Probabilistic linkage without personal information successfully linked national clinical datasets

Helen A. Blake[1,2], Linda D. Sharples[3], Katie Harron[4], Jan H. van der Meulen[1,2], Kate Walker[1,2]

1. Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, UK

2. Clinical Effectiveness Unit, Royal College of Surgeons of England, London WC2A 3PE, UK

3. Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

4. University College London (UCL) Great Ormond Street Institute of Child Health, UCL, London, WC1N 1EH, UK

Abstract

Background: Probabilistic linkage can link patients from different clinical databases without the need for personal information. If accurate linkage can be achieved, it would accelerate the use of linked datasets to address important clinical and public health questions.

Objective: We developed a step-by-step process for probabilistic linkage of national clinical and administrative datasets without personal information, and validated it against deterministic linkage using patient identifiers.

Study Design and Setting: We used electronic health records from the National Bowel Cancer Audit (NBOCA) and Hospital Episode Statistics (HES) databases for 10,566 bowel cancer patients undergoing emergency surgery in the English National Health Service.

Results: Probabilistic linkage linked 81.4% of NBOCA records to HES, versus 82.8% using deterministic linkage. No systematic differences were seen between patients that were and were not linked, and regression models for mortality and length of hospital stay according to patient and tumour characteristics were not sensitive to the linkage approach.

Conclusion: Probabilistic linkage was successful in linking national clinical and administrative datasets for patients undergoing a major surgical procedure. It allows analysts outside highly secure data environments to undertake linkage while minimising costs and delays, protecting data security, and maintaining linkage quality.

Key words:

electronic health records
national clinical datasets
patient identifiers
personal information
probabilistic linkage
record linkage

Running title: Linkage of national clinical datasets without patient identifiers using probabilistic methods

Word count: 2998

What is new?

Key findings:

Probabilistic linkage without the need for patient identifiers can be used as an alternative to deterministic linkage using patient identifiers, or as a method for enhancing deterministic linkage.

What does this add to what is known?

We developed and validated a step-by-step process for accurately linking national clinical datasets without the need for patient identifiers, providing guidance on selecting variables for linkage, estimating match weights, and choosing the probabilistic linkage threshold.

What is the implication and what should change now?

The use of probabilistic linkage without patient identifiers can increase capacity for linkage of clinical datasets and minimise costs and delays, while maintaining linkage quality and protecting data security.

1 Introduction

Linkage of electronic health records from different sources is increasingly used to address important clinical and public health questions [1, 2]. However, linking datasets requires access to unique patient identifiers and/or personal information [3]. Alternative methods that do not require such information could provide linkage whilst preserving data security [4, 5]. There is limited evidence how well these methods perform.

The two main linkage methods are deterministic, with rules to decide whether records in two datasets belong to the same individual, and probabilistic, where record pairs are given scores representing likelihoods of belonging to the same individual given strength of agreement of variables.

Deterministic linkage often uses exact agreement on unique patient identifiers, such as the NHS number in the UK's National Health Service (NHS), and personal information, such as date of birth, residential address, and postcode [6, 7]. Since different patients rarely have identical sets of these *direct* patient identifiers, such methods have high specificity. If patient identifiers are missing or misclassified, deterministic linkage has reduced sensitivity (lower probability of correctly linking records belonging to the same individual) [8].

Probabilistic linkage typically has more flexibility, enabling improved linkage when patient identifiers are missing or misclassified. It therefore tends to higher sensitivity, but lower specificity (higher probability of linking records not belonging to the same individual) than deterministic linkage [6, 9, 10, 11]. Furthermore, probabilistic linkage easily utilise a wider set of identifying variables, such as area of residence, age, treating hospital, and dates of admission/procedures/discharge [10]. These *proxy* and *indirect* identifiers can discriminate between records from different patients when used in combination, even if direct identifiers are unavailable [5]. Technically, it is possible to incorporate these variables (and similarity/distance in such variables between datasets) into a deterministic linkage strategy; however, this results in many possible deterministic rules requiring linkage decisions.

Linkage without patient identifiers has potential benefits. First, anonymised/pseudonymised datasets could be linked by a wider group of analysts, not just those working in highly secure data environments. This reduces the need to transfer patient identifiers to trusted third parties, thus minimising delays, costs, and risk of disclosure of sensitive information. Second, it would improve analysts' understanding of linkage issues [3]. Third, it allows linkage of healthcare data even if patient identifiers are unavailable [4].

We developed a step-by-step process for probabilistic linkage using indirect identifiers, clarifying methodological choices regarding selection of linkage variables, and estimating match weights. We illustrate our approach using clinical and administrative databases of bowel cancer patients undergoing emergency surgery in the NHS, validating against deterministic linkage using patient identifiers, executed by a third party.

2 Methods

2.1 Data sources and definitions

2.1.1 National Bowel Cancer Audit

National Bowel Cancer Audit (NBOCA) uses routinely collected patient characteristics, tumour pathology, processes of care and health outcomes for patients diagnosed with bowel cancer in NHS hospitals in England and Wales [12]. Our dataset includes adults (≥18 years), newly diagnosed in England with bowel cancer between 30 November 2013 and 31 March 2017. Of 75,762 identified patients, 10,566 underwent urgent or emergency surgery (major bowel resection or stoma formation) (Appendix Figure A1). All were expected to have corresponding Hospital Episode Statistics (HES) records.

2.1.2 Hospital Episode Statistics

HES records all English NHS hospital admissions for administrative/reimbursement purposes [13, 14]. Data includes dates and types of admission/procedure/discharge, patient characteristics, diagnoses, and Office of National Statistics mortality data [15].

