). Los métodos existentes en la literatura para la lim-

pieza de datos longitudinales son bastantes disímiles entre sí [1,6,7,25,27,31]. Si bien los principios utilizados en los métodos propuestos en la literatura pueden tomarse como guía para la limpieza de la base bajo estudio, la misma requiere un desarrollo original debido a las particularidades de dicha base en relación a otras, la naturaleza de los errores de carga registrados y por tratarse además de una base compuesta por más de 20 millones de registros, con actualizaciones periódicas.

El objetivo de este trabajo es presentar un método para detección de datos antropométricos de talla inconsistentes en el tiempo, tomando como base los registros de talla de la población infanto-juvenil de Argentina perteneciente al programa SUMAR. Este método podrá ser incluido en un proceso de limpieza global de la base de datos, de manera de mejorar su calidad y contar con indicadores antropométricos y de malnutrición más precisos al momento de evaluar las condiciones de salud de la población infantil y juvenil más vulnerable del país.

2. Materiales y Métodos

Fuente de datos

La fuente de datos está conformada por datos recolectados por el Programa SUMAR, durante el periodo que se extiende desde el año 2013 hasta el primer cuatrimestre del año 2020, en particular por los controles de salud y sus registros antropométricos correspondientes a niños, niñas y jóvenes de 0 a 19 años. Los datos de estos controles son sistemáticamente colectados en cada uno de los más de 6.500 efectores que forman parte del Programa y son enviados cuatrimestralmente al Ministerio de Salud de la Nación. Estos datos antropométricos conforman la fuente de datos primaria y tienen una amplia cobertura de la población más vulnerable del país. Se partió de una base pre-procesada en donde los campos de identificación única de cada individuo, sexo, fecha de nacimiento y fecha de control ya habían sido sometidos a un proceso de limpieza. Cabe destacar que parte de este pre-procesamiento consistió en retener un único control de salud por día por individuo. Sobre esta base, se calcularon los indicadores antropométricos bajo estudio, que son aquellos que involucran a la talla del individuo: talla para la edad (haz), índice de masa corporal (bmiz) y peso para la talla (wflz). Finalmente también, como parte del pre-procesamiento, se realizó una limpieza de los indicadores antropométricos, asignando NULL a aquellos valores biológicamente no plausibles, en base al puntaje-z y según los estándares de la OMS [28].

Proceso de limpieza de indicadores de talla

El proceso propuesto de limpieza se realizó en dos etapas: Etapa a) limpieza de registros de talla que mostraron **decrecimiento** entre controles sucesivos superior a un umbral establecido y Etapa b) limpieza de registros de talla que mostraron **crecimiento** entre controles sucesivos superiores a un umbral establecido.

Para ambas etapas se trabajó con: i) Controles de salud cuyo indicador antropométrico de haz era distinto a NULL, es decir que no hubiesen sido removidos previamente por tratarse de un valor no plausible individualmente según los límites establecidos por la OMS; y con ii) individuos que poseían al menos cuatro de estos controles en el tiempo. De esta forma el análisis sobre la fuente de datos primaria quedó restringido a un conjunto de datos de más de 13.000.000 de registros correspondientes a aproximadamente 2.000.000 de individuos. Cada individuo poseía un mínimo de cuatro controles de salud y un máximo de 30, aunque para aproximadamente 90% de estos individuos se registraron entre cuatro y diez controles.

Etapas a) limpieza de registros con decrecimientos de talla

Esta parte del proceso tuvo como objetivo detectar aquellos casos en que, entre dos controles consecutivos de un mismo individuo, se registre un decrecimiento superior a cierto umbral. Considerando un error de medición máximo de 1 cm en cada control, se estableció en forma conservadora un decrecimiento máximo tolerable (umbral) de 2 cm entre dos mediciones sucesivas de un mismo individuo. Para llevar adelante el proceso de limpieza se creó una nueva tabla donde se asoció a cada control de un individuo (*control n*) su siguiente control (*control n+1*), obteniéndose así una tabla con tantos registros como cantidad de controles de salud menos uno para cada individuo (debido a que el último control no posee un siguiente control). Como se mencionó, los controles utilizados para construir esta tabla fueron sólo aquellos que no poseían NULL en su indicador haz. Cada fila de esta tabla posee la información que se detalla en la Tabla 1 (*tabla de controles sucesivos*), incluyendo la diferencia de edad en días y diferencia de talla entre ambos controles.

