

Novel linkage approach to join community-acquired and national data

Tochel, Claire; Pead, Emma; McTrusty, Alice; Buckmaster, Fiona; MacGillvray, Tom; Tatham, Andrew ; Strang, Niall; Dhillon, Baljean; Bernabeu, Miguel

Published in:
BMC Medical Research Methodology

DOI:
[10.1186/s12874-024-02143-3](https://doi.org/10.1186/s12874-024-02143-3)

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in ResearchOnline](#)

Citation for published version (Harvard):

Tochel, C, Pead, E, McTrusty, A, Buckmaster, F, MacGillvray, T, Tatham, A, Strang, N, Dhillon, B & Bernabeu, M 2024, 'Novel linkage approach to join community-acquired and national data', *BMC Medical Research Methodology*, vol. 24, 13. <https://doi.org/10.1186/s12874-024-02143-3>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please view our takedown policy at <https://edshare.gcu.ac.uk/id/eprint/5179> for details of how to contact us.

RESEARCH

Open Access



Novel linkage approach to join community-acquired and national data

Claire Tochel^{1*}, Emma Pead², Alice McTrusty², Fiona Buckmaster², Tom MacGillivray², Andrew J. Tatham^{2,3}, Niall C. Strang⁴, Baljean Dhillon^{2,3} and Miguel O. Bernabeu¹

Abstract

Background Community optometrists in Scotland have performed regular free-at-point-of-care eye examinations for all, for over 15 years. Eye examinations include retinal imaging but image storage is fragmented and they are not used for research. The Scottish Collaborative Optometry-Ophthalmology Network e-research project aimed to collect these images and create a repository linked to routinely collected healthcare data, supporting the development of pre-symptomatic diagnostic tools.

Methods As the image record was usually separate from the patient record and contained minimal patient information, we developed an efficient matching algorithm using a combination of deterministic and probabilistic steps which minimised the risk of false positives, to facilitate national health record linkage. We visited two practices and assessed the data contained in their image device and Practice Management Systems. Practice activities were explored to understand the context of data collection processes. Iteratively, we tested a series of matching rules which captured a high proportion of true positive records compared to manual matches. The approach was validated by testing manual matching against automated steps in three further practices.

Results A sequence of deterministic rules successfully matched 95% of records in the three test practices compared to manual matching. Adding two probabilistic rules to the algorithm successfully matched 99% of records.

Conclusions The potential value of community-acquired retinal images can be harnessed only if they are linked to centrally-held healthcare care data. Despite the lack of interoperability between systems within optometry practices and inconsistent use of unique identifiers, data linkage is possible using robust, almost entirely automated processes.

Keywords Community optometry, Data linkage, Early disease detection, Longitudinal data, Image analysis

*Correspondence:

Claire Tochel
ctoichel@ed.ac.uk

¹Centre for Medical Informatics, University of Edinburgh, Edinburgh, UK

²Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

³Princess Alexandra Eye Pavilion, NHS Lothian, Edinburgh, UK

⁴Department of Vision Sciences, Glasgow Caledonian University, Glasgow, UK



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Retinal images, either from fundus photography or from imaging devices such as optical coherence tomography provide valuable insights into ophthalmic and systematic diseases. This has the potential to improve the ability to detect sight or life-threatening conditions at an earlier stage. Development of robustly labelled, real-world datasets is essential for obtaining the large number of images need to develop and validate tools for disease detection and risk stratification, however, to date the majority of such datasets have been derived from images obtained in secondary care and therefore may not be generalisable if tools are to be used in primary care or population-based settings.

There is widespread public support for the use of linked, routinely collected data to support health research [1]. This support is conditional on perceived public benefit and concerns about potential harms. Awareness about existing practice around data linkage is reported to be low. The need for ‘adaptive governance’ which remains responsive to both researcher and public needs and expectations (which may change over time) has been highlighted as a key element in maintaining public acceptability in this field [2].

Scotland has several decades of health data linkage between hospital discharges, cancer registrations, medications and deaths [3–5]. The process of robustly linking electronic records benefited from the development in the 1970s of a national unique identifier (Community Health Index (CHI)), initially applied to primary care data [6]. Multiple National Health Service (NHS) healthcare datasets are now routinely collated centrally by Public Health Scotland and populated with CHI [7–9]. Accurate and efficient linkage of information for the same individual from different sources, is fundamental to allow meaningful and up to date research.

