Table S1: Details of deleted attributes from the mortality dataset with HealthDataPrep

Activity	Details
DC1. Deduplication	Deleted attributes: *NATURAL, DTNASC, ESC, CAUSABAS_O* AND *ESCFALAGR1.* The data from these attributes can be extracted from corresponding attributes, such as *DTNASC* derived from *AGE*, educational level (*ESC and ESCFALAGR1*) from *ESC2010*, cause of death (*CAUSABAS_O*) from *CAUSABAS*, and the NATURAL from CODMUNNATU.
DC2. Exclusion of features with zero variance	Deleted attributes: TIPOBITO, CODIFICADO, ALTCAUSA, ASSISTMED
DC3. Handling of low-variance and low-cardinality attributes	Deleted attributes: *ORIGEM, STDOEPIDEM* AND *STDONOVA*. *STDOEPIDM*" (epidemiological death declaration status) overwhelmingly contained "No" (99.999%) with negligible "Yes" (0.001%), leading to low variance. ORIGEM (Record source) overwhelmingly contained "Ignored" (99.999%) with negligible "Others" (0.001%), leading to low variance. STDONOVA (Death Certificate Status) overwhelmingly contained "No" (99.999%) with negligible "Yes" (0.001\%), leading to low variance.
DC4. Missing data handling	Deleted attributes: *IDADEMAE, ESCMAE, ESCMAE2010, SERIESCMAE, OCUPMAE, QTDFILVIVO, QTDFILMORT, GRAVIDEZ, SEMAGESTAC, GESTACAO, PARTO, OBITOPARTO, PESO, CB_PRE, ACIDTRAB, FONTE, CAUSAMAT, ESCMAEAGR1, TPRESGINFO, DTCADINF, MORTEPARTO* AND *DTCONCASO*
DC5. Inconsistent data handling	Deleted attributes: *SERIESCFAL, TPMORTEOCO, OBITOGRAV, OBITOPUERP, EXAME, CIRURGIA, COMUNSVOIM, CIRCOBITO, DTINVESTIG, FONTEINV, NUDIASOBCO, DTCADINV, TPOBITOCOR, DTCONINV, FONTES, TPNIVELINV* Attributes with more than 75% of their values as NaN or “Ignored”
FE1.2. Manual feature selection 19	Delete temporal attributes: *DTOBITO, HORAOBITO, DTATESTADO, DTCADASTRO, DTRECEBIM, DTRECORIGA, DIFDATA* Delete irrelevant attributes: *OCUP,CODESTAB, CODMUNNATU, LINHAII, TPPOS, ATESTANTE, STCODIFICA, VERSAOSCB, VERSAOSIST, ATESTADO, LOCOCOR,CODMUNOCOR*
DP1. De-identification and suppression 2	Deleted attributes: *CONTADOR, NUMEROLOTE*