


BMJ Open Internal deterministic record linkage using indirect identifiers for matching of same-patient hospital transfers and early readmissions after acute coronary syndrome in a nationwide hospital discharge database: a retrospective observational validation study

Afonso Rocha ^{1,2}, Luis Filipe Azevedo,³ J C Silva Cardoso,⁴ Thomas G Allison,⁵ Alberto Freitas⁶

To cite: Rocha A, Azevedo LF, Silva Cardoso JC, *et al*. Internal deterministic record linkage using indirect identifiers for matching of same-patient hospital transfers and early readmissions after acute coronary syndrome in a nationwide hospital discharge database: a retrospective observational validation study. *BMJ Open* 2019;**9**:e033486. doi:10.1136/bmjopen-2019-033486

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-033486>).

Received 07 August 2019
Revised 28 November 2019
Accepted 05 December 2019



© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Afonso Rocha;
afonsomrocha@gmail.com

ABSTRACT

Objectives To assess validity of record linkage using multiple indirect personal identifiers to identify same-patient hospitalisations and definition of episode of care (EC) due to acute coronary syndrome (ACS).

Methods Using national hospital discharge data to identify all admissions due to ACS, we used six different linkage rules using indirect identifiers with increasing level of detail and compared validity against a pseudonymised unique identifier used as gold standard (GS). Contiguous hospitalisations within each matched group of hospitalizations occurring within 28 days of each other were considered one EC. We classified hospitalisations according to time between the first pair of hospitalisations as hospital transfer (HT: ≤ 1 day), early readmission (ER: 2–28 days) or recurrent cases (> 28 days).

Results There were 146 671 hospitalisations (unlinked), 121 987 ACS 28-day EC (linked GS), with 18 398 HTs (≤ 1 day), and 6286 ERs (≤ 28 days). Linkage rules using demographic and residence code variables produced linkage rates with highest validity for rule using sex, date of birth and four-digit residence code with sensitivity of 98.4 (95% CI: 98.4 to 98.5); specificity of 97.8 (95% CI: 97.6 to 98.0) and Cohen's κ of 0.9 to detect ACS-EC, compared with GS linkage rule. Similarly, validity for HT and ER was high and of similar magnitude, with sensitivity ranging between 97.2% and 98.1%, and specificity between 98.8% and 99.9%, respectively.

Conclusions Our internal linkage validation study using indirect patient identifiers will allow calibration of incidence rates and performance indicators, accounting for the effect of HT and readmissions.

BACKGROUND

Hospital administrative data provide a valuable source of information to address health-care management, resource utilisation and quality of care research questions. Strong

Strengths and limitations of this study

- Demonstrates the validity of using deterministic record linkage with indirect identifiers to link patient-level hospitalisations allowing for aggregation of hospitalisations within the same episode of care.
- Shows a valid method to overcome limitations of anonymised large administrative databases, allowing for epidemiological research with retrospective analysis and calibration of past and present ACS hospitalisation incidence trends, in-hospital mortality and performance indicators.
- This methodology is applicable in different countries and settings having high rates of hospital transfers and readmissions, such as trauma, stroke and intensive care patients.
- There was no assessment of quality of gold standard used for validation (unique pseudonymised identifier).
- Validation was done on the National Hospital Discharge Database which has very low missing/invalid rates, whereby it may not be applicable in databases with higher error rates and for external record linkage between different data sets, where stringent deterministic methods result in a high number of false negatives.

points of these databases are very wide coverage and low-cost systematic data collection.¹ On the downside, hospital administrative data are not designed for research purposes, often lack unique patient identifier and pertain to each hospitalisation not allowing linkage of multiple hospitalisations within the same episode of care (EC), being thereby susceptible to imprecisions and overestimation when patients are transferred

between hospitals or have multiple readmissions for a single EC.² This is especially problematic in the case of acute coronary syndromes (ACS) where clinical pathways and referral networks have been implemented to assure timely access to coronary angiography and revascularisation procedures, with hospital transfer (HT) rates up to 30%.^{3,4}

Identifying whether a hospital admission is a transfer from another hospital, an early readmission (ER) within the same EC, or a late readmission due to a new ACS event remains challenging and is of paramount importance for analysing and interpreting outcome data and for monitoring trends of ACS subtypes, therapeutic measures and healthcare services performance.⁵ Additionally in the US, from 2012 onwards, hospitals in which 30-day hospital readmission rates for certain conditions, including acute myocardial infarction, exceed the national average are financially penalised under the Patient Protection and Affordable Care Act.⁶

A standard approach to minimise multiple counting has been to exclude inter-HTs and readmissions but, since these are not random events, this method introduces bias and leads to loss of relevant information.^{7,8} On the other hand, treating sequential hospitalisations as independent EC results in overestimation of standardised ACS trends, lowers estimates of the proportion of patients submitted revascularisation treatment and may artificially decrease in-hospital mortality rates.^{4,5,7,9} Therefore, sequential hospitalisations for the same patient, occurring within a preset time frame, should be combined as one EC as this should be considered the preferred unit of analysis. When only unlinked data is available and there is no unique patient identifier, using an internal linkage method through demographic and event-based variables is desirable to identify and account for HTs and readmissions within the same EC.¹⁰

We aimed to build and assess the validity of a matching algorithm using secondary non-unique patient identifiers and event-based variables, using a stepwise deterministic linkage method, to identify patient-level ACS hospitalisations and contiguous hospitalisations occurring within 28 days from each other, classified as one ACS-EC, by using pseudonymised data (unique direct identifier) as gold standard (GS).