Scotland has a unique model of primary care eye examinations which, since 2006, have been funded by the NHS, removing the cost barrier to accessing eye care. Eye examinations are conducted by community optometrists and involve a thorough assessment of eye health, including fundus examination, visual field assessment, and refractive error. Retinal photography became standard for all patients over 60 years of age in 2008 [10]. This means that many community optometry practices have been capturing and storing retinal images for well over a decade and nationally optometrists now take millions of retinal images every year. The large number of retinal photographs obtained across the older population including both healthy and diseased individuals provides a valuable potential resource for longitudinal analysis, particularly as images acquired in primary care as part of the routine eye examination are likely to make the dataset

more representative of the population in which screening or early diagnostic tools would be employed.

The Scottish Collaborative Optometry-Ophthalmology Network e-research (SCONe) project was set up [11] as a result of a shared vision among clinicians, researchers and patient-supporting charities, that these retinal images could yield enormous benefit beyond the delivery of individual health care [12–14]. The SCONe resource would be a valuable addition to the current array of ophthalmic datasets many of which are created with a bias towards inclusion of cases with one particular disease [15]. To achieve this, SCONe is retrieving retinal images captured routinely at community optometry practices and linking them to NHS data within the Scottish National Safe Haven (NSH), creating a longitudinal research resource to support development of new technologies for early detection of eye disease, risk prediction, and discovery of retinal biomarkers of body and brain health [13, 14]. The NSH provides a secure technical and governance framework for linked data projects including secure access to NHS data, and ethical approval for research conducted there (<https://www.ed.ac.uk/edinburgh-international-data-facility/services/safe-haven-services/scottish-national-safe-haven>).

The benefits and risks of bringing separate datasets together to support health research are well documented [16, 17]. Ideally a unique and common identifier is available [18, 19]. Without this, it is necessary to develop a bespoke approach to matching records using robust techniques which must be guided by features in the data. Probabilistic (which calculate the likelihood of a match comparison) and deterministic (which result in a binary yes/no result) matching methods potentially bring together records with different degrees of efficiency, sensitivity and specificity [20–22]. For this work to be done at scale, manual intervention must be minimised, without introducing false matches.

Most optometry data are collected in busy patient environments and in systems which may not be interoperable with each other. Whereas these data meet the requirements of the practice to provide healthcare services, they may not meet the stringent quality requirements for research through linkage to other healthcare data, such as the routine inclusion of CHI. The SCONe Proof of Concept study aim was to test the technical feasibility of exporting colour fundus photographs for patients aged 60 and over from practice camera devices, with enough patient information to facilitate linkage to national hospital and ophthalmic data within the NSH [23].

The objectives of this study were to develop a novel automated linkage process which accurately identified PMS records for patients with retinal images captured in community optometry practices which are commonly

identified only by forename, surname and date of birth, and validate this against manual linkage.

Methods

Following SCONE's early engagement (surveys, presentations, newsletters, social and mainstream media articles) with community optometry across Scotland in 2020, some highly motivated practitioners (optometrists and practice directors) came forward to volunteer as pilot practices for the Proof of Concept study. The NHS Scotland Public Benefit and Privacy Panel for Health and Social Care approved the study in 2021. The NSH is the host of an existing imaging project which hosts radiological scans for research, and SCONE is working with the same team [24, 25].

A data sharing agreement was set up between each pilot practice's data controller and the study's co-sponsors (University of Edinburgh and NHS Lothian) and the authors arranged site visits to each practice in turn. Practices provided details of the hardware and software used by the practice to store retinal images and patient data. Test exports were run within practice in advance where possible to establish whether password access would be necessary at any point in the process, gauge the time required to carry out each task, verify that outputs would be in a usable format and explore any additional technical input which might be necessary. Additional information was sought from device manufacturers.

For robust linkage to national data to be performed within the NSH, six fields are required: patient forename, surname, date of birth (DOB), sex, address, and postcode. In optometry practices the Practice Management System (PMS) is used to register new patients, manage bookings and store clinical information; PMS records typically contains all six required linkage fields. Retinal images captured in practice are often stored in structured query language databases and the associated patient records tend to contain only the patient's forename, surname, and DOB. On site, PMS data were exported to a spreadsheet and images were exported from the databases as jpeg, tiff or png images with an associated extensible markup language file containing image filename, forename, surname, and DOB.

