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

Agustín Nicolás Dramis[1,2], María Soledad Fernández[1,2], Adriana Alicia Pérez[1] y Pablo Guillermo Turjanski[1,2]

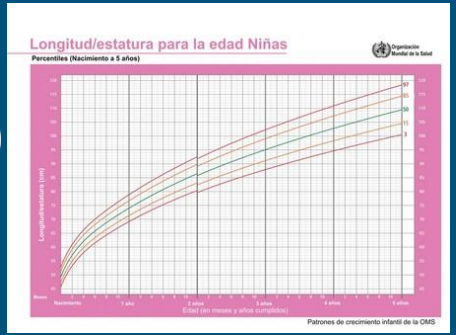
[1]Grupo de Bioestadística Aplicada, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina (GBA, FCEyN-UBA)
[2]Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina(CONICET)

Introducción

Talla

- Medida antropométrica
- Modelado del crecimiento
- Evaluación del estado nutricional (individual, poblacional)

↳ Comparación de la talla según estándares de crecimiento para la edad y sexo -> cálculo de *puntajes-z* (‘talla para la edad’)



Bases de datos nutricionales

GRAN VOLUMEN DE DATOS

- Diseño de políticas Públicas
- Evaluación de progresos y áreas que requieren mejoras
- Seguimiento dinámico de indicadores
- Orientación de decisiones en base a la evidencia.

Bases de datos nutricionales

GRAN VOLUMEN DE DATOS

- Diseño de políticas Públicas
- Evaluación de progresos y áreas que requieren mejoras
- Seguimiento dinámico de indicadores
- Orientación de decisiones en base a la evidencia

❏ Carga Manual de datos
❏ Errores de distinta naturaleza

Limpieza de medidas anómalas

"Proceso de detección, diagnóstico y edición de datos defectuosos"

Van den Broeck et al (2005): Data cleaning: detecting, diagnosing, and editing data abnormalities

Limpieza de medidas anómalas

- Pautas de la OMS para datos no plausibles (WHO 2006, De Onis 2006, 2007). *puntaje-z* menor a -6 o mayor a 6 (talla para la edad)

Limpieza de medidas anómalas

- Pautas de la OMS para datos no plausibles (WHO 2006, De Onis 2006, 2007). *puntaje-z* menor a -6 o mayor a 6 (talla para la edad)

- Para datos longitudinales: pautas de la OMS no son suficientes. Presencia de *puntajes-z* plausibles individualmente, pero inconsistentes en el tiempo si se considera su pertenencia a un mismo individuo. Se desarrolló un algoritmo que elimina datos a fin de resolver estas inconsistencias

Objetivo del trabajo

Cuantificar el impacto de cuatro tipos de errores habituales durante la carga manual de datos sobre la proporción de individuos con baja talla para la edad.

Evaluar el desempeño de dos herramientas de limpieza complementarias, aplicadas secuencialmente para identificar dichos errores y preservar datos correctos.

Metodología

Base Inicial

- Se simularon 10.000 individuos con 10 registros de talla a distintas edades.
- Se respetó la distribución de edades y sexos de una base de datos real (Programa SUMAR)
- A cada individuo se le asignó un puntaje-z constante (con un 10% al menos de individuos con malnutrición) a lo largo de sus mediciones, para cada una de las cuales se determinó la talla correspondiente.

Metodología

Base Inicial

Bases Modificadas

- Se simularon 10.000 individuos con 10 registros de talla a distintas edades.
- De esta base se generaron 4 copias, y a cada una se le aplicó un tipo de error habitual en procesos de carga manual de datos numéricos.

Metodología

Base Inicial

Bases Modificadas

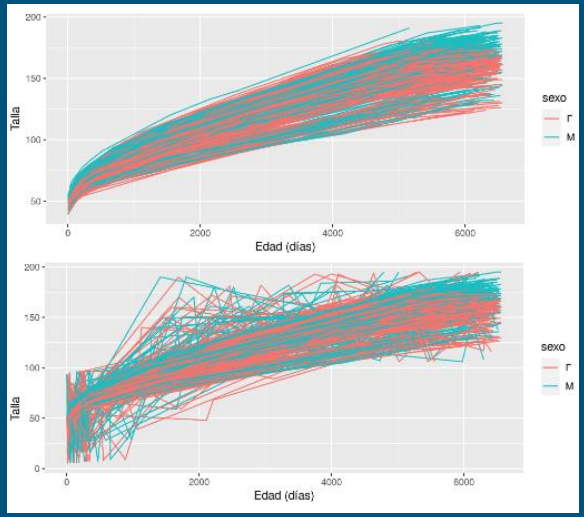
			Ejemplo numérico	
Nombre de la base	Nombre de error	Descripción	Valor original (cm)	Valor modificado (cm)
BMa	Ea	Anagrama. Se reemplazó el número por un anagrama, es decir un reordenamiento aleatorio de sus cifras	123	132
BMb	Eb	Variación de una cifra por uno. Se cambió el valor de una de sus cifras 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

- Se simularon 10.000 individuos con 10 registros de talla a distintas edades.
- De esta base se generaron 4 copias, y a cada una se le aplicó un tipo de error habitual en procesos de carga manual de datos numéricos.

Metodología

Base Inicial

Bases Modificadas



- Se simularon 10.000 individuos con 10 registros de talla a distintas edades.
- De esta base se generaron 4 copias, y a cada una se le aplicó un tipo de error habitual en procesos de carga manual de datos numéricos.
- Para cada una de ellas, el 5% de los registros fue aleatoriamente designado registro erróneo.
- A estos registros se les introdujo el error designado en la variable “talla”.
- En estas bases se aplicaron secuencialmente dos herramientas de limpieza.