METHODS

Study population and data sources

Data for the study were obtained retrospectively from the administrative national hospital discharge database provided by the Portuguese Ministry of Health's Central Administration for the Health System which includes hospitalisations occurring in all public acute care hospitals of the Portuguese National Health Service in mainland Portugal. Data providing is mandatory for every hospitalisation and used for hospital's reimbursement purposes, but also for disease prevalence estimation and healthcare utilisation assessment. Collected information

includes demographics (age, sex, residence code), hospital admission and discharge dates, discharge diagnosis in a principal diagnosis field and up to 30 secondary diagnosis fields using the International Classification of Diseases—ninth revision—clinical modification (ICD9-CM) and discharge status (deceased or alive).

Due to data privacy issues, administrative health data has traditionally been released to researchers without unique direct identifiers. From 2011 onwards, a pseudonymised unique patient identifier was provided, allowing to track same patient hospitalisations against which we aimed to assess and validate our matching algorithm. Therefore, our analysis was restricted to all hospitalisation episodes, both inpatient and outpatient, between 2011 and 2015. We followed the modified Standards for Reporting of Diagnostic Accuracy criteria to report our findings.¹¹

Event identification and classification

Coding procedures for ACS-EC vary considerably between institutions and with time, especially in the case of HTs for specialised care and treatment, ranging from both institutions (referring and receiving) coding the ACS hospitalisation and the procedure (duplicating both counts) to only the receiving institution coding the hospitalisation episode and procedure either as an inpatient or outpatient code. To capture all information pertaining to each ACS-EC, we included all hospitalisation episodes, both inpatient and outpatient, with a primary discharge diagnosis field showing ICD9-CM codes 410.x, 411.0–411.1 and 414.x and procedural codes: cardiac catheterisation (37.21, 37.22, 37.23), percutaneous coronary intervention (00.66, 36.03, 36.04; 36.06, 36.07, 36.09) and surgical coronary revascularisation (coronary artery bypass grafting (CABG): 36.10–36.17, 36.19). An exploratory analysis revealed heterogeneity of coding practices for HTs and elective readmissions among hospitals, ranging from both institutions coding admission with an ACS coding (410.x, 411.0–411.1), to first institution coding ACS and receiving institution coding hospitalisation with a 414.x code. Since we wanted to capture and aggregate all the information related to hospitalisations within each ACS-EC, including revascularisation procedures, we decided to use all codes (410.x, 411.0–411.1 and 414.x) and selected, for each matching rule, only episodes having, at least, one hospitalisation with a 410.x or 411.0–411.1 code.

Since there is no specific coding allowing identification of ACS subtypes, we used codes 410.0–410.6 and 410.8 for ST-segment elevation myocardial infarction (STEMI), codes 410.7 and 410.9 for non-STEMI (NSTEMI) and code 411.0 and 411.1 for unstable angina.¹² We used the diagnostic hierarchy method proposed by Lopez *et al* which reflects the severity of ACS subtypes, from STEMI (most severe), over NSTEMI to unstable angina (less severe). For an ACS-EC with multiple hospitalisations, the most severe category was used.²

The steps taken to select linkable inpatient and outpatient hospitalisation episodes with an ACS-related primary

diagnosis is shown in online supplementary figure 1. First, we identified redundant episodes (n=21) that had the same combination of values for all variables (60 variables), and kept only one record among duplicates. Second, we excluded records with missing or invalid values in the linking variables contained in each linkage rule (n=1095). Lastly, we restricted hospitalisation episodes to patients aged ≥ 30 years (n=171) due to concerns of unreliability of ACS estimates in younger patients.

Linkage method

We used internal deterministic data linkage requiring matches on different combinations of person-level identifiers and calculated time interval (in days) between matched hospitalisations to define a 28-day ACS-EC comprising first admission and all contiguous admissions occurring within 28-day period from each other (HT: ≤ 1 day; ER: 2–7 days; late readmission: 8–28 days).^{2 13} Cases with identical demographic identifiers (matched hospitalisations) admitted to the same hospital or in two separate hospitals within 28 days of each other were considered as belonging to the same 28-day ACS-EC, counted only once and had all their information aggregated. Matched hospitalisations occurring beyond 28 days from each other were considered as a new ACS-EC.

