

Comparison of identifiable and non-identifiable data linkage: health technology assessment of MitraClip using registry, administrative and mortality datasets

Kim Keltie ^{1,2}, Paola Cognigni ¹, Sam Gross,³ Samuel Urwin ², Julie Burn ¹, Helen Cole ⁴, Lee Berry,⁵ Hannah Patrick,⁵ Andrew Sims ^{1,2}

To cite: Keltie K, Cognigni P, Gross S, *et al*. Comparison of identifiable and non-identifiable data linkage: health technology assessment of MitraClip using registry, administrative and mortality datasets. *BMJ Health Care Inform* 2021;**28**:e100223. doi:10.1136/bmjhci-2020-100223

Received 11 August 2020
Revised 21 January 2021
Accepted 08 March 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Northern Medical Physics and Clinical Engineering, Newcastle Upon Tyne Hospitals NHS Trust, Newcastle Upon Tyne, UK

²Translational and Clinical Research Institute, University of Newcastle upon Tyne, Newcastle upon Tyne, UK

³Data Management Services, NHS Digital, Leeds, UK

⁴The Northern Health Science Alliance, Manchester, UK

⁵Observational Data Unit, National Institute for Health and Care Excellence, London, UK

Correspondence to

Dr Andrew Sims;
andrew.sims5@nhs.net

ABSTRACT

Objectives The UK MitraClip registry was commissioned by National Health Service (NHS) England to assess real-world outcomes from percutaneous mitral valve repair for mitral regurgitation using a new technology, MitraClip. This study aimed to determine longitudinal patient outcomes by linking to routine datasets: Hospital Episode Statistics (HES) Admitted Patient Care (APC) and Office of National Statistics.

Methods Two methods of linkage were compared, using identifiable (NHS number, date of birth, postcode, gender) and non-identifiable data (hospital trust, age in years, admission, discharge and operation dates, operation and diagnosis codes). Outcome measures included: matching success, patient demographics, all-cause mortality and subsequent cardiac intervention.

Results A total of 197 registry patients were eligible for matching with routine administrative data. Using identifiable linkage, a total of 187 patients (94.9%) were matched with the HES APC dataset. However, 21 matched individuals (11.2%) had inconsistencies across the datasets (eg, different gender) and were subsequently removed, leaving 166 (84.3%) for analysis. Using non-identifiable data linkage, a total of 170 patients (86.3%) were uniquely matched with the HES APC dataset. Baseline patient characteristics were not significantly different between the two methods of data linkage. The total number of deaths (all causes) identified from identifiable and non-identifiable linkage methods was 37 and 40, respectively, and the difference in subsequent cardiac interventions identified between the two methods was negligible.

Conclusions Patients from a bespoke clinical procedural registry were matched to routine administrative data using identifiable and non-identifiable methods with equivalent matching success rates, similar baseline characteristics and similar 2-year outcomes.

INTRODUCTION

Evidence-based clinical guidelines, such as those published by the UK National Institute for Health and Care Excellence (NICE),

Summary

What is already known?

- ▶ Few, if any, published studies report on the strengths and limitations of different methods of linking registries to routine administration datasets, to inform health technology assessments.
- ▶ Both identifiable and non-identifiable linkage of registries with routine data sets are possible.

What does this paper add?

- ▶ Two independent methods of linking a clinical registry with Hospital Episode Statistics data enabled reliable longer term outcomes to be obtained.
- ▶ Identifiable and non-identifiable linkage are complementary.
- ▶ Transparent description of the two methods facilitates use of the techniques by other researchers.

prioritise randomised controlled trials (RCTs) over other study types. However, because RCTs are costly, lengthy and may fail to identify rare adverse events, alternative study designs using real-world data have a role to play in technology assessment.^{1 2}