Metodología

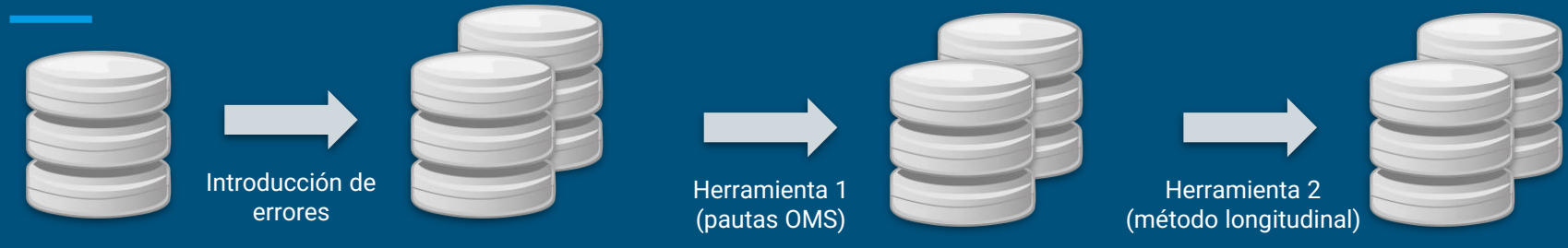
Base Inicial

Bases Modificadas

Bases Post H1

- La H1 consistió en la aplicación de las pautas de la OMS para la remoción de datos considerados biológicamente no plausibles (valores de puntaje-z superiores a 6 o inferiores a -6).

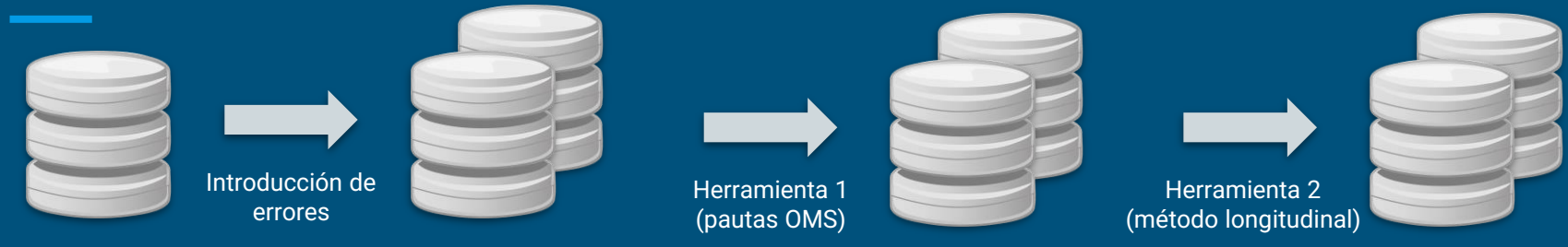
Metodología



- La H1 consistió en la aplicación de las pautas de la OMS para la remoción de datos considerados biológicamente no plausibles (valores de puntaje-z superiores a 6 o inferiores a -6).
- Es un método transversal que considera a cada registro de forma aislada.
- La H2 consistió en la aplicación de un método de desarrollo propio*, que contempla la plausibilidad de los cambios entre registros de un mismo individuo.

* 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).

Metodología



- A cada una de las bases se la determinó la prevalencia de baja talla para la edad, calculada como la proporción de individuos con al menos un registro con puntaje-z < -2.
- A cada una de las herramientas de limpieza se le determinó su sensibilidad (# registros erróneos eliminados / # registros erróneos totales) y su especificidad (# registros sin errores retenidos/ # registros sin errores totales), para cada uno de los tipos de error introducido.

Resultados - Prevalencia de baja talla

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 %

- La prevalencia de baja talla aumentó entre 13,22 y 27,22 puntos porcentuales según tipo de error
- Aplicar las herramientas de limpieza resultó en una baja de la prevalencia, acercándose a una distancia de entre 0,12 y 3,27 puntos porcentuales de la inicial.
- Si bien el mayor impacto lo tuvo la H1, la H2 produjo un acercamiento adicional de entre 0,19 y 0,53 puntos porcentuales.

Resultados - Desempeño de la herramienta

	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 %

- La sensibilidad del protocolo de limpieza fue de entre 44,49% y 83,56%, aumentando siempre al incorporar la H2.
- La especificidad del protocolo fue alta (entre 99,80% y 99,89%), disminuyendo ligeramente al incorporar la H2.

Discusión

- La prevalencia de baja talla aumentó de manera consistente al introducir errores.

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 %

	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 %

Discusión

- La prevalencia de baja talla aumentó de manera consistente al introducir errores.
- Se observó un mayor aumento de la prevalencia al errar una cifra de la talla por uno.

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 %

	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 %

Discusión

- La prevalencia de baja talla aumentó de manera consistente al introducir errores.
- Se observó un mayor aumento de la prevalencia al una cifra de la talla por uno.
- La H1 eliminó una mayor proporción de errores.

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 %

	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 %

Discusión

- La prevalencia de baja talla aumentó de manera consistente al introducir errores.
- Se observó un mayor aumento de la prevalencia al una cifra de la talla por uno.
- La H1 eliminó una mayor proporción de errores.
- Se distinguieron dos tipos de errores en base a su acción sobre distinta cantidad de cifras.

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 %

	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 %

Conclusión

El trabajo muestra el impacto de la introducción de errores habituales de carga manual en registros de talla, generando amplias variaciones en la prevalencia de un indicador de relevancia en salud como la baja talla.

Muestra la importancia de realizar un proceso de limpieza de bases de datos previo al análisis, y el potencial de aporte de las herramientas longitudinales.

Respecto a estos métodos, trabajos futuros se centrarán en optimizar los criterios de éste método, y el análisis de errores individualmente plausibles, y sobre otras variables.

Muchas gracias