De la *tabla de controles sucesivos* se seleccionaron sólo aquellos registros (pares de controles) que presentaron un decrecimiento mayor a 2 cm (es decir, `talla_diferencia < -2`). Denominamos a los pares y a los controles que los conforman **pares conflictivos**. Se desarrolló un procedimiento para detectar, en cada par, cuál de los dos controles resultaba menos fiable (o si lo eran ambos), y en dicho caso asignarles NULL a todos los indicadores para dicho control que dependan del campo talla (haz, bmiz y wflz).

El procedimiento aplicado, para identificar cuál de los dos controles resultaba menos fiable, se basó en computar la cantidad de **inconsistencias** de cada control del par conflictivo en relación a los controles no involucrados. A continuación se detalla el procedimiento implementado:

Paso 1: **Cantidad de inconsistencias para el control n .** Se computa la cantidad de decrecimientos entre el control n y todo el resto de los controles no involucrados en un par conflictivo. Aquellos controles no conflictivos que posean valor NULL en su indicador haz no participan de este paso.

Paso 2: **Cantidad de inconsistencias para el control $n+1$.** Se realiza el mismo procedimiento anterior, pero esta vez para el control $n+1$.

Paso 3: **Asignación de NULL a indicadores del control con mayor cantidad de inconsistencias.** Al control que posee la mayor cantidad de inconsistencias (*control n , control $n+1$* o a ambos en caso de empate), se le asigna en el dataset original NULL a todos sus indicadores bajo estudio que dependen del campo talla (*haz, bmiz y wflz*).

Paso 4: **Volver a ejecutar el Paso 1.** Al asignar NULL en el paso previo, es necesario volver a generar la *tabla de controles sucesivos*, seleccionar los nuevos pares conflictivos y ejecutar el proceso desde el Paso 1 (iteración). El proceso termina cuando, en esta última tabla, no se detectan más pares conflictivos.

Cuadro 1. Campos de la *tabla de controles sucesivos* utilizados por ambas etapas del proceso de limpieza.

Campo	Descripción
id_individuo	Identificador único de individuo
fecha_control_actual	Fecha del control n
edad_actual	Edad del individuo al momento del control n (días)
talla_actual	Talla del individuo al momento del control n (cm)
fecha_control_siguiete	Fecha del control $n+1$
edad_siguiete	Edad del individuo al momento del control $n+1$ (días)
talla_siguiete	Talla del individuo al momento del control $n+1$ (cm)
edad_diferencia	Diferencia de edad del individuo entre ambos controles ($edad_dias_siguiete - edad_dias_actual$) (días)
talla_diferencia	Diferencia de talla del individuo entre ambos controles ($talla_siguiete - talla_actual$) (cm)
tasa_crecimiento	$talla_diferencia / edad_dias_diferencia$ (cm/días)

Etapa b) limpieza de registros con crecimientos excesivos

Esta parte del proceso tiene como objetivo detectar aquellos casos en que, entre dos controles consecutivos de un mismo individuo, se registre un crecimiento excesivo. Para establecer un valor de corte de crecimiento excesivo, se utilizaron las curvas de crecimiento de talla para la edad para la población infantil argentina [24]. Se consideró como máxima tasa de crecimiento anual plausible aquella que se obtiene al pasar del percentil 3 al 97 en un año, que resulta, en promedio, en un crecimiento de 24 cm/año. Se incluyó asimismo el error de medición de 1 cm en cada registro. Se establecieron dos valores de corte en función de la edad de los individuos, considerando que para menores de un año se exigen controles al menos mensuales, pudiendo inclusive efectuarse controles quincenales, mientras que para los mayores de un año se exige un control anual. Para menores de un año, se calculó el crecimiento plausible quincenal más 2 cm de error de medición, resultando un crecimiento máximo tolerable entre dos mediciones de 0,2 cm diarios. En niños mayores de un año, se calculó un crecimiento anual plausible más 2 cm de error de medición, lo que representa 0.07 cm diarios.