We set six test linkage rules using various combinations and granularity of the following linkage variables: sex; date of birth and residence code. Deterministic linkage rules require, for identification of matched hospitalisations (hospitalisations pertaining to the same patient) an exact match on values of all linkage variables specified on each matching the rules (table 1). Residence code consists of a sequential combination of six digits according to the administrative level of detail: two identifying districts (total of 18); two for municipalities within each district (total of 278) and two for parishes within each district and municipality (4050 up to 2012; and 2882 after the administrative reform of 2013).¹⁴ A direct pseudonymised identifier (unique patient ID nine-digit combination derived from national identification number) was used

as the GS.¹⁵ We sequentially tested rules with increasing level of granularity to assess validity and linkage error rate compared with the GS.

Despite contiguous admissions, within 28 days from each other, being counted only once as a single 28-day ACS-EC, we aggregated information from different hospitalisations regarding revascularisation procedures, severity indicators, comorbidities and in-hospital mortality.

Statistical analysis

For each matching rule we calculated total matching rate (proportion of total hospitalisations successfully linked) and matching rate for ACS-EC. Using unique ID as GS we calculated the number of matching errors (missed matches; false matches) for each matching rule. Comparative linkage quality was assessed by calculating sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) with their 95% CIs using the one-sample Clopper–Pearson and the standard logit methods, respectively.^{16 17} Chance-weighted proportional agreement between matching rules and GS was calculated using Cohen's κ and classified as poor if $\kappa \leq 0.20$; fair if $0.21 < \kappa \leq 0.40$; moderate if $0.41 < \kappa \leq 0.60$; high $0.61 < \kappa \leq 0.80$ and excellent agreement if $\kappa > 0.80$.¹⁸ We compared baseline characteristics between true matches and false matches and between missed matches and true non-matches, using independent samples t-test and Chi-square for continuous and categorical variables, respectively. We then described the characteristics of the study population with a single hospital admission compared with those having multiple hospitalisations within a 28-day ACS-EC. Analyses were performed using IBM SPSS Statistics V.25 and Microsoft Excel V.16.30.

Patient and public involvement

There was no patient or public involvement in any step of this study.

RESULTS

During the study period there were 146671 hospitalisations due to ACS with mean age 67.7 (12.3) years and 68.9% were men. Median length of stay was 3 days (IQR 6). Unlinked data revealed 26842 (18.3%) hospitalisations with STEMI, 36597 (24.9%) with NSTEMI, 10347 (7.5%) with unstable angina and 72885 (49.6%) classified as other acute and subacute forms of ischaemic heart disease. Cardiac catheterisation was performed in 70% of hospitalisation episodes, percutaneous coronary intervention in 38.2% and CABG in 6.3%, while in 23.8% of hospitalisations no cardiac procedure was performed. Heart failure was present in 19699 (13.4%), 15019 (10.2%) had atrial fibrillation, 2981 (2.0%) ventricular fibrillation, along with 1469 (1.0%) cardiac arrests and a total of 6241 (4.3%) in-hospital deaths (online supplementary table 1).

The linkage rule requiring an exact match on the unique patient ID (GS) identified 34948 matched

Table 1 Stepwise deterministic matching algorithms according to detail of identifying variables

Matching rules	Matching variables algorithm
Rule 1	Sex, YearBirth
Rule 2	Sex, YearBirth, MonthBirth
Rule 3	Sex, YearBirth, MonthBirth, DayBirth
Rule 4	Sex, YearBirth, MonthBirth, DayBirth, ResidCode-2digits
Rule 5	Sex, YearBirth, MonthBirth, DayBirth, ResidCode-4digits
Rule 6	Sex, YearBirth, MonthBirth, DayBirth, ResidCode-6digits
Gold standard	Unique patient ID

ID, identification number; ResidCode, residence code.

**Table 2** Total number and proportion of matched hospitalisations and 28-day ACS episode of care using each matching rule

Matching rules	Number of matched hospitalisations	% of matched hospitalisations	Number of same-patient contiguous hospitalisations	% of total HE identified as same-patient contiguous hospitalisations*
Rule 1	146 518	99.9	145 909	99.5
Rule 2	144 978	98.8	126 472	86.2
Rule 3	113 789	77.6	36 113	24.6
Rule 4	62 399	42.5	27 391	18.7
Rule 5	47 923	32.7	26 064	17.8
Rule 6†	40 909	28.1	22 290	15.3
Gold standard	34 948	23.8	24 684	16.8

*Contiguous corresponds to sum of HE identified as hospital transfer, early readmission or late readmission.

†exclusion of 1151 invalid/missing fifth/sixth digit residence code (leaving a total of 139 863 hospitalisation episodes). ACS, acute coronary syndrome; HE, hospitalisation episode.

hospitalisations corresponding to 23.8% of all hospitalisations, with 16.8% readmissions within 28 days from initial hospitalisation. Among the test rules based on indirect identifiers, matching rate decreased from 99.9% for rule 1 to 28.1% for rule 6, and the matching rate of ACS-EC with multiple hospitalisations decreased from 99.5% to 15.3%, from rules 1 to 6 (table 2).