With its Commissioning through Evaluation (CtE) scheme, National Health Service (NHS) England enabled a limited number of patients to access promising interventions that were not currently funded by the NHS, while collecting clinical and patient experience data within a formal evaluation programme.³ Analysis of this real-world evidence, combined with evidence from clinical trials, informed NHS England's commissioning decision. Included in the CtE programme was percutaneous mitral valve repair for mitral regurgitation using MitraClip (Abbott, Illinois, USA). Data collection was mandated in a bespoke procedural registry (UK MitraClip registry), from which in-hospital and short-term outcomes

for the procedure were reported. To obtain longer term outcome data (readmission rates, subsequent interventions, adverse events and mortality outcomes), linkage of identifiable registry data to routinely collected Hospital Episode Statistics (HES) and Office of National Statistics (ONS) mortality data was also conducted by NHS Digital.⁴

However, transcription errors in patient identifiers and missing data in any data source are known to influence data linkage and may bias results.⁵ Therefore, in order to assess the quality of the data linkage, and potential impact of any linkage errors, this study aimed to develop an alternative method, linking non-identifiable data from the MitraClip registry to pseudonymised extracts from HES Admitted Patient Care (APC) and ONS mortality datasets. This allowed a comparison of matching success rates between the two linkage methods (using identifiable and non-identifiable data) and additionally, a comparison of longitudinal outcomes such as subsequent cardiac intervention and all-cause mortality.

METHODS

Registry data collection

The MitraClip registry was commissioned by NHS England and opened on 1 October 2014. Patients were eligible for the MitraClip CtE scheme if they had stage

2, 3 or 4 mitral regurgitation of functional/ischaemic or degenerative aetiology (ie, excluding rheumatic heart disease), and were deemed high risk or were turned down for conventional mitral valve surgery. MitraClip implantation was either standalone or alongside percutaneous coronary intervention (staged or concurrent).

Identifiable data were collected in the MitraClip registry without explicit patient consent, via section 251 of the National Health Service Act 2006 (17/CAG/0153 (Previously CAG 10-07(b)/2014)).

Linkage using identifiable data

Patient identifiers were extracted from the MitraClip registry by the data controller (the National Institute for Cardiovascular Outcomes Research, NICOR) on 5 April 2018 and sent to NHS Digital (data supplied under DARS-NIC-151212-B5Z3R agreement). Records were linked by NHS Digital to the HES APC dataset and the ONS mortality dataset, using an eight-step deterministic matching proprietary algorithm based on NHS number, date of birth, gender and postcode (figure 1A).⁶ Data from HES included all admissions from matched patients with hospital discharge dates between 1 April 2008 and 1 March 2018. Data from ONS included all reported deaths from matched patients until 4 April 2018. Records from patients having registered type 2