Data preparation

Patient forename and surname from the image device and PMS were first cleaned by converting all characters to lower case and removing non-letter characters or spaces using R v4.1.3 [26]. Match comparisons were created using "phonics" and "stringdist" packages.

Developing "ground truth" for match comparison

Within each of two test practices, an inner join operation was applied between image and PMS data, with

perfect matches identified and removed from subsequent matching steps. Assessment of the PMS data and imaging records which did not match, revealed many potential missed matches due to slight differences. These appeared to include both intentional and unintentional differences with apparent 'errors'. Perceived intentional differences included truncation or expansion of names and unintentional ones included incorrect dates of birth, spelling errors, missing or extra spaces, hyphens and apostrophes. For the purpose of matching these were all considered 'errors' in patient information. They were more prevalent on the camera device exports compared to the more complete PMS records.

Review and discussion of these cases by the authors, supplemented by professional experience in optometry practices led to a consensus around the extent of error which should be tolerated and therefore which potential matches to include and reject. To create optimal matching rules for the algorithm, several different deterministic and probabilistic matching techniques were explored iteratively. We evaluated the range of acceptable typographical errors and alternative names identified without inadvertently linking records erroneously. Based on the literature, deterministic rules were considered preferable to probabilistic rules to minimise the need for manual review of scores in the algorithm [22].

A list of manual matches was created for each practice (i.e. the "ground truth") based on individual review of camera device records which had not matched to a PMS patient record, but for which the author perceived that the mismatch fitted within the agreed rules and that the mismatch was likely due to an error. Whereas this approach identified many more linkable images, it was very time consuming (approximately three person-days for a single practice with 2,000 unmatched images) and included the risk of human error.

Based on the information gathered from the first two practices, we developed a sequential set of rules to match patients from the image device data to the PMS patient list with a high true positive rate and low false positive rate compared to the manual matching process. We tested this same sequence in three further practices.

The following definitions were used to calculate true and false positive rates when comparing the matching done via the automated sequence to the manual process (Fig. 1):

- true positive: image was associated with the same individual in the automated and manual match lists
- false positive: image was associated with a different individual in automated and manual match lists
- false negative: image was not associated with an individual in automated list but was in manual list
- true negative: image was not associated with an individual in automated or manual match lists.

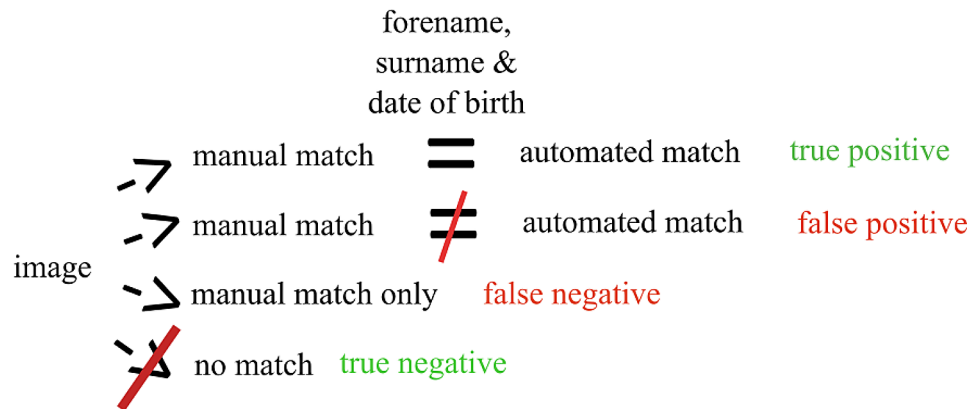


Fig. 1 Schematic showing definition of the four potential outcome matches for an image comparing manual and automated matching processes

Results

Development of automated matching sequence

The most useful techniques for matching names were selected based on iterative exploration of the early pilot practices. Variations such as Muhammad and Mohamed were matched by converting each string to a Soundex code (M530) based on a phonetic comparison [26]. Truncations or expansions (e.g., “Elizabeth” to “Eliza”) could be detected using the Longest Common Substring (LCS) calculation [27]. This is based on the longest string that can be obtained by pairing characters in the two names while keeping the order of characters intact. The LCS distance is calculated from the number of unpaired characters meaning that low scores usually indicate stronger matches.