The proportion of ACS-EC with multiple hospitalisations increased from 16.9% in 2011 to 17.9% in 2015, being less frequent in women and with advancing age compared with single hospitalisation ACS-EC. The rate of multiple hospitalisations within same EC was lower for those with unstable angina, and higher in those submitted to cardiac procedures, especially CABG. There was considerable geographical heterogeneity in incidence of ACS hospitalisations and proportion of ACS-EC with multiple hospitalisations with major coastal districts (Lisbon and Porto) being responsible for 40.8% of all ACS-EC, but smaller inland districts depicting the highest

rate of ACS-EC with multiple hospitalisations ranging up to 37.4% (online supplementary table 2).

All test rules overestimated the number of recurrent ACS and underestimated first ACS hospitalisation compared with the GS. Rule 6 had the lowest detection of HTs (11.0% vs 12.5% for GS) but the second highest proportion of first ACS hospitalisation identification (71.9% vs 76.2% for GS) (table 3).

As level of detail of variables included in matching rules increased, the number of false matches decreased from 121 255 (82.7% of matches) for rule 1 to 1490 for rule 6 (3.6%) and, inversely, the proportion of missed matches increased from 0 (0.0%) for rule 1 to 3343 (8.1%) for rule 6. Furthermore, validity measures showed that adding residence code to demographic variables in matching rules significantly increased validity against the GS, with sensitivity decreasing only slightly from 100% for rule 1 to 97.8% for rule 5, with a steeper decrease to 86.2% for rule 6, with both specificity and PPV increasing as matching

Table 3 Total number and proportion of ACS hospitalisations according to time between first and subsequent hospital admission for matched hospitalisations using each matching rule

Linkage rules	First ACS hospitalisation*	Hospital transfers (≤1 day)	Early readmissions (>1 day and ≤7 days)	Late readmissions (>1 day and ≤28 days)	Recurrence (>28 days)
Rule 1	153 (0.1)	128 450 (87.6)	15 178 (10.3)	2281 (1.6)	609 (0.4)
Rule 2	1693 (1.2)	59 799 (40.8)	30 151 (20.6)	36 522 (24.9)	18 506 (12.6)
Rule 3	32 882 (22.4)	22 141 (15.1)	4123 (2.8)	9849 (6.7)	77 676 (53.0)
Rule 4	84 272 (57.5)	20 079 (13.7)	2482 (1.7)	4830 (3.3)	35 008 (23.9)
Rule 5	98 748 (67.3)	19 855 (13.5)	2055 (1.4)	4154 (2.8)	21 859 (14.9)
Rule 6†	104 442 (71.9)	16 036 (11.0)	2216 (1.5)	4038 (2.8)	18 619 (12.8)
Gold standard	111 723 (76.2)	18 398 (12.5)	2243 (1.5)	4043 ^{2,8}	10 264 (7.0)

*Includes both non-matched ACS hospitalisation (single hospital admission) and first hospitalisation in ACS episodes of care with multiple hospitalisations.

†Exclusion of 1151 invalid/missing fifth/sixth digit residence code (leaving a total of 139 863 hospitalisations). ACS, acute coronary syndrome.

rule granularity increases. Cohen's κ depicted an excellent agreement between rules using sex, date of birth and residence codes (linkage rules 4 to 6) and the GS for the detection of 28-day ACS-EC, with rule 5 showing the highest degree of agreement ($\kappa=0.941$), closely followed by rule 4 ($\kappa=0.927$), then decreasing for rule 6 ($\kappa=0.876$) (table 4). Table 5 shows the matching quality according to time between first and subsequent hospital admission for matched hospitalisations identified using matching rule 5, with somewhat lower PPV for HTs and late readmissions.

Using matching rule 5 to identify multiple hospitalisations within the same ACS-EC, 98.3% of episodes were correctly classified, while there was a false match rate of 7.3% and a false non-match rate of 0.4%. Table 6 compares the characteristics of ACS hospitalisation erroneously classified by rule 5 as a match (false match) or non-match (missed match) compared with true match and true non-matches, respectively. False match rate was more common in those presenting with unstable angina or coded as other ACS and subacute ACS (ICD9-CM 414) and in patients submitted to either cardiac catheterisation or percutaneous coronary intervention. Missed matches were more common in younger age, in hospitalisations coded as ICD9-CM 414 and in patients submitted to coronary artery bypass surgery (table 6). When analysing mismatch rates at district level, Lisbon had an exceedingly high proportion of false matches (21.8%) compared with the other districts (4.8%), with three municipalities alone being responsible for 77.2% of all false matches. Exclusion of these three municipalities from the analysis resulted in a drop in false-match rate in Lisbon district from 21.8% to 6.6%, approaching the district and national average.