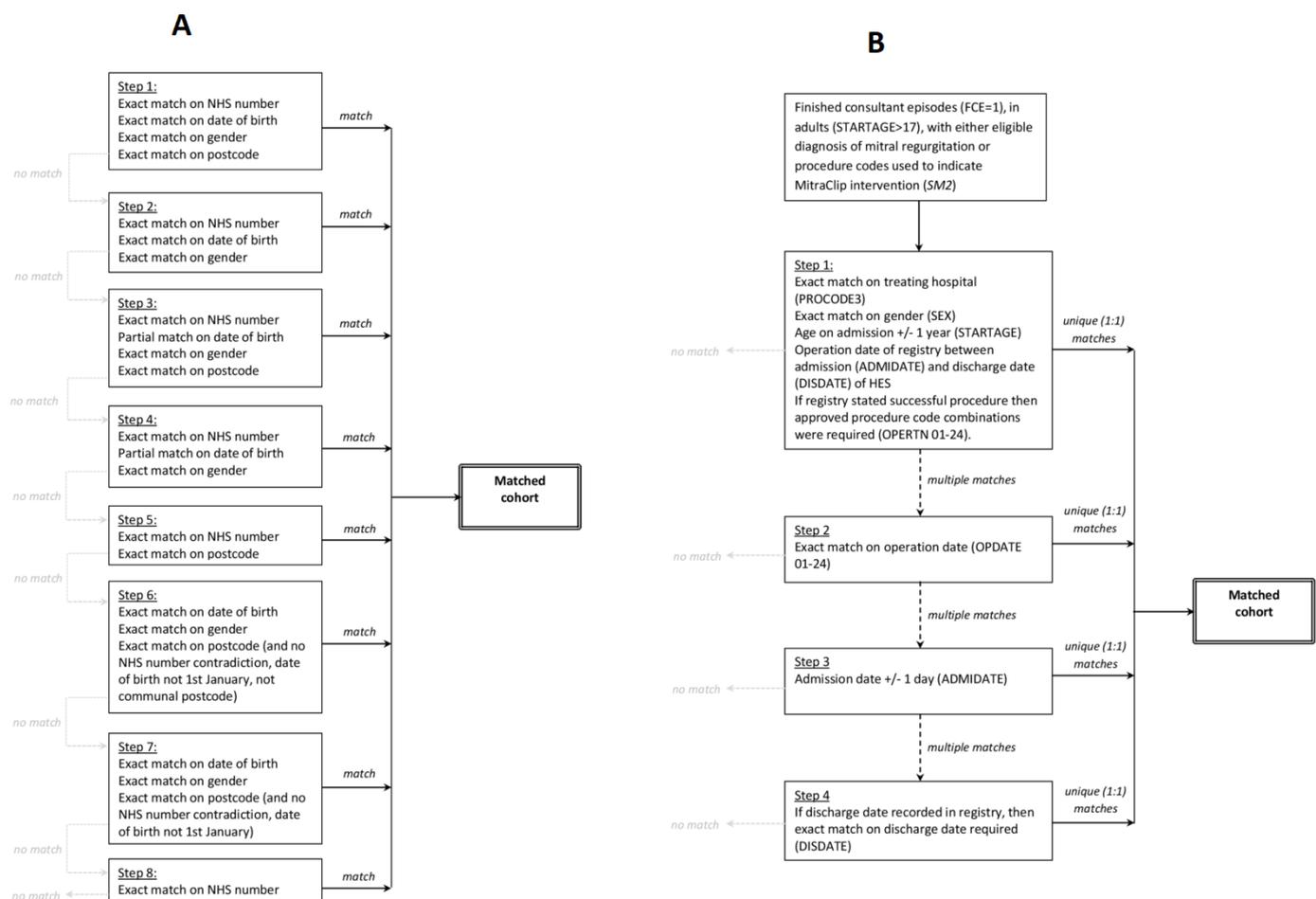


Figure 1 Matching steps used during (A) identifiable and (B) non-identifiable linkage. NHS: National Health Service.

opt-outs (ie, those not wishing their confidential patient information to be used for purposes other than their individual care) were removed from both extracts by NHS Digital before releasing the linked data back to the data controller. Patient identifiers (in registry, HES and ONS data) were then replaced with a 'Study ID' by NICOR (as data controller) and submitted to the study team (as the named co-data processor), [figure 1A](#).

Matched records from the registry, HES and ONS were reviewed by the study team, and those with conflicting demographic and administrative details were flagged to indicate potential errors in matching (ie, matching to an incorrect patient). The demographics of those with and without inconsistencies were then compared to confirm that exclusion of potentially mismatched records would not introduce bias into the linked data results. All patient-level information provided by NHS Digital resulting from the identifiable linkage was deleted by the study team on expiry of the data sharing agreement prior to the commencement of linkage using non-identifiable data, enabling an independent assessment of methods.

Linkage using non-identifiable data

Separately, the study team had access to pseudonymised HES and ONS mortality datasets for all patients admitted to hospital in England, under Data Access Request Service (DARS) agreement DAR-NIC-170211-Z1B4J. These data were supplied via NHS Digital's managed extract service and were saved on a secure SQL server within Newcastle upon Tyne Hospitals.

An anonymised data extract was taken by the data controller from the UK MitraClip registry on 5 April 2018 and sent to the study team. For patients with more than one MitraClip procedure entered into the registry, the most recent MitraClip procedure was used for matching.

From the HES dataset, individual episodes of care satisfying the following criteria were deemed eligible for matching to the registry: (1) finished consultant episodes with a discharge date between 1 April 2014 and 1 March 2018 (to match the time frame of identifiable linkage); (2) a diagnosis of mitral insufficiency or a procedure code indicating mitral valve repair (see online supplemental file 1); (3) age over 17 years; (4) treatment at specific NHS Trust who submitted data to the UK MitraClip registry (see online supplemental file 2).