Para llevar adelante el proceso de limpieza se creó nuevamente la Tabla 1 (*tabla de controles sucesivos*). Se definió **par conflictivo** al par de controles (*control n* y *control n+1*) que presente un *crecimiento excesivo* definido como:

- La tasa de crecimiento entre ambos controles es mayor o igual a 0.07 cm/días si la edad del individuo al momento del *control n* es mayor o igual a 365 días (`edad_dias_actual ≥ 365` y `tasa_crecimiento ≥ 0,07`)
- La tasa de crecimiento entre ambos controles es mayor o igual a 0.20 cm/días si la edad del individuo al momento del *control n* es menor a 365 días (`edad_dias_actual < 365` y `tasa_crecimiento ≥ 0,20`)

Además se incluyó la restricción de que, para ser conflictivos, la diferencia entre ambos controles debía superar un umbral de 2 cm (`talla_diferencia > 2`), por considerarse un límite en el error de medición en la detección de pares conflictivos, al igual que en la Etapa a.

Se desarrolló un procedimiento para detectar, en cada par, cuál de los dos controles resultaba menos fiable (o si lo eran ambos), y en dicho caso asignarles NULL a todos los indicadores para dicho control que dependan del campo talla (`haz`, `bmiz` y `wflz`). Esto se llevó a cabo, primero, i) para los controles realizados luego del año de vida (`edad_actual ≥ 365`) y, luego, ii) para la totalidad de los controles.

En ambos casos, el procedimiento aplicado para identificar cuál de los dos controles resultaba menos fiable se detalla a continuación:

- Paso 1: **Cantidad de inconsistencias para el control n .** Se computa la cantidad de crecimientos excesivos entre el control n y todo el resto de los controles no involucrados en un par conflictivo. Aquellos controles no conflictivos que posean valor NULL en su indicador haz no participan de este paso.
- Paso 2: **Cantidad de inconsistencias para el control $n+1$.** Se realiza el mismo procedimiento anterior, pero esta vez para el control $n+1$.
- Paso 3: **Asignación de NULL a indicadores del control con mayor cantidad de inconsistencias.** Al control que posee la mayor cantidad de inconsistencias (*control n , control $n+1$* o a ambos en caso de empate), se le asigna en el dataset original NULL a todos sus indicadores bajo estudio que dependen del campo talla (*haz, bmiz y wflz*).
- Paso 4: **Volver a ejecutar el Paso 1.** Al asignar NULL en el paso previo pueden aparecer nuevos conflictos, por lo tanto es necesario volver a generar la *tabla de controles sucesivos*, seleccionar los nuevos pares conflictivos y ejecutar el proceso desde el Paso 1 (iteración). El proceso termina cuando, en esta última tabla, no se detectan más pares conflictivos.

Una vez finalizadas ambas etapas de limpieza, se revisó manualmente para un grupo de individuos con distinta cantidad de controles y posiciones de los pares conflictivos en la serie de registro, si el algoritmo para la identificación y resolución de pares conflictivos por decrecimiento y por crecimiento excesivo había operado correctamente.

Luego de aplicados los procesos de limpieza descritos, los individuos podrían quedar con menos de cuatro controles de salud con valores de indicadores distintos a NULL. En dichos casos - individuos con menos de cuatro registros - se estableció un proceso de limpieza específico. Al detectarse pares conflictivos siguiendo la metodología propuesta anteriormente, la selección de cuál control sería retenido y cual removido, se realizó al azar.

3. Resultados

En el proceso de limpieza por **decrecimiento** se les asignó NULL a los indicadores antropométricos que involucran talla (talla para la edad, peso para la talla e índice de masa corporal) de un 8.5% de los controles procesados y en el caso de la limpieza por **crecimiento excesivo** este valor fue de 5.1%.