Other ‘edit distance’ calculations comparing names such as Levenshtein and Jaro-Winkler distance did not improve detection of matches [27]. The Jaro-Winkler distance (whereby a perfect match between two strings gives a score of one) in some cases returned a low score for what appeared to be clear matches during manual review. The Levenshtein distance (where zero indicates a perfect match, but the maximum score depends on the number of characters in the longer of the two strings) performed similarly to LCS, but as the latter was a simpler measure on which to apply a cut-off given the observed naming errors, it was used. A score of 6 was found to be a suitable threshold.

With respect to DOB, after review of commonly occurring errors, it was decided that an error in one number (day, month, year) or the reversal of day and month was acceptable in the presence of other matching details.

Validation

In the three validation practices, the first step (identical forename, surname and DOB) matched between 84% and 91% of images to a patient in the PMS. These cases were not further scrutinised. Manual matching typically added

a further 8-15% of images successfully linked to a PMS patient record.

The success of each automated matching step compared to the manual match list in each practice was calculated, with those matches removed from subsequent steps (Table 1).

1. Exact forename, surname and DOB
2. Exact forename and surname, DOB error tolerated
3. Exact DOB, Soundex forename and surname
4. Soundex match on forename and surname with DOB error tolerated
5. Exact DOB and surname, LCS forename under threshold (6 used in these tests)
6. Exact DOB and forename, LCS surname under threshold (6 used in these tests).

All of the false positive matches identified during the final two stages in practice 1 were manually reviewed, and were deemed to be true matches, but for which information had been manually corrected in the PMS data to fill missing details from another source. By eliminating this process from the 2nd and 3rd practices, no false positives were created using the automated linkage steps.

Beyond the pilot

Work on SCONE has continued with seven more practices visited at the time of writing. Further deterministic steps have been found to be necessary to catch obvious potential matches where new local anomalies in data entry were evident after the final step described above. For example, some practices include middle names with forename and double-barreled surnames in the PMS but not the image record. The additional steps added (but not validated with a full manual check) were: forename from image list contained entirely within PMS forename; surname from image list contained entirely within PMS surname and forename and surname reversed in image list. To minimise the two probabilistic matching steps

Table 1 Number of true and false matches made between image and Patient Management System records within each practice by each rule when compared to manual matching after step 1. True positive and false positive rates provided in brackets

Practice (number of patients)	Match step	True positive (rate)	True negative	False positive (rate)	False negative	Number of images
Practice #1 (2,339)						
	1	18,699 (0.859)	4,890	0 (0)	3,079	26,668
	2	19,338 (0.888)	4,881	0 (0)	2,449	26,668
	3	20,696 (0.95)	4,881	0 (0)	1,091	26,668
	4	20,735 (0.951)	4,866	0 (0)	1,067	26,668
	5	21,495 (0.986)	4,823	43 (0.009)	307	26,668
	6	21,547 (0.988)	4,788	78 (0.016)	255	26,668
Practice #2 (309)						
	1	1,003 (0.836)	803	0 (0)	197	2,003
	2	1,043 (0.869)	803	0 (0)	157	2,003
	3	1,133 (0.944)	803	0 (0)	67	2,003
	4	1,133 (0.944)	803	0 (0)	67	2,003
	5	1,155 (0.963)	803	0 (0)	45	2,003
	6	1,185 (0.988)	803	0 (0)	15	2,003
Practice #3 (133)						
	1	247 (0.911)	5	0 (0)	24	276
	2	249 (0.919)	5	0 (0)	22	276
	3	263 (0.97)	5	0 (0)	8	276
	4	263 (0.97)	5	0 (0)	8	276
	5	266 (0.982)	5	0 (0)	5	276
	6	269 (0.993)	5	0 (0)	2	276

Match steps:

1. Exact forename, surname and DOB
2. Exact forename and surname, DOB error tolerated
3. Exact DOB, Soundex forename and surname
4. Soundex match on forename and surname with DOB error tolerated
5. Exact DOB and surname, LCS forename < 6
6. Exact DOB and forename, LCS surname < 6

True positive rate = true positive / (true positive + false negative)

False positive rate = false positive / (false positive + true negative)

(and therefore the number of matches requiring manual review) these three steps have been added before using LCS. In subsequent practices the ranked match scores have been reviewed, and the score of 6 remains a useful threshold to minimise false positives, however occasionally obvious outliers can be matched by making a manual

intervention. All useful steps identified to date are shown schematically in Fig. 2.