Individual episodes of care were aggregated into admissions (also known as spells) using the spell identification number (SUSSPELLID). Data cleaning was carried out to remove admissions if the spell number was missing or invalid. Non-unique admissions (where the same spell number was assigned to different patients) were also removed. Age was determined using the age on admission. Spells with missing discharge date were assigned the end date of the last episode in the admission (see online supplemental file 3).

Individual patient matching between the anonymised registry extract and the eligible admissions from HES was performed by a four-step algorithm using the following

variables: treating hospital, gender, age, admission date, procedure codes, procedure dates, discharge date ([figure 1B](#)). At each step, patients with no matches were excluded while those with multiple matches proceeded to the next step. Unique (1:1) matches from all steps were combined to give the final matching HES cohort.

The cohort was followed longitudinally by extracting all subsequent episodes of care from HES APC (discharged on or before 1 March 2018), and ONS mortality records (dated on or before 4 April 2018), using the unique patient identifier (ENCRYPTED_HESID) determined from the non-identifiable matching process. This ensured that the study period was the same for the identifiable and the non-identifiable linkage methods.

Extended follow-up of the cohort was also determined using the latest pseudonymised HES APC and ONS data available (discharge date or date of death up to and including 31 March 2020) to Newcastle upon Tyne Hospitals under the managed extract service.

Statistical analysis

Storage and querying of HES data were conducted using SQL (MariaDB). All scripts for case ascertainment, cleaning, processing and statistical analyses were written in the statistical programming language R.⁷

Descriptive statistics were used to describe the baseline characteristics of patients matched using the identifiable and non-identifiable data linkage methods, including: gender, age, diabetes status, critical pre-op status, mitral regurgitation aetiology, admission method.

Long-term outcomes were determined for all matched patients. Each patient was followed in the HES APC and ONS datasets from the date of the MitraClip procedure until the date of discharge (or for patients where MitraClip device was not successfully implanted, until the end of the procedural admission), the date of death or the latest date included in the linked HES data. Kaplan-Meier analysis was conducted for total all-cause mortality. Specific cardiac interventions following mitral valve implantation (eg, mitral valve intervention, cardiac pacemaker insertion and implantation of a cardioverter defibrillator) were identified from readmissions by searching for the Office of Population Censuses and Surveys (OPCS) procedure codes described in online supplemental file 4.

Role of the funding source

NHS England commissioned and funded a fixed number of MitraClip procedures within the CtE scheme. NHS England also commissioned NICE to facilitate the evaluation through Newcastle External Assessment Centre. Staff at NICE contributed to the design and conduct of the study, interpretation of the results, review and approval of the manuscript. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

RESULTS

Records of 278 MitraClip procedures were extracted from the registry by the data controller; 79 did not meet the study inclusion criteria (multiple reasons for exclusion may apply: 57 with non-eligible indications, 56 with concomitant treatment, 7 with unknown/none/moderate mitral regurgitation, 4 due to rheumatic aetiology, 4 missing/inconsistent procedural dates) leaving 199 procedures from 197 patients eligible for matching. The earliest procedure was conducted in January 2015 and the latest in January 2018.

From identifiable linkage, a total of 187 patients (187/197, 94.9%) were matched with the HES APC dataset, however 21 of the matched patients (11.2%) had inconsistencies across the datasets (eg, different gender) (online supplemental file 5). There were no significant differences in baseline characteristics between patients who passed the internal quality checks (n=166) and those who did not (n=21) in terms of age, sex, body surface area, diabetes, critical pre-op status, Canadian Cardiovascular Society angina status, New York Heart Association classification dyspnoea status, Killip class of heart failure, Canadian Study of Health and Aging frailty score, aetiology of mitral regurgitation and Katz index of independence. Therefore, following identifiable linkage, a total of 166 patients (84.2% matching) remained for subsequent analysis.

Using non-identifiable data, a total of 169 patients (169/197, 85.8%) were uniquely matched with the HES APC dataset using a four-step matching algorithm (online supplemental file 5). Any patients with conflicting information across data sources would not have passed the matching algorithm, and therefore all patients matched via this method remained for subsequent analysis.