Discussion

Using the National Hospital Discharge Database, we built, tested and compared the validity of deterministic internal record linkage using different combinations of indirect identifiers for the identification of 28-day EC consisting of patient-level sequential hospitalisations occurring within 28 days from each other. We found that linkage rules which include demographic and residence code variables showed comparable linkage rates and high validity compared with the GS. We found that false match rate was significantly reduced by increasing the level of detail of residence code, from district to municipality and to parish but, in case of inclusion of parish coding (rule 6), at the expense of an increase in missed matches, loss of sensitivity and agreement with the GS. To our knowledge this is the first study to validate a matching algorithm, without direct identifiers, for matching and identification of ACS patient-level hospitalisations, incorporating all subtypes of ACS and including a wider range of hospitalisations (eg, code 414.x) in order to capture and aggregate information from all sequential hospitalisations within the same ACS-EC, and to assess the impact of aggregating information on ACS hospitalisation

Table 4 Measures of matching quality for each matching rule in the detection of 28-day ACS episode of care

Linkage rules	Missed matches	False matches	Cohen's κ (95% CI)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive predictive value % (95% CI)
Rule 1	0	121 225	0.002 (0.000 to 0.004)	100.00 (99.9 to 100.0)	0.62 (0.58 to 0.67)	16.92 (16.91 to 16.92)
Rule 2	17	101 805	0.062 (0.059 to 0.065)	99.93 (99.90 to 99.96)	16.54 (16.34 to 16.75)	19.50 (19.46 to 19.54)
Rule 3	15	11 444	0.764 (0.760 to 0.768)	99.94 (99.91 to 99.97)	90.62 (90.45 to 90.78)	68.31 (67.93 to 68.69)
Rule 4	207	2914	0.927 (0.092 to 0.930)	99.16 (99.05 to 99.28)	97.61 (97.53 to 97.70)	89.36 (89.00 to 89.70)
Rule 5	542	1922	0.941 (0.939 to 0.944)	97.80 (97.62 to 97.99)	98.42 (98.35 to 98.49)	92.63 (92.30 to 92.94)
Rule 6*	3343	1490	0.876 (0.873 to 0.880)	86.15 (85.72 to 86.59)	98.77 (98.71 to 98.83)	93.32 (92.99 to 93.62)
Gold standard						

*Exclusion of 1151 invalid/missing fifth/sixth digit residence code (leaving a total of 139 863 hospitalisation episodes). ACS, acute coronary syndrome.

**Table 5** Measures of matching quality according to time between first and subsequent hospital admission for matched hospitalisations identified using matching rule 5

% (95% CI)	Primary ACS*	Hospital transfers	Early readmissions	Late readmissions
Sensitivity	98.44 (98.37 to 98.51)	97.21 (96.96 to 97.44)	98.13 (97.48 to 98.65)	98.14 (97.68 to 98.54)
Specificity	97.80 (97.61 to 97.98)	98.75 (98.69 to 98.81)	99.94 (99.92 to 99.95)	99.79 (99.77 to 99.82)
PPV	99.55 (99.51 to 99.59)	91.78 (91.40 to 92.14)	95.99 (95.12 to 96.70)	93.10 (92.33 to 93.80)
NPV	92.70 (92.40 to 93.00)	99.60 (99.56 to 99.63)	99.97 (99.96 to 99.98)	99.95 (99.93 to 99.96)
Cohen's κ	0.942 (0.940 to 0.944)	0.934 (0.933 to 0.939)	0.970 (0.965 to 0.975)	0.954 (0.950 to 0.959)

*Refers to ACS episodes of care (primary non-matched ACS plus matched hospitalisation episodes classified as recurrent >28 days). ACS, acute coronary syndrome; NPV, negative predictive value; PPV, positive predictive value.

counts, characterisation of ACS patients and on indicators of performance.

Most record linkage studies using indirect identifiers have been designed to externally link different data sets, namely clinical registries with claims data, whereby two records are considered a true match, given agreement or disagreement on a set of partial identifiers.^{19 20} For our study we took a different perspective, we aimed to internally link same-patient hospitalisation episodes due to ACS to build patient-level data on consecutive hospitalisations using event-based variables to define a time frame to build an EC. We chose deterministic linkage for its simplicity and appropriateness in scenarios in which missing and invalid values in matching variables are rare and these matching variables are sufficiently discriminative, as is often the case in large administrative data sets.²¹ By doing an analysis of different sets of identifiers against the GS, we demonstrated that combination of demographic and residence code (at district and municipality levels) variables showed the highest validity. Westfall and McGloin⁷ found similar results in a subanalysis of 120 206 myocardial infarction hospital admissions where they used a matching algorithm of indirect identifiers (age or month-year of birth, sex, zip code, ICD9 code) to detect HTs as same-patient hospitalisations occurring within 7 days of first hospitalisation, and found a sensitivity of 96.7% and specificity of 98.7%.