We selected patients in HES to match NBOCA inclusion criteria as closely as possible. Surgical urgency was not recorded in HES, so all surgery patients were retained. We identified 1,434,135 records for 103,094 patients with a hospital admission for bowel cancer between 30 November 2013 and 31 March 2017, and no record of a bowel cancer diagnosis admission in the preceding five years. We defined primary procedure as the earliest major bowel resection or stoma formation within this period and not earlier than 30 days before date of diagnosis. The admission record containing the primary procedure was used to capture patient, tumour, and surgical procedure information from available records. The final cohort used for linkage consisted of 69,759 patients (Appendix Figure A2).

2.1.3 Data item definitions

The following data were obtained from both NBOCA and HES databases: age at diagnosis (in years), sex, Lower Super Output Area (LSOA; defined below), emergency admission, date of surgery, surgical procedure, responsible surgeon, hospital trust, Cancer Alliance, surgical

approach, cancer site, and distant metastases. American Society of Anesthesiologists (ASA) grade, performance status and cancer stage were obtained from NBOCA. Outcomes (90-day mortality, two-year mortality, length of stay) and number of comorbidities were obtained from HES.

LSOA represent small geographical areas in England and Wales, defined by postcode (average 672 households per LSOA) [16]. NHS hospital trusts (groups of hospitals) coordinate cancer care provision within 20 defined geographical areas called Cancer Alliances [17]. ASA grade categorises a patient's physical status from one (healthy) to five (moribund) [18]. Performance status categorises functional ability from zero (normal activity) to four (no self-care) [19].

In both datasets, diagnostic information used ICD-10 codes [20], categorised by cancer site, and surgical procedure used OPCS-4 codes [21] (see Appendix Table A1). Surgical approach (open or laparoscopic) was recorded in NBOCA and derived from OPCS-4 codes in HES (Appendix Table A1). Age at diagnosis was recorded in NBOCA and derived in HES from earliest admission with a bowel cancer diagnosis. Emergency admission was derived in each from method of admission. Cancer stage in four categories and distant metastases were derived from final pathology TNM staging in NBOCA [22]. Number of comorbidities was defined using ICD-10 codes in HES, according to the Royal College of Surgeons of England Charlson Score [23].

2.1.4 Classification of variables

Linkage variables were categorised as patient identifiers, proxy identifiers, or indirect identifiers.

Patient identifiers are variables containing information that is explicitly collected so individuals can be identified. Deterministic linkage used NHS number, date of birth, and postcode.

Proxy identifiers are variables derived from but less precise than patient identifiers. We considered age at diagnosis for date of birth and LSOA for postcode.

Indirect identifiers are variables, not derived from patient identifiers, that can discriminate between patients for the purpose of record linkage. We considered sex, date of surgery, surgical procedure, responsible surgeon, hospital trust, surgical approach, cancer site, distant metastases, and emergency admission.

2.2 Deterministic linkage using patient identifiers

NHS Digital conducted deterministic linkage using NHS number, sex, date of birth, and postcode. A sequence of eight deterministic rules were applied (Appendix Figure A3). The stage at which records are linked, *match rank*, ranges from 1 (records agree on all four patient identifiers) to 8 (records agree on NHS number only).

2.3 Probabilistic linkage methods

Mimicking the situation where investigators have no patient identifiers, probabilistic linkage used indirect and proxy identifiers, taking NBOCA as the master dataset, and linking to HES.

Probabilistic linkage uses two key quantities, *m-probability* (measure of data quality), and *u-probability* (measure of chance agreement); definitions in Appendix B. Using subscripts *1* for NBOCA and *2* for HES, m-probability is the probability that a pair of records agree for linkage variable x, given records belong to the same individual, $prob(x_1 = x_2 | I_1 = I_2)$ [24]. The u-probability is the probability that a pair of records agree for x, given records belong to different individuals, $prob(x_1 = x_2 | I_1 \neq I_2)$ [24]. The *match weight* is the ratio of these quantities and reflects how well each variable discriminates between individuals [25].

For record pairs agreeing on an identifier, the m/u ratio,

$$prob(x_1 = x_2 | I_1 = I_2) / prob(x_1 = x_2 | I_1 \neq I_2),$$

provides an *agreement contribution* to the match weight.

For record pairs disagreeing, the m/u ratio,

$$prob(x_1 \neq x_2 | I_1 = I_2) / prob(x_1 \neq x_2 | I_1 \neq I_2),$$

provides a *disagreement contribution* to the match weight.

In practice, to simplify computations, we use the log(base2)-transformation of these ratios, with the benefit that one unit increase in logged weight corresponds to doubling the ratio [25].

Most probabilistic linkage approaches assume linkage variables are independent, conditional on match status of an individual [11, 24]. Hence transformed match weights can be summed over all linkage variables to obtain *overall match weight*.

2.3.1 Calculation of overall match weights for each proxy and indirect identifier

For linkage, we selected candidate identifiers, based on ability to discriminate between matches and non-matches, and completeness of records. First, for each identifier, x, we

selected record pairs that agreed on all other proxy/indirect identifiers and estimated overall m-probability as the proportion of these record pairs that agreed exactly on x. For overall u-probability, if x had $K \leq 10$ possible categories, we calculated proportions in each category in NBOCA and HES and summed the products of these proportions across all possible values:

$$\sum_{k=1}^{K} prob(x_1 = k) \times prob(x_2 = k).$$

For identifiers with $K > 10$ categories, we estimated overall u-probability using the reciprocal of number of distinct values, $1/K$.

Overall agreement/disagreement contributions were estimated using these overall m/u-probabilities. Selection of which proxy/indirect identifiers to use for linkage was based on the difference between overall agreement and disagreement contributions, amongst variables that were missing for <5% individuals.

2.3.2 Linking individuals using match weights

The match weights in the linkage algorithm were calculated as above, except for age, date of surgery, sex, and surgical procedure. For age and date of surgery, we estimated m/u-probabilities for exact agreement and three levels of disagreement, to allow for varying differences in dates (Appendix C). For sex and surgical procedure, as individuals were not distributed evenly across categories, we estimated m/u-probabilities for each category (Appendix C). Match weight contribution was set to zero for missing values.