Table 1 Baseline characteristics for matched cohorts

	Identifiable linkage (n=166)	Non-identifiable linkage (n=169)
Female gender	52 (31.3%)	50 (26.6%)
Median age, years (Q1:Q3)	78.5 (69:85)	78 (68:85)
Diabetes	30 (18.1%)	34 (20.1%)
Critical pre-op status	3 (1.8%)	4 (2.4%)
Functional/ischaemic aetiology	100 (61.7%)	103 (60.9%)
Urgent/emergency procedure	20 (12.0%)	19 (11.2%)

*Note that diabetes status which included any of the following was converted to presence of diabetes: diabetes (dietary control), diabetes (oral medicine), diabetes (insulin), newly diagnosed diabetes.

Table 2 Outcomes for matched cohorts

	Identifiable linkage (n=166)	Non-identifiable linkage (n=169)
Total patient days followed	99 241	89 736
Follow-up per patient, days		
Mean (SD)	597.8 (283.1)	531.0 (307.1)
Median (Q1:Q3)	610 (409:807)	560 (355:764)
Range	2–1128	0–1128
All-cause deaths	37	42
1-year survival	0.873 (0.823 to 0.925) (n=133)	0.833 (0.778 to 0.893) (n=122)
2-year survival	0.773 (0.706 to 0.847) (n=58)	0.725 (0.652 to 0.807) (n=50)
Subsequent intervention, patients		
Mitral valve replacement	6	6
Mitral valve repair	2	2
Other mitral intervention	1	0
Cardiac pacemaker insertion	10	12
Cardioverter defibrillator	2	3

Baseline patient characteristics for matched patients using the two methods of data linkage is described in [table 1](#).

Median length of follow-up was 610 and 560 days following identifiable (n=166) and non-identifiable (n=169) linkage methods, respectively. The total number of deaths (all causes) was 37 and 42 across the two linked datasets, respectively, and the difference in subsequent cardiac interventions identified between the two methods was negligible, [table 2](#).

Extended follow-up using non-identifiable linkage and latest available data enabled a median (Q1:Q3) length of follow-up of 1161 (597:1458) days, where a total of 78 deaths were recorded. Survival at 1-year (n=136) 0.841 (0.786 to 0.899), 2-year (n=119) 0.736 (0.671 to 0.807), 3-year (n=96) 0.617 (0.546 to 0.697) and 4-year (n=42) 0.499 (0.422 to 0.590) follow-up are shown in [figure 2](#).

DISCUSSION

Key points

Our study transparently describes the identifiable and non-identifiable linkage of a bespoke clinical procedural registry to routine healthcare data sets. Equivalent matching rates were achieved using the two methods, and similar baseline characteristics and 2-year outcomes were demonstrated for patients matched by either technique.

Strengths and limitations of linkage using identifiable data

To collect identifiable patient data, explicit approvals are required by an independent body: either by an

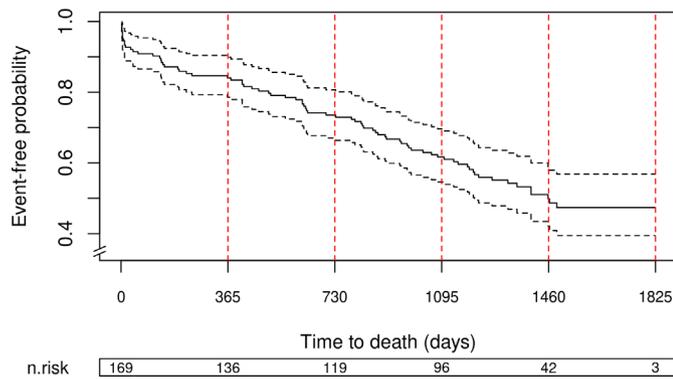


Figure 2 Kaplan-Meier survival curve derived from extended follow-up using non-identifiable linkage.

ethics board alongside a Caldicott Guardian with patient consent, or as in the case of the MitraClip registry, section 251 of the National Health Service Act 2006 approval such that the common law duty of confidentiality is temporarily lifted. All of these processes ensure that minimal patient identifiable information is collected for a specified purpose and safeguarding of those data. Although there are additional governance steps to undertake and the application process for obtaining matched data can be time consuming and costly, a key benefit of using patient identifiers (like NHS number) for subsequent matching to other data sources is that it does not require any clinical or clinical coding knowledge on the part of the researcher, with the process being fully automated and conducted independently by a trusted third party.