Discussion

We have developed an algorithmically-defined matching sequence with deterministic and probabilistic rules which, applied to limited data (three identifying fields







	PMS data		Image device data	Match rule	Probabilistic (P) or deterministic (D)
	claire smith		claire smith	exact	D
validated rules	dob: yyyy/mm/dd		dob: yyyy/dd/mm ----/mm/dd yyyy/- -/dd yyyy/mm/- -	tolerated errors	D
	claire smith		clare smyth	Soundex: phonetic	D
	claire smitherson		claire smith	Longest Common Substring	P
additional	claire smith		smith claire	fname & sname reversed	D
	claire marie smith		marie smith	name contained	D

Fig. 2 Schematic of deterministic (D) and probabilistic (P) rules used in the matching algorithm. Green indicates an exact match, amber indicates that the match between the image and Patient Management System (PMS) record met a pre-defined rule. Date of birth (dob) is based on year (yyyy), month (mm) and day (dd)

for each image), facilitated rapid matching to a more detailed patient record, on a different and unconnected system with high true and low false positive rates. The first step (perfect match of three fields) yielded at least an 84% match and application of the rules increased this to 99% with no false positives compared to ground truth. These rules were selected, based on a thorough review of the datasets in multiple practices, considering the known behavioural and technological factors at play in practice to assess the likelihood of error. Assuming that the manual match is adequately accurate, these examples suggest that the algorithm rules can deliver a robust and efficient matching process for future practices including those with very large patient lists. The process continues to be developed and refined in response to new experiences in practice.

Real world data challenges

The lack of interoperability between image capture devices and PMS in most community optometry practice is challenging on multiple levels. It means that staff must manually enter patient details in order to take a retinal image, despite the full patient record already existing just metres away. This is inefficient and inevitably leads to the potential for error. From a research perspective, it makes the data harder to incorporate into healthcare datasets, as these must be linked to the patients’ unique identifier which needs accurate and detailed patient information.

The Scottish Government’s Digital health and care strategy, refreshed in 2021, emphasised the need for data to be used to benefit the country’s citizens, and acknowledged that people are frustrated at the lack of co-ordination and links between different parts of the healthcare system [28]. This work, and the SCONE project generally, is an example of the innovation required to securely

bring together elements of that system which are currently fragmented, to harness benefits which are impossible with data stored in isolation.

Patients whose images were not matched to the full PMS record, risk not being matched to their unique identifier (CHI) within the NSH and will therefore be missing from the linked research resource. If these individuals are lost at random from practice lists, then this will not be problematic other than the loss of beneficial data to any work done on the images. However, if there is a systematic reason behind the lack of matches, the cohort may not be representative of the community. For example we may disproportionately lose married women who have changed their surname (in systems where this is not retained in their record), or people from ethnic groups among which names don’t neatly fit into systems with a single forename and surname format. If an image is incorrectly matched to the wrong CHI, the wrong individual’s health records will be included in the linked dataset. Clearly this has the potential to lead to inaccurate research findings which would undermine the project, so methods for quality assessment need to be built in [16, 29]. It also poses a problem from the governance perspective, in that the project has permission to use records from a defined subset of the population. Both adverse implications must be avoided at all costs, and we have therefore taken a strict approach whereby the risk of losing true matches is preferable to including false matches.

Limitations

The risk of matching a patient’s image with a different patient’s record within each practice based on the three available fields, was considered to be much lower than a similar matching process conducted nationally (for which six fields are considered necessary). The relatively (and

in some cases absolutely) small number of patients and localised geographical area from which each practice population was drawn meant that this novel approach with the agreed error tolerance was considered acceptable to the authors. All practices involved in this exercise were independently owned with small numbers of staff inputting information into the image device and PMS. Different rules and error tolerance may be required if matching involved larger practices with more staff and more potential variation in data input. The data collection and systems are also likely to vary between practices meaning that there is likely no single set of rules which would apply optimally in every practice. However, these rules, based on a logical sequence of low risk measures, worked in practices with very different size of patient list (20,000 vs. 27,000) and we have continued to develop the process as we visit more sites. We currently do not have a method to verify image matching, i.e. does any given retinal image actually belong to the named patient.