Choosing the appropriate matching rule is highly dependent on the aim of the analysis, and on the type, quality and completeness of data pertaining to the matching variables chosen.^{15 22} Moreover, use of a matching rule for record linkage should ideally be preceded by a pilot study, where validity against a given GS (usually a unique patient identifier) is assessed. In our study, we found that stricter residence code matching rule (six digits) resulted in a higher proportion of missed matches and loss of agreement with GS compared with more relaxed rules (four digits), possibly because it is more susceptible to coding errors, changes of residency and to administrative reforms with change in parishes' number and codes overtime.²³

Our matching algorithm with highest face validity (rule using sex, date of birth and four-digit residence code—rule 5) showed a low missed-match rate of 0.4% and

higher false-match rate of 7.4%. We found high regional heterogeneity with clustering of false matches in three municipalities within the same district. It possibly reflects regional variations in data quality, reporting or coding procedures,²⁴ and it reinforces the need for detailed analysis of characteristics associated with linkage error when doing validation studies for matching algorithms used in record linkage studies. We found missed matches to be more common in younger ages and those with planned procedures (ICD9-CM 414; surgical revascularisation); while false matches were more frequent in those with unstable angina and submitted to catheterisation and/or percutaneous revascularisation procedures. Nonetheless, these linkage errors had limited impact on the overall performance of the matching algorithm with specificity above 98% in detection of all contiguous hospitalisations' subtypes. Linkage methods that maximise specificity lead to the most robust study results and should therefore be the main focus when building matching rules for record linkage studies.²⁵

Our study has some limitations. Although we used a pseudonymised unique identifier as GS, it consists of a long string of numbers and is therefore susceptible to errors, the impact of which has not been assessed. In our study, we did an internal record linkage to identify patient-level contiguous hospitalisations, classified according to time elapsed between sequential hospitalisations, using indirect identifiers with low missing/invalid rates.

Different studies have compared linkage rates for a linkage rule using indirect identifiers with one using direct identifiers to link records from registries to Medicare claims data and showed, like we did in our study, highly valid linkages compared with the GS rule(s) that included direct identifiers.^{10 15} We used a deterministic linkage method and required exact matches on >3 variables in our rules. The expected error rates are low, and the rate for false-positive linkages is anticipated to be small. However, false-negative linkages are a concern in all rules, including the GS. The degree of bias from the imperfect GS depends on the number of false-negative matches in the GS and the prevalence of the true linkage.

Our results are likely generalisable to attempts that link hospitalisation-level records, but both expected error rates of linkage variables and prevalence of the condition

Table 6 Characteristics of matching errors of 28-day ACS-EC identified, using matching rule 5

	Matched as 28-day ACS-EC		P value	Non-matched 28-day ACS-EC		P value
	False match (n=1900)	True match (n=24 142)		Missed match (n=542)	True non-match (n=1 20 087)	
Sex			0.21			0.52
Male	1386 (7.4)	17 285 (92.6)		363 (0.4)	81 995 (99.6)	
Female	514 (7.0)	6857 (93.0)		179 (0.5)	38 092 (99.5)	
Age groups			0.42			0.02
30–44	74 (7.4)	930 (92.6)		33 (0.8)	4228 (99.2)	
45–54	251 (7.5)	3100 (92.5)		67 (0.5)	14 066 (99.5)	
55–64	464 (7.3)	5925 (92.7)		131 (0.5)	27 516 (99.5)	
65–74	593 (7.7)	7111 (92.3)		141 (0.4)	34 565 (99.6)	
75–84	431 (6.7)	5976 (93.3)		132 (0.4)	29 474 (99.6)	
85+	87 (7.3)	1100 (92.7)		38 (0.4)	10 238 (99.6)	
District level			<0.001			<0.001
Lisboa	892 (21.8)	3195 (78.2)		108 (0.9)	29 362 (99.6)	
Guarda	139 (13.0)	928 (87.0)		36 (0.3)	2297 (98.5)	
C. Branco	186 (8.9)	1911 (91.1)		15 (0.2)	4696 (99.7)	
Faro	31 (6.4)	452 (93.6)		13 (1.5)	4384 (99.7)	
Coimbra	37 (5.8)	603 (94.2)		11 (0.3)	6442 (99.8)	
Portalegre	44 (5.3)	792 (94.7)		17 (0.5)	1884 (99.1)	
Bragança	35 (5.1)	655 (94.9)		12 (0.3)	1841 (99.4)	
V Real	13 (4.8)	256 (95.2)		3 (0.1)	2249 (99.9)	
Viseu	19 (4.6)	394 (95.4)		4 (0.4)	3620 (99.9)	
Évora	11 (4.5)	235 (95.5)		9 (0.3)	2833 (99.7)	
Beja	40 (4.4)	876 (95.6)		16 (0.4)	5753 (99.7)	
Leiria	32 (4.4)	702 (95.6)		10 (1.0)	2244 (99.6)	
Braga	74 (4.1)	1733 (95.9)		74 (0.6)	7621 (99.0)	
Porto	157 (3.8)	3973 (96.2)		100 (0.5)	19 418 (99.5)	
Aveiro	64 (3.0)	2100 (97.0)		22 (0.4)	6174 (99.6)	
Setúbal	58 (2.6)	2181 (97.4)		53 (0.3)	10 577 (99.5)	
V Castelo	16 (2.3)	672 (97.7)		7 (0.1)	2774 (99.7)	
Santarém	52 (2.1)	2484 (97.9)		32 (0.5)	5918 (99.5)	
ACS subtypes			0.02			0.05
STEMI	165 (5.7)	2709 (94.3)		102 (0.4)	23 866 (99.6)	
NSTEMI	232 (4.8)	4588 (95.2)		123 (0.4)	31 654 (99.6)	
UA	108 (8.9)	1100 (91.1)		35 (0.4)	9104 (99.6)	
Cardiac procedures (aggregated)			<0.001			<0.001
No procedure	195 (4.8)	3872 (95.2)		102 (0.3)	30 672 (99.7)	
Catheterisation	673 (8.6)	7110 (91.4)		166 (0.4)	38 937 (99.6)	
PCI	866 (8.3)	9576 (91.7)		202 (0.4)	45 130 (99.6)	
CABG	166 (4.4)	3584 (95.6)		72 (1.3)	5348 (98.7)	
Comorbidity burden			0.05			0.65
Number of comorbidities						
0–3	1796 (7.4)	22 535 (92.6)		497 (0.5)	109 452 (99.5)	