However, matching rates using identifiable sources may be reduced by certain factors; type 2 patient opt-outs are applied to identifiable HES data, with a rate of 2.7% reported nationally in March 2019.⁸ Furthermore, matching multiple data sources which rely solely on patient identifiers is sensitive to manual transcription errors (eg, mistyping of NHS number, date of birth and/or postcode) or missing data.⁹ Therefore, as demonstrated in this study, following the results of automated matching, an additional sense check/validation stage is still required to ensure matching to the correct person.

Strengths and limitations of linkage using non-identifiable data

To conduct linkage using non-identifiable data, variables including demographic details (gender, age) and administrative data (treating hospital dates, type of procedure) are used, thus combining matching and validation all within a single process. There is no requirement for identifiable information, which is beneficial from a data security and information governance perspective. If no identifiable information is being collected, then there is also no requirement to seek section 251 approval where explicit consent has not been obtained.

However, additional skills are required to conduct linkage using non-identifiable information. Access to clinical and clinical coding expertise is necessary to provide insight into the clinical pathway, to determine

relevant procedure and diagnostic codes for analyses and to identify the relevant subgroup for matching. Knowledge of HES data quality and cleaning processes is also required. For example, an individual patient is not always assigned the same identifier and, rarely, different patients may be assigned the same identifier. Hagger-Johnson *et al* previously demonstrated that using the patient identifier 'HESID' generated by NHS Digital resulted in a false match rate of 0.2% and missed match rate of 4.1% in paediatric intensive care records in England, leading to an under-estimate in readmission rates.¹⁰ Additionally, analysis of spells (collections of care episodes within a single admission) reveals inconsistencies that point to underlying data quality problems, such as duplications of records, missing information or inconsistencies, or activity recorded after death.¹¹ This necessitates additional cleaning of HES data to identify duplicated, inconsistent or missing spell information, and overlapping spells all of which allow removal of ineligible patients prior to matching.

We have demonstrated that non-identifiable data linkage works well for procedures with well-defined OPCS/International Classification of Diseases (ICD) coding combinations such as the MitraClip procedure, but this may not be the case for all clinical interventions; for example, those which have less specific clinical coding, but also those conducted out with the inpatient/day-case setting (where quality and completeness of clinical coding may differ), and high volume procedures where the likelihood of having multiple patients treated with same age and gender on same day are high. Further research is required to confirm that the matching success reported for the NHS England MitraClip registry is achievable for a range of other interventional procedures. Access to national pseudonymised administrative data was required to conduct the non-identifiable data linkage. With this comes responsibility for protecting the confidentiality of information for all potentially matching patients as well as for the cohort of interest. Each study applying data linkage of multiple data sources using non-identifiable information must recognise the potential for re-identification.¹² Our linkage algorithm shows that unique matching is possible with non-identifiable fields, suggesting that pseudonymised extracts carry confidentiality implications comparable to identifiable datasets. For this reason, all uses of HES data, including anonymous and pseudonymous matching proposals are reviewed by an independent panel (Independent Group Advising on the Release of Data, IGARD) to ensure safeguarding of patient data and subsequent data handling and processing by the approved institution are subject to audit by NHS Digital. Additionally, the guidelines for publishing results of such studies should be followed.¹³ Other initiatives for safe data linkage of identifiable data do exist, for example, the 'Separation of functions' offered by the Scottish Informatics and Linkage Collaboration.¹⁴

Overall strengths and limitations of linking to administrative datasets

The overall strengths of data linkage between clinical registries and routine data have been well documented, such as richer clinical information and the estimation of reporting bias.^{15–17} The main benefit of data linkage in this study, however, was the ability to conduct long-term comprehensive follow-up across all NHS Trusts in England. It has been reported that MitraClip appears to confer immediate improvements in cardiac indices of patients in line with published trial data,⁴ but long-term outcomes have not been published. Our techniques have delivered longer term outcome data and offer the capability for analyses to be repeated at 5 and 10 years as required, demonstrating a way of conducting active surveillance of medical procedures using routine data sources.¹⁸ This has the potential to improve understanding of the safety and efficacy of an intervention (particularly where long-term complications are likely to be detected outside the centre responsible for an intervention) and thereby to inform refinement of procedural selection to maximise long-term effectiveness.