Conclusion

To date the SCONE team have delivered retinal images, captured over many years, for 4,000 patients to NSH, where they were matched to CHI. This study describes the necessary intermediate step of linking retinal images stored with just three personal data fields, to the six fields required for CHI linkage, which facilitated the creation of a cohort of 28,947 community-acquired images linked to routinely-collected healthcare data within the NSH for the first time. Evaluating the images linked to NHS data will allow us to assess the potential they contain, and work towards establishing a rich, longitudinal retinal image repository, which could grow year on year with the full support of Scottish Government [30]. This would realise SCONE's potential to support the detection of pre-symptomatic disease and the development of improved diagnostic tools and treatments, directly benefiting the public whose images are its defining and most valuable feature.

Acknowledgements

We are extremely grateful to the optometrists who came forward to participate as pilot practices to support this early work, and their practice staff who facilitated our site visits. The authors would like to acknowledge the support of the electronic Data Research and Innovation Service Team (Public Health Scotland) for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven. We are also grateful to EPCC who provide the IT infrastructure of the National Safe Haven and have collaborated with the SCONE team to support the ingestion of images. The SCONE steering group includes representation from Scottish Government and Action Against AMD and the Foresight Consortium UK.

Author contributions

BD, NS, MB, AT and TMacG conceived of the SCONE project and established its scope and objectives. CT, EP, AMcT and FB deduced the required processes for data extraction from optometric systems and developed the necessary understanding of the processes involved in order to interpret the data. CT and

EP developed the matching algorithm and tested it. CT wrote the first draft of the manuscript, all co-authors reviewed it and contributed to the final version.

Funding

The authors gratefully acknowledge funding from: Fondation Leducq Transatlantic Network of Excellence (17 CVD 03); EPSRC grant no. EP/X025705/1; British Heart Foundation and The Alan Turing Institute Cardiovascular Data Science Award (C-10180357); Diabetes UK (20/0006221); Fight for Sight (5137/5138); the SCONE projects funded by Chief Scientist Office, Edinburgh & Lothians Health Foundation, Sight Scotland, the Royal College of Surgeons of Edinburgh, the RS Macdonald Charitable Trust, and Fight For Sight; the Neurii initiative which is a partnership among Eisai Co., Ltd, Gates Ventures, LifeArc and HDR UK.

Data availability

The datasets analysed during the current study are not publicly available because they were acquired from community optometrists who authorised their use through Data Sharing Agreements (once processed into the necessary specification) only within the National Safe Haven. After delivery to the National Safe Haven the images and linking data were deleted from University of Edinburgh systems as agreed in the permissions for the study. In the future we hope that the SCONE dataset (retinal images linked to pseudonymised healthcare data) will be available for use by researchers via the Public Health Scotland National Safe Haven. We are still in the early stages of developing this resource. The corresponding author is happy to prepare an anonymised version of the scripts used in the data processing to share on request.

Declarations

Ethics approval and consent to participate

SCONE is covered by the National Safe Haven ethical approval (from the East of Scotland NHS Research Ethics Service), as confirmed in writing by Public Health Scotland before the study commenced. The use of unconsented data within Safe Havens is approved by Scottish Government who set out guidelines for this in their Charter for Safe Havens [31]. The SCONE dataset (community-acquired retinal images linked to centrally-held healthcare data) is pseudonymised within the National Safe Haven, prior to being made available for research. The SCONE dataset is a novel resource which we are creating for the first time from data that were not previously linked. This paper describes part of the process of creating that novel resource. In order to link retinal images from the community to centrally held healthcare data, by definition, identifiable data must be used. The data were managed carefully and kept secure at all times, in accordance with all relevant guidelines and regulations. The study was co-sponsored by University of Edinburgh and NHS Lothian who oversaw the governance underpinning the work and endorsed its protocols. Data Sharing Agreements were set up between the study sponsors and every data controller (community optometrist) who agreed to participate. Approval to work with patient-level healthcare data was provided by the NHS Scotland Public Benefit and Privacy Panel for Health and Social Care (<https://www.informationgovernance.scot.nhs.uk/pbpphsc/>). This panel, which reports directly to Scottish Government, is the most senior NHS governance body in Scotland and assesses applications to work with healthcare data within the National Safe Haven.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 April 2023 / Accepted: 5 January 2024

Published online: 17 January 2024

References

1. Aitken M, De St. Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics*. 2016;17(1):70. <https://doi.org/10.1186/s12910-016-0153-x>.