In practice, comparing each NBOCA-HES record pair involved estimating over 737 million (10,566×69,759) potential links. We used three sequential blocking steps to reduce computational burden. First, records with exact agreement on Cancer Alliance underwent linkage, then unlinked NBOCA records with age difference ≤10 years, and finally, remaining unlinked NBOCA records with date of surgery within 180 days.

In each blocking step, match weights were calculated for all possible record pairs and summarised by histogram, resulting in two overlapping distributions, one for true matches and one for non-matches. A threshold was chosen as the point where the distributions intersected.

2.4 Validation

Taking deterministic linkage as gold-standard, we calculated sensitivity and specificity of probabilistic linkage based on proxy/indirect identifiers and plotted a Receiver Operating Characteristic (ROC) curve across alternative thresholds of match weights. Where records

were linked by only one method, agreement on individual linkage variables were examined to assess likelihood of *false links* (non-matches that are linked) and *missed links* (true matches that are not linked).

To explore potential sources of bias arising from linkage error, we compared characteristics for deterministically linked record pairs according to whether they were probabilistically linked.

To assess sensitivity of statistical analyses to linkage method we fitted regression models to: 90-day mortality (logistic), survival from surgery up to two years (Cox proportional hazards), and length of stay (LOS; linear). Independent variables included in all models were: age at diagnosis, comorbidities, ASA grade, cancer site, emergency admission, cancer stage, and surgical approach.

3 Results

3.1 Probabilistic linkage

Table 1 shows overall m/u-probabilities for candidate linkage variables. Patient and administrative variables (LSOA, dates, treatment providers, age, and sex), tended to discriminate better than clinical variables (cancer site/stage). Figure 1 plots overall m-probability against one minus overall u-probability for linkage variables considered. For linkage, we included variables with difference in log-transformed overall contributions for agreement/disagreement of around 5 or more: LSOA, hospital trust, date of surgery, responsible surgeon, age, sex, and surgical procedure. Calculated m-probabilities, u-probabilities, and match weight contributions by linkage variable are in Appendix C.

After reviewing a histogram of match weights (Figure 2), a threshold of 25 was chosen, resulting in 81.4% (8,603/10,566) of eligible NBOCA records linked to HES using probabilistic linkage (Figure 3). This compares to 82.8% (8,748/10,566) of eligible NBOCA records linking deterministically, with >99% of these agreeing on NHS number, sex, and date of birth (Appendix Table D1).

Table 1: Measures of discrimination and proportion of missing values for candidate linkage variables

Candidate linkage variable	Number of distinct values	% missing data in NBOCA (N=10,566)	% missing data in HES (N=69,759)	Overall m-probability (data quality)	Overall u-probability (chance	Log-transformed overall contribution for agreement	Log-transformed overall contribution for disagreement	Difference in log-transformed overall

				agreement)				contributions	
NHS number[1]	2	N/A	N/A	0.992	<0.001	15.91	-6.99	22.90	Higher discrimination
LSOA	27,582	0.13%	0.69%	0.953	<0.001	14.68	-4.40	19.08	
Hospital trust	145	0%	0%	0.999	0.007	7.18	-10.57	17.75	
Date of surgery	1,503	0%	0%	0.858	0.001	10.33	-2.81	13.14	
Responsible surgeon	3,149	0.38%	0.32%	0.672	<0.001	11.05	-1.61	12.65	
Age	88	0%	0.32%	0.973	0.011	6.42	-5.17	11.59	
Sex	2	0.04%	<0.01%	0.997	0.502	0.99	-7.58	8.57	
Surgical procedure	10	0%	0%	0.893	0.225	1.99	-2.86	4.84	
Cancer site	3	0%	0%	0.939	0.607	0.63	-2.69	3.32	
Surgical approach	2	1.28%	0%	0.858	0.501	0.78	-1.82	2.59	Lower discrimination
Emergency admission	2	1.15%	0.13%	0.741	0.491	0.59	-0.97	1.56	
Distant metastases	2	25.3%	0%	0.628	0.479	0.39	-0.48	0.87	

[1] Whether or not NHS number matched, according to the match rank variable supplied by NHS Digital in HES (NHS number did not match if match rank was 6, 7, or missing).

Figure 1: Comparing data quality and chance agreement using overall m-probabilities and u-probabilities of candidate linkage variables