Para la limpieza por decrecimiento fue necesario iterar siete veces el algoritmo. Para la limpieza por crecimiento excesivo, en los casos que el primer control del par conflictivo correspondiera a un individuo mayor de un año, alcanzó con tres iteraciones; y en el caso de que el primer control del par conflictivo correspondiera a menores de un año solo hicieron falta dos iteraciones para llegar al final del proceso.

La revisión manual de la resolución de los conflictos por parte del algoritmo resultó exitosa en todos los casos revisados, independientemente del proceso (decrecimiento o crecimiento excesivo), la edad del individuo (menor o mayor a 365 días) y la posición del par conflictivo en la serie longitudinal. A modo de ejemplo, en la Figura 1 se muestra para un individuo con 15 registros de talla en el tiempo, la resolución de dos pares conflictivos por decrecimiento que se detectaron al aplicar el proceso descrito en la etapa 'a' del procedimiento. El primer par conflictivo involucra a los controles n°11 y n°12, y en este caso se retiene el n°12 ya que no presenta ninguna inconsistencia respecto a controles no involucrados en pares conflictivos (puntos verdes), mientras que el control n°11 presentó una inconsistencia respecto al control n°13. Para el segundo par conflictivo, el control n°14 no presentó inconsistencias mientras que el control n°15 presentó una en relación al control n°13, por lo tanto se remueve. Para este mismo individuo, una vez finalizado el proceso de limpieza por decrecimiento se encontró en la etapa 'b' del proceso (limpieza por crecimiento excesivo) un par conflictivo (ver Figura 2). En este caso el control n°13 no presentó inconsistencias mientras que el control n°14 presentó una respecto al control n°12 (tasa de crecimiento superior a 0.07 cm/día y una diferencia mayor a 2 cm entre esos controles, según las reglas establecidas), por lo tanto fue removido.

4. Discusión

En este trabajo se presenta una propuesta para la detección de errores en series longitudinales, específicamente en el campo talla, en una gran base de datos de importancia en salud pública y con cobertura nacional. Si bien trabajos a gran escala sobre los registros antropométricos en Argentina se han llevado a cabo previamente y han contemplado una limpieza de la base [16,17], el desarrollo e implementación de este método particular es la primera vez que se utiliza, según nuestro conocimiento.

Una de las principales ventajas del método planteado radica en que el mismo permite conservar los registros individuales de talla, pero al detectarse valores no confiables asigna NULL a los indicadores antropométricos que utilizan la talla del individuo. De esta forma se evita que registros no confiables ingresen al cómputo de prevalencias de indicadores de malnutrición, sin modificar los registros originales.

Idealmente, los métodos de limpieza de datos deberían priorizar reconstruir datos por sobre la eliminación de los mismos [32], y utilizar métodos informáticos