Continued



Table 6 Continued

	Matched as 28-day ACS-EC		P value	Non-matched 28-day ACS-EC		P value
	False match (n=1900)	True match (n=24 142)		Missed match (n=542)	True non-match (n=1 20087)	
≥3	104 (6.1)	1607 (93.9)		45 (0.4)	10 635 (99.6)	
Charlson index			0.9			0.41
0–3	1716 (7.3)	21 825 (92.7)		478 (0.5)	104 462 (99.5)	
≥3	184 (7.4)	2317 (92.6)		64 (0.4)	15 625 (99.6)	

Results report absolute number and percentage for each variable category, unless stated otherwise.

.ACS, acute coronary syndrome; ACS-EC, acute coronary syndrome episode of care; CABG, coronary artery bypass grafting; ICD-9, International Classification of disease ninth revision; NSTEMI, non ST-segment elevation myocardial infarction; PCI, percutaneous coronary intervention; STEMI, ST-segment elevation myocardial infarction; UA, unstable angina.

should be considered. We have used standard demographic variables as linking variables, such as gender, date of birth and residence code, which are much less prone to errors and missing data, since most of these are automatically uploaded to the database. Nonetheless, our results may not be applicable in settings in which databases have high error rates in these linkage variables, since they will produce a large number of false-negative links warranting for the addition of probabilistic linkage methods.

CONCLUSION

Deterministic linkage using multiple indirect identifiers allows for accurate and valid internal linkage of patient-level contiguous hospitalisations in a preset time frame defining an EC, comparable with linkage with direct identifiers in hospital administrative data. Most data on nationwide or large-scale trends of ACS incidence, management and mortality have been abstracted from unlinked administrative health data and released to researchers without a unique patient identifier, and even in those jurisdictions that have recently introduced pseudonymised databases, longer-term trends analysis still relies heavily on unlinked records.²⁶ Therefore, our method of identifying, classifying and aggregating information of contiguous hospitalisations within the same EC will allow calibration of incidence rates and performance indicators to the number of EC and not to hospitalisations, and will be of value in different countries. Furthermore, it might also be useful in other clinical conditions that have high rates of transfers and readmissions, such as trauma,²⁷ stroke²⁸ and intensive care patients.²³

Author affiliations

¹Center for Health Technology and Services Research (CINTESIS), University of Porto-Faculty of Medicine, Porto, Portugal

²Cardiovascular Rehabilitation Unit, Physical Medicine and Rehabilitation, Centro Hospitalar Universitário Sao Joao EPE, Porto, Portugal

³Department of Health Information and Decision Sciences (CIDES) & Center for Health Technology and Services Research (CINTESIS), University of Porto-Faculty of Medicine, Porto, Portugal

⁴Department of Cardiology, Centro Hospitalar Universitário São João, University of Porto-Faculty of Medicine, Porto, Portugal

⁵Department of Cardiovascular Medicine and Cardiovascular Surgery, Mayo School of Medicine, Rochester, Minnesota, USA

⁶Department of Health Information and Decision Sciences (CIDES) & Center for Health Technology and Services Research (CINTESIS), University of Porto-Faculty of Medicine, Porto, Portugal

Contributors AR, LFA and AF conceived the study, developed the study design, made data analysis and interpretation and drafted the manuscript. LFA and AF provided statistical expertise for data analysis. JCSC and TGA critically revised the work for important intellectual content. All authors contributed to refinement of the study protocol and approved the final manuscript.

Funding This article was supported by National Funds through Fundação para a Ciência e a Tecnologia within CINTESIS, R&D Unit (reference UID/IC/4255/2019), and by project NORTE-01-0145-FEDER-000026—Symbiotic technology for societal efficiency gains: Deus ex Machina, financed by NORTE2020 under PORTUGAL2020.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethical approval was not required for the present study because it uses anonymised secondary data obtained during routine care, systematically reported by all public health hospitals in mainland Portugal and publicly available.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data source for this study is a retrospective administrative health database with data from all hospitalisations that were already anonymised. These are anonymised secondary data obtained during routine care, systematically reported by all public health hospitals in mainland Portugal and publicly available. Data were provided by the ACSS, an official organ of the Ministry of Health, through a cooperation protocol with the Faculty of Medicine, University of Porto and CINTESIS for research purposes.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Afonso Rocha <http://orcid.org/0000-0003-0824-4598>