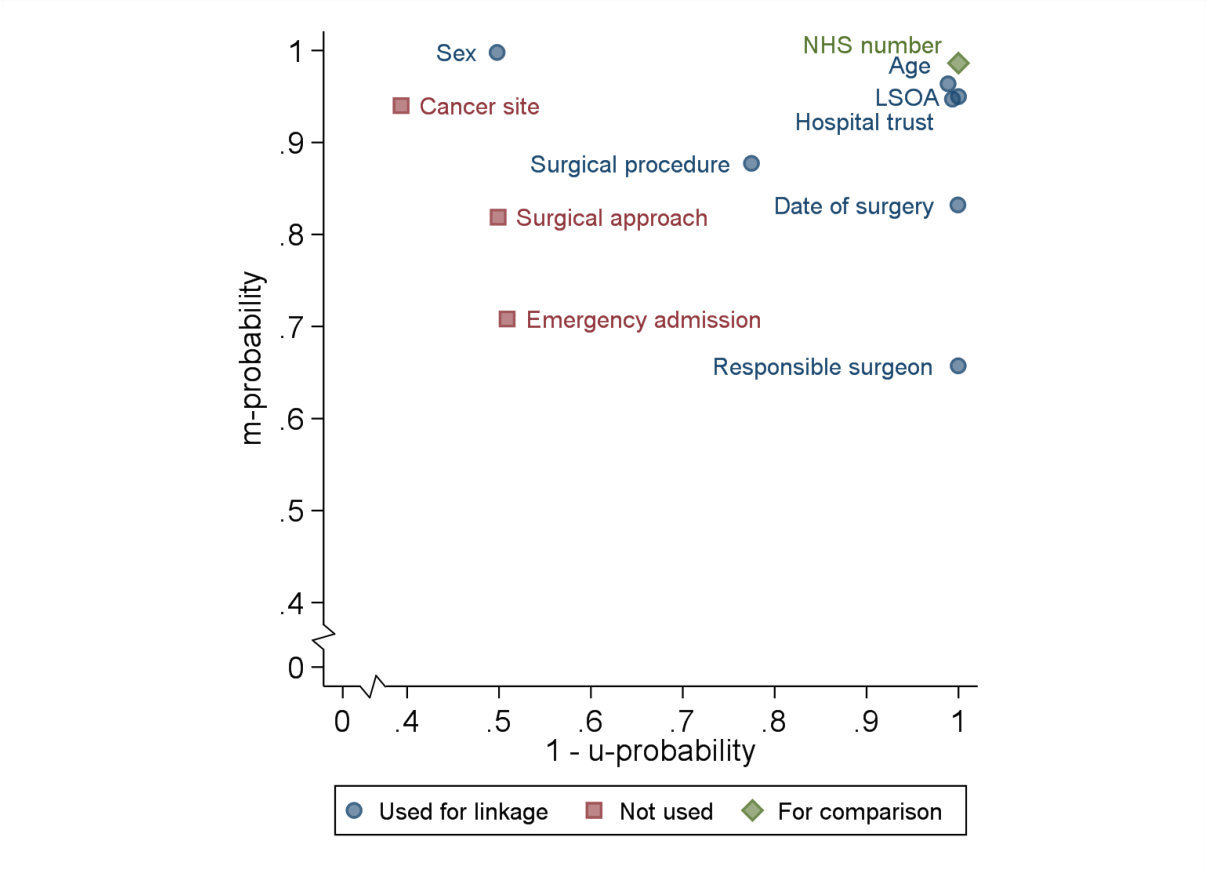


Figure 2: Distribution of match weights computed in blocking step 1 after deterministically linking on Cancer Alliance (threshold for linkage, T, of 25 chosen at the point where two distributions intersect).

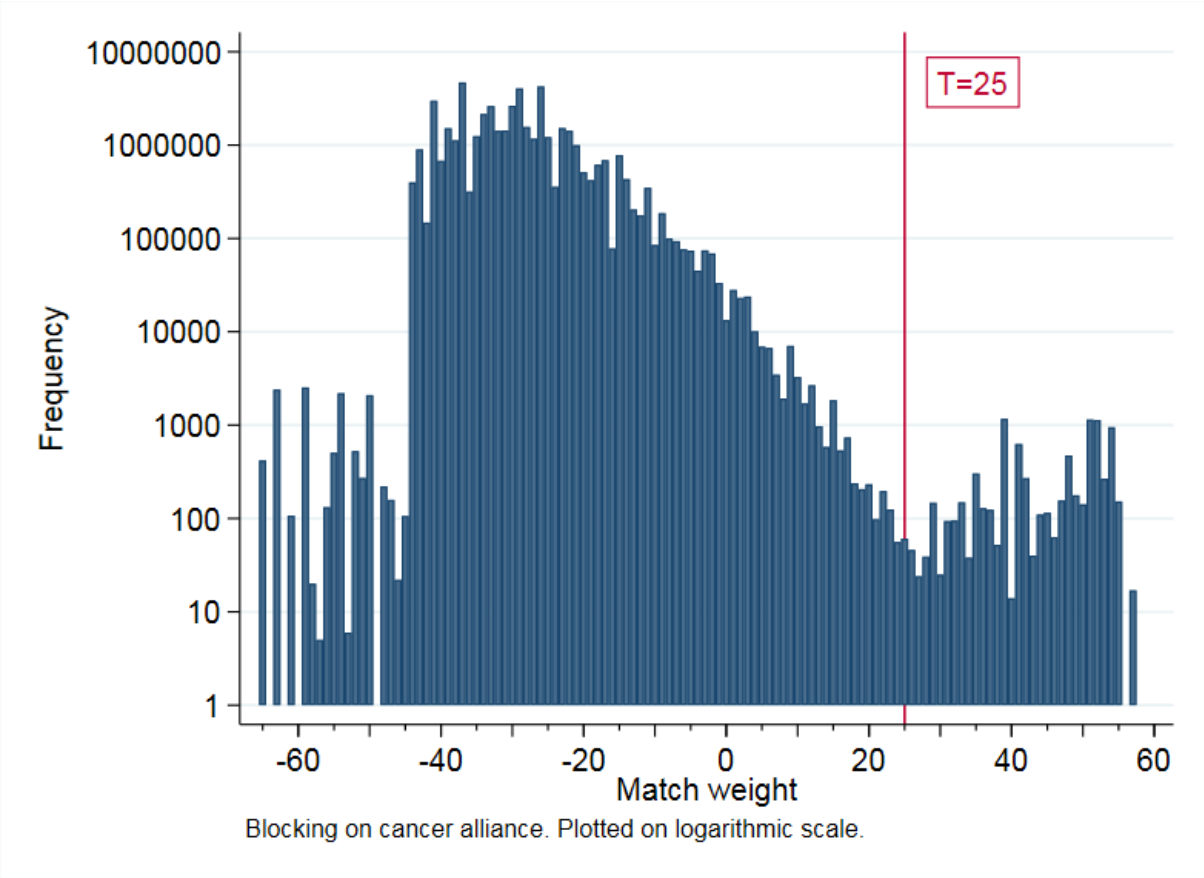
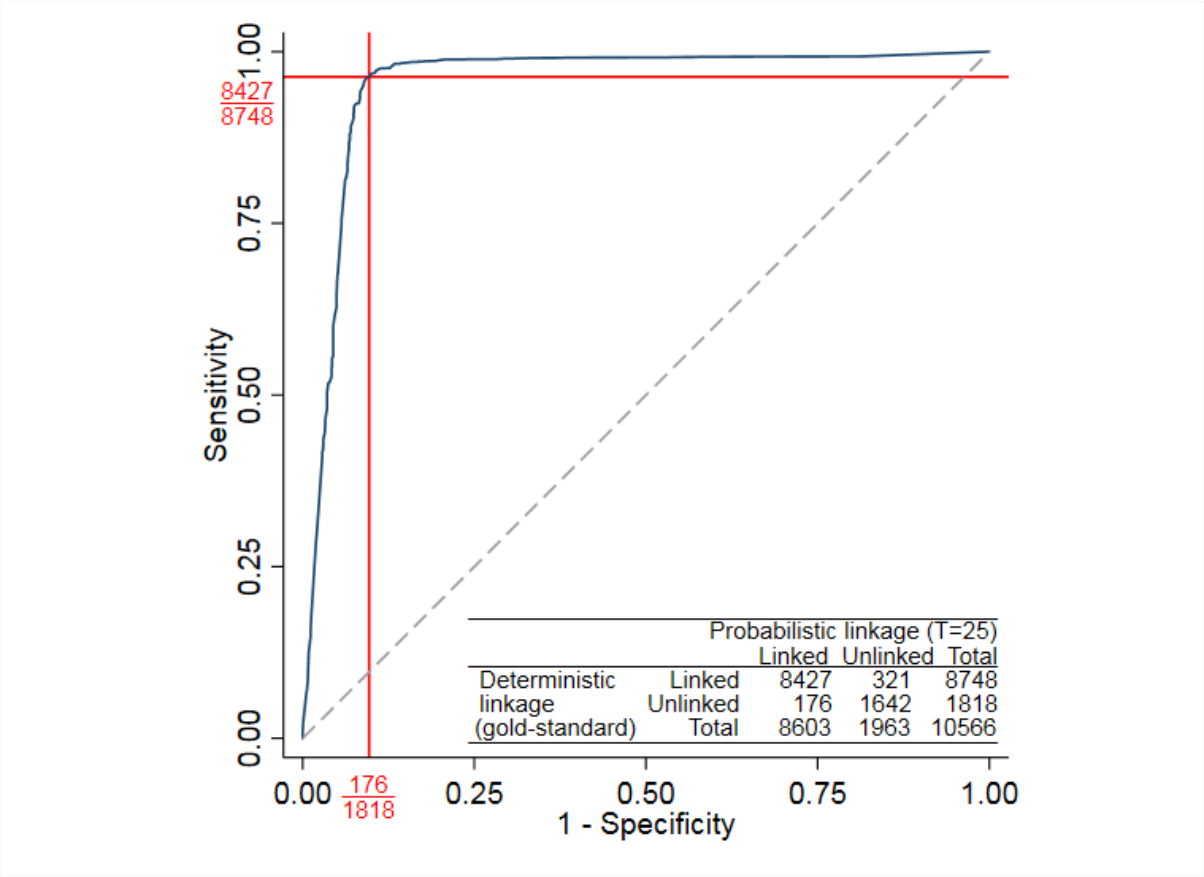