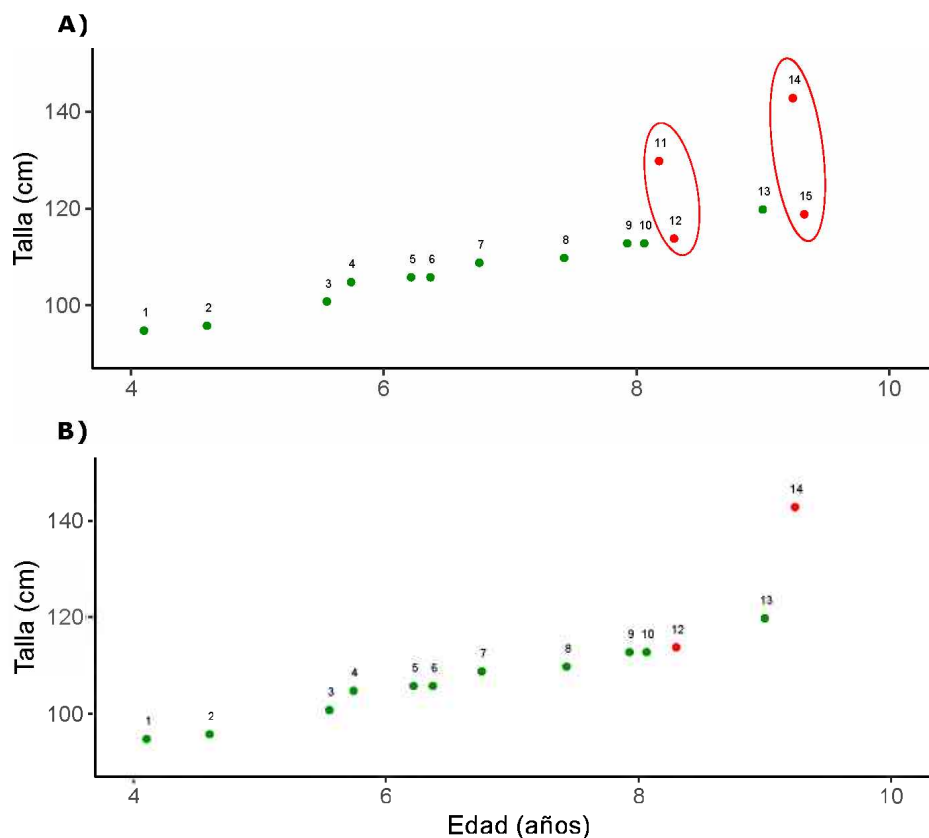


Figura 1. Ejemplo del funcionamiento del algoritmo de limpieza para decrecimiento (etapa 'a' del método). A) Detección de los pares conflictivos, B) Limpieza de controles. En el caso del primer par conflictivo, el control n°11 presentó una inconsistencia respecto a los demás controles no involucrados (puntos verdes) y el n°12 ninguna, por lo tanto se remueve el n°11 y se retiene el n°12. En el caso del segundo par conflictivo, el control n°14 no presenta inconsistencias mientras que el n°15 presenta una y es removido.

