Gac Sanit. 2025;39(S2):734

Gaceta Sanitaria

https://www.gacetasanitaria.org

809 - ESTIMATE THE RECORD LINKAGE SPECIFICITY USING INDIRECT IDENTIFIERS – A BIRTH COHORT GERAÇÃO XXI

I. Viseu, P. Meireles, M. Severo

ISPUP; ITR.

Resumen

Background/Objectives: The use of healthcare data in research requires balancing utility and privacy, often relying on pseudonymization for record linkage. A key challenge is ensuring accurate identifiers, as missing or inaccurate data can lead to false or missed matches. While careful design helps minimize errors, they remain inevitable, with short codes increasing the risk of false matches. Even small linkage errors can introduce bias, making quality assessment crucial. Studies show that adding more indirect identifiers, like demographic data and residence codes, improves specificity but reduces sensitivity. However, the influence of alternative identifiers, such as name initials and sociodemographic information, on linkage accuracy remains unexplored. This study aimed to assess the validity of matching algorithms using indirect identifiers, including participant's sex, birth date, the first letter of the mother's name, the first letter of the participant's name, and maternal education, comparing their performance against a gold standard.

Methods: For this study we used data from 8619 participants from the Generation XXI birth cohort –a population-based cohort of newborns from public maternity wards in Porto between 2005 and 2006. Participants were identified with a unique ID (gold standard), and we pseudonymized records based on sex, birth date, and initials of the mother's and participant's first names. A deterministic linkage method was used to link records of each strategy –one using the pseudonymized code and a second strategy using the pseudonymized code and the maternal education Linkage accuracy was assessed by comparing the generated codes to the gold standard, evaluating false matches, specificity, positive predictive value (PPV), and Cohen's Kappa. The analysis assumed 100% sensitivity.

Results: The first strategy identified 458 false matches, with 94.95% (95%CI: 94.48, 95.40) PPV and a specificity of 99.9988% (95%CI: 99.99, 99.99). Adding maternal education (strategy 2) improved PPV to 99.00% (95%CI: 98.77, 99.20) and agreement from 0.97 to 0.99.

Conclusions/Recommendations: This study found that combining the first letters of the mother's and child's first names, sex, and the child's full birth date resulted in high specificity, but also a 5% false match. Adding the mother's education improved both PPV and agreement, reducing the false match to 1%. These results suggest that adding stable socio-demographic information can be used to improve record linkage, particularly in healthcare contexts.

Funding: FCT: UIBD/04750/2020 and LA/P/0064/2020 (https://doi.org/10.54499/UIDB/04750/2020 and https://doi.org/10.54499/LA/P/0064/2020). Inês Viseu: PhD grants UI/BD/154374/2022 funded by the FCT.