Figure 3: Receiver Operating Characteristic curve evaluating the sensitivity and specificity of probabilistic linkage match weights compared to assumed gold-standard of deterministic linkage (with probabilistic linkage threshold T=25 marked).



3.2 Validation results

3.2.1 Evaluating sensitivity and specificity of probabilistic linkage

Most NBOCA records were linked to HES using both methods (8,427/10,566) (Figure 3). Sensitivity and specificity of probabilistic linkage, with deterministic linkage as gold-standard, were 96.3% (8427/8748) and 90.3% (1642/1818) respectively. Figure 3 shows that reducing the threshold would yield small gains in sensitivity for substantial reductions in specificity. Conversely, increasing the threshold would improve specificity but substantially reduce sensitivity.

Table 2 shows agreement patterns for 176 records that linked probabilistically but not deterministically. 143 (81%) agreed on LSOA and ≥4 other identifiers, suggesting most are true links and specificity of the probabilistic linkage is therefore likely underestimated (as these links were missed in deterministic linkage).

Table 2: Agreement patterns for record pairs linked by probabilistic linkage but not deterministic linkage. • indicates exact agreement.

Agreement patterns											
LSOA	Date of surgery	Responsible surgeon	Hospital trust	Age	Sex	Surgical procedure	N. agree	Frequency (%)		Match weight Mean	Match weight Range
•	•	•	•	•	•	•	7	56	(31.82)	52.12	40.38-55.08
•	•		•	•	•	•	6	36	(20.45)	38.14	27.82-43.70
•	•	•	•	•	•		6	13	(7.39)	47.62	38.38-49.40
•	•	•	•		•	•	6	12	(6.82)	43.21	36.13-48.61
•		•	•	•	•	•	6	8	(4.55)	45.73	34.95-50.46
•			•	•	•	•	5	5	(2.84)	33.20	25.08-38.68
•	•		•	•	•		5	5	(2.84)	35.07	33.23-37.47
•		•	•	•	•		5	3	(1.70)	35.87	31.29-44.78
•	•		•		•	•	5	2	(1.14)	34.59	33.41-35.76
•	•	•	•			•	5	1	(0.57)	34.99	34.99-34.99
•		•	•	•		•	5	1	(0.57)	26.39	26.39-26.39
•		•	•		•	•	5	1	(0.57)	28.00	28.00-28.00
•			•	•	•		4	3	(1.70)	29.19	26.03-32.28
•		•	•		•		4	3	(1.70)	25.60	25.11-26.00
	•	•	•	•	•	•	6	4	(2.27)	33.62	33.15-35.05
	•	•	•		•	•	5	11	(6.25)	30.67	26.97-32.97
		•	•	•	•	•	5	4	(2.27)	28.13	25.30-29.19
	•	•	•	•	•		5	3	(1.70)	29.21	26.80-30.77
	•		•	•	•	•	5	1	(0.57)	26.53	26.53-26.53
	•	•	•		•		4	2	(1.14)	26.21	25.08-27.34
		•	•	•	•		4	1	(0.57)	26.15	26.15-26.15
		•	•		•	•	4	1	(0.57)	29.38	29.38-29.38

All 321 records that linked deterministically but not probabilistically matched on at least NHS number, date of birth, and sex (Appendix Table D1), suggesting they are also true links (missed links in the probabilistic linkage). 96% of these disagreed on at least one of LSOA, hospital trust, and date of surgery.