10 MS Fernández et al.

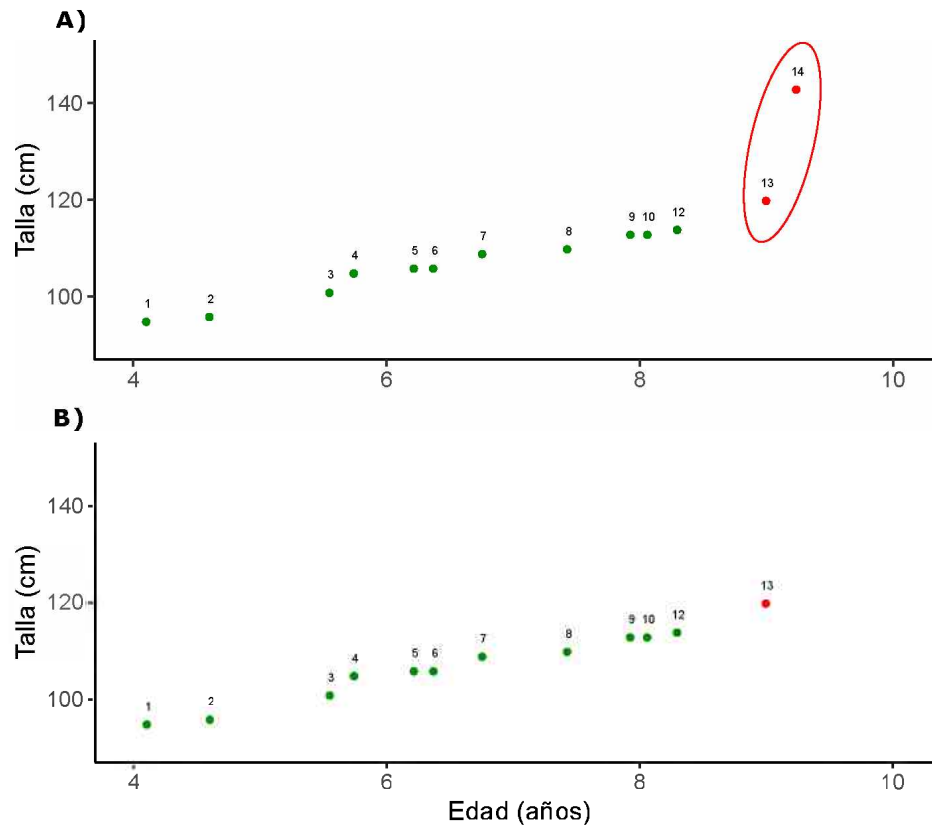


Figura 2. Ejemplo del funcionamiento del algoritmo de limpieza para crecimiento excesivo (etapa 'b' del método). A) Detección de los pares conflictivos, B) Limpieza de los pares conflictivos. En este caso el control n°13 no presenta inconsistencias respecto a los demás controles no involucrados (puntos verdes), mientras que el control n°14 presenta una inconsistencia respecto al control n°12, por lo tanto es removido

que permitan su reproducibilidad [8]. En el caso de la metodología propuesta en este trabajo, dada la naturaleza del campo estudiado, no fue posible reconstruir dichos datos, pero sí aplicar un método reproducible. Por otro lado, el algoritmo planteado permite detectar y cuantificar el tipo de inconsistencias en el campo bajo estudio (decrecimiento o crecimiento excesivo) y la magnitud de cada uno de ellos, lo cual a su vez permite implementar mejoras para reducir los errores. En este sentido, a futuro se plantea realizar un análisis detallado de sesgos por grupos (de edad, sexo, agrupamientos jurisdiccionales, etc.) con el objetivo de cuantificar la magnitud de dichos errores y estudiar su impacto en la estimación de la prevalencia de distintos indicadores de malnutrición para diferentes grupos y escalas. Esto permitirá dirigir de manera más eficiente las propuestas en la mejora del proceso de recolección y reporte de los datos.

Este método se utilizó para individuos con cuatro o más controles de salud válidos. Para individuos con tres o menos controles de salud, se consideró que la información longitudinal era insuficiente para la detección de inconsistencias, por lo que la decisión acerca de qué registro eliminar en caso de detectarse pares conflictivos fue aleatoria.

Existen aspectos que deben ser tenidos en cuenta al momento de realizar una limpieza. Por un lado, el orden en el que se realizan las etapas de la limpieza pueden alterar los porcentajes de registros afectados. En el caso de este trabajo, cuando el procedimiento se aplicó para crecimiento excesivo, ya había sido aplicado el proceso de limpieza por decrecimiento entre controles sucesivos. Por otro lado, las reglas establecidas tanto para la detección de pares conflictivos como para el cómputo de inconsistencias afectan la limpieza, y estas pueden ser modificadas en función de los objetivos del trabajo, la precisión deseada y los errores que se deseen tolerar, entre otros.

Se propone en un futuro, analizar el desempeño del método presentado en relación a otros ya existentes [4,25,27]. Para realizar esta comparación, se proyecta trabajar con una base de datos de crecimiento longitudinal, en la cual se simulen errores de carga que generen casos de decrecimiento y crecimiento excesivo. Sobre esta base con errores simulados conocidos, se evaluará la capacidad de detección de errores de diferentes métodos, pudiéndose comparar su desempeño con medidas tales como la sensibilidad y especificidad. Adicionalmente, los métodos serán evaluados respecto a su demanda computacional.

En resumen, el acceso a grandes volúmenes de datos, junto con la incorporación de técnicas y herramientas informáticas y el incremento en la capacidad computacional, brinda actualmente la posibilidad de implementar un abordaje desde la ciencia de datos a las problemáticas vinculadas con la salud pública. La ciencia de datos en salud es un área de investigación emergente que promete contribuir al diseño de políticas públicas [2,12,15], describiendo, monitoreando o creando sistemas de alertas para el seguimiento dinámico de indicadores,

de malnutrición en este caso, a distintas escalas. La base de datos de registros antropométricos del Programa SUMAR proporciona una oportunidad para el desarrollo de métodos computacionales al servicio de la salud, permitiendo sugerir la implementación de mejoras que redunden en la mejora de la calidad de los indicadores de malnutrición de la población infanto-juvenil más vulnerable del país. La herramienta planteada en este trabajo podría ser útil para otras instituciones así como en futuros estudios.