Our study validated outcomes following MitraClip intervention by two separate techniques of data linkage between a clinical registry and routine healthcare data. However, the quality of collected data is crucial to the success of any data linkage. Clinical registries and routine administrative datasets both require high quality data submission as a single mismatch between the two can force the exclusion of a patient from the study. This work has highlighted the many benefits of data linkage to routine databases and thus strongly advocates the adoption of high quality data entry protocols and data validation in registries intended for health technology assessment.

This study has highlighted several lessons to be learnt for future linkage of clinical data to routine administrative data, whichever linkage method is used. As far as registry design is concerned, improvements could be achieved by incorporating input data validation. Also, to ensure easier identification during matching, mandatory coding of procedures using pre-specified OPCS codes should be used by treating hospitals contributing data to clinical registries. Comorbidities and adverse events could also be captured in registries using ICD codes, which would also be beneficial when conducting matching. Interim data linkage to identify potential data entry errors before final linkage is advisable.

Conclusion

This study has demonstrated the linkage of patient data from a bespoke clinical registry with routine healthcare databases via two equivalent methods to gain comprehensive and reliable follow-up from a cardiac intervention conducted in the ‘real world’ NHS in England. Linking to administrative data, by either method, would limit the administrative burden of future observational research. Here we have described a novel method that uses non-identifiable data, and in cases where robust clinical coding

can be specified, it could be applied to other hospital interventions. Studies using the technique must recognise the potential for re-identification. In this study, both data sources were pseudonymised but in cases where the study team may have access to identifiable patient information consent must be obtained to undertake linkage.

Furthermore, to ensure robust and generalisable results from clinical registries and routine databases, data validation at the point of data entry and data cleaning following linkage are essential steps in the analysis methods. Interim data linkage to identify and correct potential data quality issues during the course of data collection are also strongly recommended.

Acknowledgements HES and ONS data held by NHS Digital (formerly the UK NHS Health and Social Care Information Centre, HSCIC) have been used to help complete the analysis © 2019. Reused with the permission of the NHS Digital/HSCIC. All rights reserved. Clinical coding advice was provided by the clinical coding department within Newcastle upon Tyne Hospitals NHS Foundation Trust. NHS England’s specialised services Clinical Panel for granting permission to use the Commissioning through Evaluation data.

Contributors All authors conceived and designed the study. KK, PC, SU, JB, HC and AS analysed the data and were responsible for the statistical analysis. All authors interpreted the data. All authors contributed to the manuscript.

Funding The UK MitraClip registry and the identifiable data linkage with NHS Digital under DARS-NIC-151212-B523R agreement was funded by NHS England. NICE were contracted by NHS England to oversee the Commissioning through Evaluation scheme. The Newcastle upon Tyne Hospitals NHS Foundation Trust hosts an External Assessment Centre funded by NICE. The Newcastle upon Tyne Hospitals NHS Foundation Trust funded the national anonymised extracts provided under DAR-NIC-170211-Z1B4J agreement.

Competing interests The Newcastle upon Tyne Hospitals NHS Foundation Trust, the employing institution of KK, PC, SU, JB, HC and AS, is contracted as External Assessment Centre to the NICE Medical Technologies Evaluation Programme (MTEP) and is contracted by Academic Health Science Network North East and North Cumbria to develop methodologies and case studies relating to ‘evaluation in practice’ in the context of using routine healthcare datasets and, where appropriate, clinical registries, to assess outcomes and adoption of novel interventions. AS reports grants from NIHR and Wellcome Trust and outside the submitted work. KK reports grants from NIHR outside the submitted work. HP and LB are employed by NICE and were contracted by NHS England to oversee the Commissioning through Evaluation scheme. SG is employed by NHS Digital. HP and LB are employed by NICE and were contracted by NHS England to oversee the Commissioning through Evaluation scheme. No other financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Hospital Episodes Statistics data to reproduce results are available from NHS Digital via a formal application process. The MitraClip CTE registry data is owned by NHS England; applications to access the data can be made to NHS England’s specialised services Clinical Panel.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