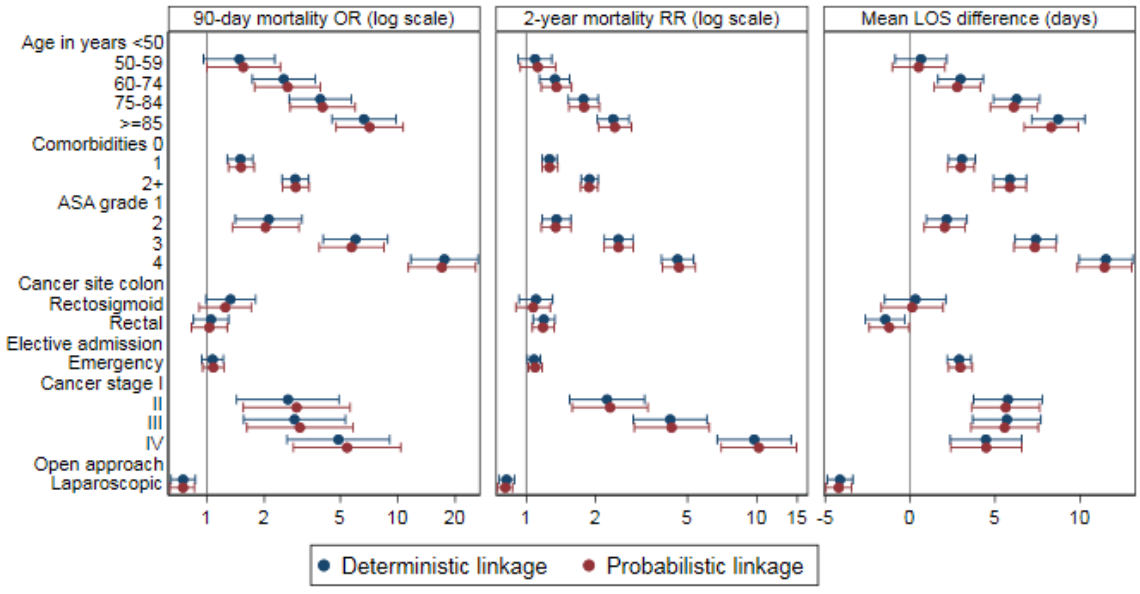
3.2.2 Comparing characteristics of probabilistically and deterministically linked records

Patients that linked deterministically, but not probabilistically, were younger and more likely to have emergency admission (Appendix Table D2). Otherwise, patients had similar characteristics.

3.2.3 Comparing regression coefficients from analyses using deterministically and probabilistically linked datasets

Figure 4 summarises results of regression models for factors affecting 90-day mortality, survival up to two years and LOS for patients linked deterministically and probabilistically (Appendix Tables D3-D5). Overall, 90-day mortality was 12.41% (95% CI: 11.72-13.14) for probabilistically linked patients, compared to 12.36% (95% CI: 11.67-31.06) for deterministically linked patients. Two-year mortality was 39.57% (95% CI: 38.53-40.63) versus 39.55% (95% CI: 38.52-40.58) respectively, and mean LOS was 15.4 days (95% CI: 15.1-15.8) versus 15.5 days (95% CI: 15.2-15.9). There were no substantive differences in results between sets of models, with all variables having similar point estimates and overlapping confidence intervals. There were small differences in effects of age, ASA grade, and cancer stage on 90-day mortality but little impact for other outcomes.

Figure 4: Crude estimates and 95% confidence intervals (CI) for 90-day mortality odds ratios (OR), 2-year mortality hazard ratios (HR), and crude mean difference in length of stay (LOS) using patients linked deterministically or patients linked probabilistically. Mortality outcomes plotted using logarithmic scale.



Discussion

4.1 Summary

Illustrated using bowel cancer patients undergoing emergency major surgery, we provided a step-by-step process for linking clinical datasets without personal information, with guidance on selecting variables for linkage and calculating match weights. The approach had over 96% sensitivity and 90% specificity compared to deterministic linkage using patient identifiers. There were no systematic differences between linked and unlinked patients. Regression analyses for mortality and LOS were not sensitive to the linkage approach.

There is currently limited evaluation of linkage without patient identifiers, or guidance on implementation. Previous studies have linked data without a unique patient identification number, however many used other patient identifiers, such as patient name [7] and date of birth [9, 26]. Other studies have linked clinical and administrative datasets with only indirect identifiers but without a gold-standard for validation [4, 5].

4.2 Limitations

Our example comprised patients undergoing major surgery in an acute setting with diagnosis and treatment in the same admission. Linkage variables were related to patient and procedure, rather than diagnosis and earlier investigations, and included a mixture of identifiers generalisable across all settings (e.g. geographical area, age, sex, date of event/ diagnosis/procedure) and application-specific identifiers (e.g. surgical procedure, responsible surgeon). Even for common events such as childbirth, there will rarely be more than one person in the same LSOA of the same age having an event/procedure on the same day (approximately 1750 births per day in England and Wales across 34,753 LSOAs)[4, 27]. When multiple people have the same combination of these generalisable identifiers, additional application-specific identifiers can be used to differentiate between likely and unlikely links. We therefore expect our approach to apply for all major events or treatments requiring admission to hospital. For elective procedures, healthcare may spread over multiple admissions and further work should explore generalisability to these scenarios.

