Gac Sanit. 2026;40(S4):685

Gaceta Sanitaria

https://www.gacetasanitaria.org

736 - LINKING BRANDED FOOD DATABASES: AN EMPIRICAL STUDY COMPARING RULE-BASED, SUPERVISED AND PROBABILISTIC APPROACHES

A. Vicente, D. Lopes, I. Castela, I. Figueira, J. Carriço, M. Figueira, M.J. Gregório

Local Health Unit of Vila Nova de Gaia/Espinho; National Programme for the Promotion of Healthy Eating, Directorate-General of Health; EPIUnit-Institute of Public Health of University of Porto; Faculty of Nutrition and Food Sciences of the University of Porto.

Resumen

Background/Objectives: Branded food databases support the monitoring and evaluation of nutrition policies yet record linkage across time remains challenging due to rapid changes in the food supply. Although deterministic matching based on Global Trade Item Numbers is generally considered the most reliable, it may be insufficient when the same food products attributes change over time highlighting the need for complementary probabilistic and fuzzy matching methods. This study aims to empirically compare record linkage strategies between 2022 and 2024 to identify a robust, scalable approach for branded food product matching.

Methods: Data were collected as part of the EU JA Best-ReMaP (2022) and the EU JA PreventNCD (2024). Two food categories-Breakfast Cereals (BC n = 615) and Fresh Dairy Products and Desserts (FDPD n = 1,768) -were included using a hybrid deterministic-fuzzy pipeline. Exact matching on GTIN and semantic-key were applied first. Remaining records (BC n = 270; FDPD n = 888) were processed using three rule-based fuzzy configurations (v1-v3) differing in weight assignments, use of gatekeepers (v2), and without quantity consideration (v3); a supervised logistic regression model (v4) and an unsupervised probabilistic linkage model (v5). Models produced normalized match scores and categorical decisions (manual review or no-match). Outputs were expert-validated to establish ground truth.

Results: The conservative rule-based model v1 achieved the highest ROC-AUC (BC 0.95; FDPD 0.80), indicating strong global discrimination, but generated many false positives (precision BC 0.53; FDPD 0.14). Model v2 provided the most operationally viable trade-off, with high sensitivity (BC ?1; FDPD 0.86), improved precision (BC 0.79; FDPD 0.16), and a reduction in records requiring manual review compared to v1 (-21% in BC and -23% in FDPD). Model v3 had similar ROC-AUC but lower F1 (BC 0.67; FDPD 0.20). Model v4 performed poorly, with low sensitivity (BC 0.18; FDPD 0), while model v5 had unstable precision and very low F1 (BC 0.09; FDPD 0.11). Overall, F1 scores remained low due to extreme class imbalance but v2 consistently achieved the highest.

Conclusions/Recommendations: While v1 maximized sensitivity, the constrained rule-based fuzzy model v2 offered a more viable balance between sensitivity and precision, substantially reducing human validation burden while preserving matching quality. These findings support transparent hybrid deterministic-fuzzy pipelines as practical and generalizable solutions for large-scale branded food product linkage.