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 iDs

Kim Keltie <http://orcid.org/0000-0001-5108-6279>
 Paola Cognigni <http://orcid.org/0000-0001-5418-3103>
 Samuel Urwin <http://orcid.org/0000-0003-4304-1358>
 Julie Burn <http://orcid.org/0000-0002-2300-2725>
 Helen Cole <http://orcid.org/0000-0003-3036-3374>
 Andrew Sims <http://orcid.org/0000-0002-9553-7278>

REFERENCES

- Charlton V. NICE and Fair? Health Technology Assessment policy under the UK's National Institute for Health and Care Excellence, 1999-2018. *Health Care Analysis* 2019;1–35.
- Bowman L, Baras A, Bombien R, *et al.* Understanding the use of observational and randomized data in cardiovascular medicine. *Eur Heart J* 2020;41:2571–8 <https://doi.org/10.1093/eurheartj/ehaa020>
- NHS England. Commissioning through evaluation. Available: <https://www.england.nhs.uk/commissioning/spec-services/npc-crg/comm-eval/> [Accessed 12th Aug 2019].
- Willits I, Keltie K, Urwin S. Nice observational data unit: commissioning through evaluation (CTE) percutaneous mitral valve leaflet repair for mitral regurgitation (MitraClip), 2018. Available: <https://www.england.nhs.uk/commissioning/wp-content/uploads/sites/12/2019/04/Percutaneous-mitral-valve-leaflet-repair-for-mitral-regurgitation-Commissioning-through-Evaluation-Report.pdf> [Accessed 12th Aug 2019].
- Harron K, Dibben C, Boyd J, *et al.* Challenges in administrative data linkage for research. *Big Data Soc* 2017;4:205395171774567–12.
- Padmanabhan S, Carty L, Cameron E, *et al.* Approach to record linkage of primary care data from clinical practice research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol* 2019;34:91–9.
- R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria. Available: <http://www.R-project.org/> [Accessed 13th Aug 2019].
- NHS Digital. National data Opt-out statistics, 2019. Available: <https://digital.nhs.uk/data-and-information/publications/statistical/national-data-opt-out/march-2019/ndop-mar19> [Accessed 13th Sep 2019].
- Hagger-Johnson G, Harron K, Goldstein H, *et al.* Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. *BMJ Health & Care Informatics* 2017;24:891.
- 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.
- Keltie K, Elneil S, Monga A, *et al.* Complications following vaginal mesh procedures for stress urinary incontinence: an 8 year study of 92,246 women. *Sci Rep* 2017;7:12015.
- Problems with care.data and other stories.... *British Medical Journal* 2015;351:h4613.
- NHS Digital. Hospital episode statistics (Hes) analysis guide, December, 2019. Available: https://digital.nhs.uk/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes_analysis_guide_december_2019-v1.0.pdf [Accessed Last accessed 15th January 2021].
- Scottish Informatics and Linkage Collaboration. Data linkage safeguards. available from. Available: <http://www.datalinkagescotland.co.uk/data-linkage-safeguards> [Accessed Last accessed 15th January 2021].
- Boyd JH, Randall SM, Ferrante AM, *et al.* Accuracy and completeness of patient pathways--the benefits of national data linkage in Australia. *BMC Health Serv Res* 2015;15:312.
- Burn J, Sims AJ, Keltie K, *et al.* Procedural and short-term safety of bronchial thermoplasty in clinical practice: evidence from a national registry and hospital episode statistics. *Journal of Asthma* 2017;54:872–9.
- Field K, Kosmider S, Johns J, *et al.* Linking data from hospital and cancer registry databases: should this be standard practice? *Intern Med J* 2010;40:566–73.
- Garies S, Youngson E, Soos B, *et al.* Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. *BMJ Health Care Inform* 2020;27:e100161.