Patient and administrative variables were used for probabilistic linkage rather than clinical variables, as they contributed more to linkage and were less subjective. Hence linkage without patient identifiers works better when more administrative information is available across both datasets.

If available, free-text information such as patient name, address, and clinician notes would allow a more accurate gold-standard to be generated through manual review. However,

deterministic linkage using NHS number and other patient identifiers was available and it is rare that different patients would have identical NHS numbers. Therefore, this approach likely to has 100% specificity, but less than 100% sensitivity. Our analysis suggested that both linkage methods missed some links and deterministic linkage was not 100% sensitive. Thus, the specificity of the probabilistic linkage will have been underestimated. Reducing the threshold reduces missed links from probabilistic linkage, but at the cost of reduced specificity.

The methods used to estimate match weights assumed conditional independence between linkage variables and correct specification of m/u-probabilities, although linkage results tend not to be sensitive to mis-specification [28]. However, if many linkage variables are used, issues with missing data, mismeasurement, or dependence may be amplified [29, 30]. We note that all our linkage variables had low frequencies of missing data.

For probabilistic linkage, we selected the record with the highest match weight. If patients in one dataset had multiple potential links to the other dataset, some information may have been discarded. Alternative approaches include prior-informed imputation, which may perform better in some cases [31, 32], and multiple imputation methods deal with uncertainty in linkage [33].

To reduce computational burden, we linked records using blocking, as previously recommended [3, 5]. This strategy has been criticised for increasing missed links [2], but sequential blocking steps should minimise this. Parallelisation is a potential alternative solution.

4.3 Implications

With increasing availability of large clinical datasets, there is potential to build more complete pathways of patient care through data linkage. Our results demonstrate that probabilistic linkage of anonymised/pseudonymised datasets using indirect and proxy identifiers has the potential to increase capacity for data linkage and minimise costs and delays, while preserving data security and maintaining linkage quality.

Our findings also demonstrate that probabilistic linkage using indirect and proxy identifiers can recover links missed deterministically, due to missing or misclassified patient identifiers. This suggests probabilistic linkage using indirect and proxy identifiers can enhance deterministic linkage methods.

Published guidance recommends providing transparency throughout linkage [2], as we have illustrated. For example, we demonstrated that the difference in the log-transformed overall agreement/disagreement contributions of each linkage variable can identify variables that would contribute the strongest to linkage. Furthermore, we followed recent guidance which proposes evaluating linkage quality by comparing the characteristics of individuals linked using different linkage approaches and assessing sensitivity of analyses to the linkage approach [34].

4.4 Conclusion

Probabilistic linkage without patient identifiers was successful in linking national clinical datasets for patients undergoing a major surgical procedure. It has important implications as it allows analysts outside highly secure data environments to carry out linkage while protecting data security and maintaining – and potentially improving – linkage quality.

Acknowledgements

This study/project is funded by the National Institute for Health Research (NIHR) HS&DR Project:17/05/45. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

The National Bowel Cancer Audit is commissioned by the Healthcare Quality Improvement Partnership (HQIP) as part of the National Clinical Audit and Patient Outcomes Programme, and funded by NHS England and the Welsh Government (www.hqip.org.uk/national-programmes). Neither HQIP nor the funders had any involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

KH is funded by the Wellcome Trust (Grant 212953/Z/18/Z). This research was supported in part by the NIHR Great Ormond Street Hospital Biomedical Research Centre and the Health Data Research UK (grant No. LOND1), which is funded by the UK Medical Research Council and eight other funders.

This work uses data provided by patients and collected by the NHS as part of their care and support.

As the National Bowel Cancer Audit involves analysis of data for service evaluation, it is exempt from UK National Research Ethics Committee approval. Section 251 approval was obtained from the Ethics and Confidentiality Committee for the collection of personal health data without the consent of patients. The study was performed in accordance with the Declaration of Helsinki.

Declarations of interest:

None.

Author contributions:

HAB: data curation, formal analysis, methodology, writing - original draft, review and editing

LS: funding acquisition, methodology, writing - original draft, review and editing

KH: funding acquisition, methodology, writing - original draft, review and editing

JvdM: conceptualisation, methodology, funding acquisition, writing - original draft, review and editing

KW: conceptualisation, methodology, funding acquisition, writing - original draft, review and editing

Availability of data statement:

The data used in this study are available from NHS Digital and Public Health England's Office for Data Release but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. We do not have permission to share the patient-level records used in our analysis.

References

[1] M. A. Bohensky, D. Jolley, V. Sundararajan, S. Evans, D. V. Pilcher, I. Scott, and C. A. Brand, "Data linkage: A powerful research tool with potential problems," *BMC Health Services Research*, vol. 10, no. 1, p. 346, 2010.

[2] R. Gilbert, R. Lafferty, G. Hagger-Johnson, K. Harron, L.-C. Zhang, P. Smith, C. Dibben, and H. Goldstein, "GUILD: GUidance for Information about Linking Data sets," *Journal of Public Health*, vol. 40, pp. 191–198, 03 2017.

[3] K. Harron, C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M. L. Barreto, and H. Goldstein, "Challenges in administrative data linkage for research," *Big data & society*, vol. 4, no. 2, p. 2053951717745678, 2017.

[4] K. Harron, R. Gilbert, D. Cromwell, and J. van der Meulen, "Linking data for mothers and babies in de-identified electronic health data," *PLOS ONE*, vol. 11, no. 10, p. e0164667, 2016.

[5] E. H. Lawson, C. Y. Ko, R. Louie, L. Han, M. Rapp, and D. S. Zingmond, "Linkage of a clinical surgical registry with Medicare inpatient claims data using indirect identifiers," *Surgery*, vol. 153, no. 3, pp. 423–430, 2013.