Referencias

1. Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics* **15**(4), 603–619 (2014)
2. Bellazzi, R.: Big data and biomedical informatics: a challenging opportunity. *Yearbook of medical informatics* **9**(1), 8 (2014)
3. Black, R.E., Victora, C.G., Walker, S.P., Bhutta, Z.A., Christian, P., De Onis, M., Ezzati, M., Grantham-McGregor, S., Katz, J., Martorell, R., Others: Maternal and child undernutrition and overweight in low-income and middle-income countries. *The lancet* **382**(9890), 427–451 (2013)
4. 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)
5. den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., Herbst, K.: Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLOS Medicine* **2**(10) (2005). <https://doi.org/10.1371/journal.pmed.0020267>, <https://doi.org/10.1371/journal.pmed.0020267>
6. Cole, T.J., Donaldson, M.D.C., Ben-Shlomo, Y.: SITAR—a useful instrument for growth curve analysis. *International journal of epidemiology* **39**(6), 1558–1566 (2010)
7. 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)
8. De, E.E., Pierson, D.: Data handling: cleaning and quality control. NETLAKE toolbox for the analysis of high-frequency data from lakes (Factsheet 1) Technical report NETLAKE COST Action ES1201 pp. 2–6 (2016)
9. De Onis, M., Onyango, A.W., Borghi, E., Garza, C., Yang, H., Group, W.H.O.M.G.R.S., Others: 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)
10. Grantham-McGregor, S., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., Group, I.C.D.S., Others: Developmental potential in the first 5 years for children in developing countries. *The lancet* **369**(9555), 60–70 (2007)
11. 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)
12. Khoury, M.J., Ioannidis, J.P.A.: Big data meets public health. *Science* **346**(6213), 1054–1055 (2014)

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. M. Sabignoso, L. Zanazzi, S. Sparkes, I.M.: Strengthening the purchasing function through results-based financing in a federal setting: lessons from argentina’s programa sumar. Tech. rep., World Health Organization (2020)
15. Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. *Jama* **309**(13), 1351–1352 (2013)
16. 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)
17. Nuñez, P.A., Fernández-Slezak, D., Farall, A., Szretter, M.E., Salomón, O.D., Vallenggia, C.R.: Impact of universal health coverage on child growth and nutrition in Argentina. *American journal of public health* **106**(4), 720–726 (2016)
18. de Onis, M., 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**, 660–667 (2007)
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? *Frontiers in psychology* **4**, 370 (2013)
21. Pelletier, D.L., Frongillo, E.A.: Changes in child survival are strongly associated with changes in malnutrition in developing countries. *The Journal of nutrition* **133**(1), 107–119 (2003)
22. Prendergast, A.J., Humphrey, J.H.: The stunting syndrome in developing countries. *Paediatrics and international child health* **34**(4), 250–265 (2014)
23. 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)
24. SAP: Guía para la evaluación del crecimiento físico. Buenos Aires: Sociedad Argentina de Pediatría (2013)
25. 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)
26. Victora, C.G., Adair, L., Fall, C., Hallal, P.C., Martorell, R., Richter, L., Sachdev, H.S., Maternal, Group, C.U.S., Others: Maternal and child undernutrition: consequences for adult health and human capital. *The lancet* **371**(9609), 340–357 (2008)
27. Welch, C., Petersen, I., Walters, K., Morris, R.W., Nazareth, I., Kalitzi, E., White, I.R., Marston, L., Carpenter, J.: Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. *Pharmacoepidemiology and drug safety* **21**(7), 725–732 (2012)
28. WHO: Physical status: The use of and interpretation of anthropometry. Report of a WHO Expert Committee p. 452. (1995)
29. WHO: 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)
30. WHO: Acabar con la obesidad infantil. Tech. rep., Organización Mundial de la Salud (2016)

14 MS Fernández et al.

31. Yang, S., Hutcheon, J.A.: Identifying outliers and implausible values in growth trajectory data. *Annals of epidemiology* **26**(1), 77–80 (2016)
32. Zhang, A., Song, S., Wang, J., Yu, P.S.: Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment* **10**(10), 1046–1057 (2017)