REFERENCES

- Mazzali C, Paganoni AM, Ieva F, *et al*. Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Serv Res* 2016;16:234.
- Lopez D, Nedkoff L, Knuijan M, *et al*. Exploring the effects of hospital transfers and readmissions on trends in population counts of hospital

- admissions for coronary heart disease: a Western Australian data linkage study. *BMJ Open* 2017;7:e019226.
- 3 Ibáñez B, James S, Agewall S, *et al.* 2017 ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Rev Esp Cardiol* 2017;70.
 - 4 Khera R, Jain S, Pandey A, *et al.* Comparison of readmission rates after acute myocardial infarction in 3 patient age groups (18 to 44, 45 to 64, and ≥65 years) in the United States. *Am J Cardiol* 2017;120:1761–7.
 - 5 Fransoo R, Yogendran M, Olafson K, *et al.* Constructing episodes of inpatient care: data infrastructure for population-based research. *BMC Med Res Methodol* 2012;12:133.
 - 6 Services CfMM. *Readmissions reduction program*, 2012.
 - 7 Westfall JM, McGloin J. Impact of double counting and transfer bias on estimated rates and outcomes of acute myocardial infarction. *Med Care* 2001;39:459–68.
 - 8 Gurwitz JH, Goldberg RJ, Malmgren JA, *et al.* Hospital transfer of patients with acute myocardial infarction: the effects of age, race, and insurance type. *Am J Med* 2002;112:528–34.
 - 9 Insam C, Paccaud F, Marques-Vidal P. Trends in hospital discharges, management and in-hospital mortality from acute myocardial infarction in Switzerland between 1998 and 2008. *BMC Public Health* 2013;13:270.
 - 10 Bohensky MA, Jolley D, Sundararajan V, *et al.* Empirical aspects of linking intensive care registry data to hospital discharge data without the use of direct patient identifiers. *Anaesth Intensive Care* 2011;39:202–8.
 - 11 Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64:821–9.
 - 12 Yeh RW, Sidney S, Chandra M, *et al.* Population trends in the incidence and outcomes of acute myocardial infarction. *N Engl J Med* 2010;362:2155–65.
 - 13 Peng M, Li B, Southern DA, *et al.* Constructing episodes of inpatient care: how to define hospital transfer in hospital administrative health data? *Med Care* 2017;55:74–8.
 - 14 Número de municípios e freguesias no Continente e Regiões Autónomas a 31 Dezembro [Internet], 2017. Available: <https://www.google.pt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=2ahUKEwjM3qeA1aPdAhVHxokHTLjCJgQFjADegQICBAC&url=http%3A%2F%2Fsmi.ine.pt%2FVersao%2FDownload%2F59&usg=AOvVaw1jLFIIsOWX6VFlmPHyAdOB> [Accessed 5 Sep 2018].
 - 15 Setoguchi S, Zhu Y, Jalbert JJ, *et al.* Validity of deterministic record linkage using multiple indirect personal identifiers: linking a large registry to claims data. *Circ Cardiovasc Qual Outcomes* 2014;7:475–80.
 - 16 Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med* 1998;17:2635–50.
 - 17 Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med* 2007;26:2170–83.
 - 18 Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology* 2010;73:1167–79.
 - 19 Méray N, Reitsma JB, Ravelli ACJ, *et al.* Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol* 2007;60:883.e1–11.
 - 20 Roos LL, Wajda A, Nicol JP. The art and science of record linkage: methods that work with few identifiers. *Comput Biol Med* 1986;16:45–57.
 - 21 Li B, Quan H, Fong A, *et al.* Assessing record linkage between health care and vital statistics databases using deterministic methods. *BMC Health Serv Res* 2006;6:48.
 - 22 Harron K. Introduction to data linkage: administrative data research network, 2016. Available: https://adrn.ac.uk/media/174200/data_linkage_katieharron_2016.pdf
 - 23 Hagger-Johnson G, Harron K, Fleming T, *et al.* Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open* 2015;5:e008118.
 - 24 Hagger-Johnson G, Harron K, Gonzalez-Izquierdo A, *et al.* Identifying possible false matches in anonymized hospital administrative data without patient identifiers. *Health Serv Res* 2015;50:1162–78.
 - 25 Moore CL, Amin J, Gidding HF, *et al.* A new method for assessing how sensitivity and specificity of linkage studies affects estimation. *PLoS One* 2014;9:e103690.
 - 26 Flabouris A, Hart GK, George C. Outcomes of patients admitted to tertiary intensive care units after interhospital transfer: comparison with patients admitted from emergency departments. *Crit Care Resusc* 2008;10:97–105.
 - 27 Vu T, Day L, Finch CF. Linked versus unlinked hospital discharge data on hip fractures for estimating incidence and comorbidity profiles. *BMC Med Res Methodol* 2012;12:113.
 - 28 George BP, Doyle SJ, Albert GP, *et al.* Interfacility transfers for US ischemic stroke and TIA, 2006–2014. *Neurology* 2018;90:e1561–9.