[6] K. Harron, E. Mackay, and M. Elliot, "An introduction to data linkage," 2016.

[7] E. S. Paixão, K. Harron, K. Andrade, M. G. Teixeira, R. L. Fiaccone, M. d. C. a. N. Costa, and L. C. Rodrigues, "Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in brazil," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, p. 108, 2017.

[8] Y. Zhu, Y. Matsuyama, Y. Ohashi, and S. Setoguchi, "When to conduct probabilistic linkage vs. deterministic linkage? a simulation study," *Journal of Biomedical Informatics*, vol. 56, pp. 80 – 86, 2015.

[9] N. Méray, J. B. Reitsma, A. C. J. Ravelli, and G. J. Bonsel, "Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number," *Journal of Clinical Epidemiology*, vol. 60, no. 9, pp. 883–891, 2007.

[10] J. C. Doidge and K. Harron, "Demystifying probabilistic linkage: Common myths and misconceptions," *International journal of population data science*, vol. 3, no. 1, pp. 410–410, 2018. 30533534[pmid] PMC6281162[pmcid].

[11] M. Tromp, A. C. Ravelli, G. J. Bonsel, A. Hasman, and J. B. Reitsma, “Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage,” *Journal of Clinical Epidemiology*, vol. 64, no. 5, pp. 565–572, 2011.

[12] National Bowel Cancer Audit, “Annual Report 2019.” Accessed 31st March 2020. Available from: www.nboca.org.uk/reports/annual-report-2019/.

[13] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, “Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC),” *International Journal of Epidemiology*, vol. 46, no. 4, pp. 1093–1093i, 2017.

[14] NHS Digital, “Hospital Episode Statistics (HES).” Accessed 25th May 2020. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.

[15] NHS Digital, “Linked HES-ONS mortality data.” Accessed 1st December 2020. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data.

[16] Office for National Statistics, “2011 Census: Population and household estimates for small areas in England and Wales, March 2011.” Accessed 31st March 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuspopulationandhouseholdestimatesforsmallareasinenglandandwales/2012-11-23#output-areas-super-output-areas-and-electoral-wardsdivisions.

[17] NHS England, “Cancer Alliances - improving care locally.” Accessed 7th May 2020. Available from: https://www.england.nhs.uk/cancer/cancer-alliances-improving-care-locally/.

[18] M. Daabiss, “American Society of Anaesthesiologists physical status classification,” *Indian Journal of Anaesthesia*, vol. 55, no. 2, pp. 111–115, 2011.

[19] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, and P. P. Carbone, “Toxicity and response criteria of the eastern cooperative oncology group,” *American journal of clinical oncology*, vol. 5, no. 6, pp. 649–656, 1982.

[20] NHS Digital, “International statistical classification of diseases and health related problems (ICD-10) 5th Edition.” Accessed 24th September 2019. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-

collections/scci0021-international-statistical-classification-of-diseases-and-health-related-problems-icd-10-5th-edition.

[21] NHS Digital, “Hospital Episode Statistics (HES).” Accessed 24th September 2019. Available from: https://datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions_and_procedures.html.

[22] “Colorectal cancer staging,” *CA: A Cancer Journal for Clinicians*, vol. 54, no. 6, pp. 362–365, 2004.

[23] J. N. Armitage and J. H. van der Meulen, “Identifying co-morbidity in surgical patients using administrative data with the royal college of surgeons charlson score,” *BJS*, vol. 97, no. 5, pp. 772–781, 2010.

[24] I. P. Fellegi and A. B. Sunter, “A theory for record linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[25] A. Sayers, Y. Ben-Shlomo, A. W. Blom, and F. Steele, “Probabilistic record linkage,” *International Journal of Epidemiology*, vol. 45, no. 3, pp. 954–964, 2015.

[26] B. G. Hammill, A. F. Hernandez, E. D. Peterson, G. C. Fonarow, K. A. Schulman, and L. H. Curtis, “Linking inpatient clinical registry data to medicare claims data using indirect identifiers,” *American Heart Journal*, vol. 157, no. 6, pp. 995–1000, 2009.

[27] Office for National Statistics, “Births in England and Wales: 2019.” Accessed 19th April 2021. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2019.

[28] H. Xu, X. Li, C. Shen, S. L. Hui, S. Grannis, *et al.*, “Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter?,” *The Annals of Applied Statistics*, vol. 13, no. 3, pp. 1753–1790, 2019.

[29] W. E. Winkler, “Overview of record linkage and current research directions,” in *Bureau of the Census*, Citeseer.

[30] W. Sauerbrei, A. Buchholz, A.-L. Boulesteix, and H. Binder, “On stability issues in deriving multivariable regression models,” *Biometrical Journal*, vol. 57, no. 4, pp. 531–555, 2015.

[31] K. Harron, H. Goldstein, A. Wade, B. Muller-Pebody, R. Parslow, and R. Gilbert, “Linkage, evaluation and analysis of national electronic healthcare data: Application to

providing enhanced blood-stream infection surveillance in paediatric intensive care," *PLOS ONE*, vol. 8, no. 12, p. e85278, 2013.

[32] K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein, "Evaluating bias due to data linkage error in electronic healthcare records," *BMC Medical Research Methodology*, vol. 14, no. 1, p. 36, 2014.

[33] H. Goldstein, K. Harron, and A. Wade, "The analysis of record-linked data using multiple imputation with data value priors," *Statistics in Medicine*, vol. 31, no. 28, pp. 3481–3493, 2012.

[34] K. L. Harron, J. C. Doidge, H. E. Knight, R. E. Gilbert, H. Goldstein, D. A. Cromwell, and J. H. van der Meulen, "A guide to evaluating linkage quality for the analysis of linked data," *International Journal of Epidemiology*, vol. 46, no. 5, pp. 1699–1710, 2017.