2. Laurie G, Ainsworth J, Cunningham J, Dobbs C, Jones KH, Kalra D, et al. On moving targets and magic bullets: can the UK lead the way with responsible data linkage for health research? *Int J Med Informatics*. 2015;84(11):933–40. <https://doi.org/10.1016/j.ijmedinf.2015.08.011>.
3. Evans JM, MacDonald TM record-linkage for pharmacovigilance in Scotland. *Br J Clin Pharmacol*. 1999;49(1):105–10. <https://doi.org/10.1046/j.1365-2125.1999.00853.x>.
4. Fleming M, Kirby B, Penny KI. Record linkage in Scotland and its applications to health research. *J Clin Nurs*. 2012;21(19pt20):2711–21. <https://doi.org/10.1111/j.1365-2702.2011.04021.x>.
5. Harron K. In: Mackay E, Elliot M, editors. Introduction to data linkage. Administrative Data Research Network; 2016.
6. Scottish Government - Health Informatics Research Advisory Group, Health A, and Biomedical Informatics Research Strategy for Scotland. 2015; Available from: <https://www.gov.scot/publications/health-biomedical-informatics-research-strategy-scotland/pages/4/>.
7. Kendrick S, editor. The Development of Record Linkage in Scotland: The Responsive Application of Probability Matching. Record Linkage Techniques; 1997 04/12; Arlington, VA: Federal Committee on Statistical Methodology.
8. Lambourg EJ, Gallacher PJ, Hunter RW, Siddiqui M, Miller-Hodges E, Chalmers J, Pugh D, Dhaun N, Bell S. Cardiovascular outcomes in patients with chronic kidney disease and COVID-19: a multi-regional data-linkage study. *Eur Respir J*. 2022;60(5). <https://doi.org/10.1183/13993003.03168-2021>.
9. Robertson L, Ayansina D, Johnston M, Marks A, Black C. Measuring multimorbidity in hospitalised patients using linked hospital episode data: comparison of two measures. *J Popul Data Sci*. 2019;21. <https://doi.org/10.23889/ijpds.v4i1.461>.
10. Scottish Government - Directorate of Primary Care. General Ophthalmic Services - NHS Eye Examinations. Memorandum to NHS:PCA(O)(2008)3. 2008; Available from: [https://www.sehd.scot.nhs.uk/pca/PCA2008\(O\)03.pdf](https://www.sehd.scot.nhs.uk/pca/PCA2008(O)03.pdf).
11. SCONE. Scottish Collaborative Optometry-Ophthalmology Network e-research. [2023-03-27] Available from: <https://www.ed.ac.uk/clinical-sciences/ophthalmology/scone>.
12. Hagiwara Y, Koh JEW, Tan JH, Bhandary SV, Laude A, Ciaccio EJ, Tong L, Acharya UR. Computer-aided diagnosis of glaucoma using fundus images: a review. *Comput Methods Programs Biomed* 2018 165: 1–12 <https://doi.org/10.1016/j.cmpb.2018.07.012>.
13. Patton N, Aslam TM, MacGillivray T, Deary IJ, Dhillon B, Eikelboom RH, Yegesan K, Constable IJ. Retinal image analysis: concepts, applications and potential. *Prog Retin Eye Res*. 2006;25(1):99–127. <https://doi.org/10.1016/j.preteyeres.2005.07.001>.
14. Sengupta S, Singh A, Leopold HA, Gulati T, Lakshminarayanan V. Ophthalmic diagnosis using deep learning with fundus images - a critical review. *Artificial intelligence in medicine*. 2020 102 <https://doi.org/10.1016/j.artmed.2019.101758>.
15. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, Keane PA, Sebire NJ, Burton MJ, Denniston AK. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3(1):e51–e66. [https://doi.org/10.1016/s2589-7500\(20\)30240-5](https://doi.org/10.1016/s2589-7500(20)30240-5).
16. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol*. 2014;14(1):36. <https://doi.org/10.1186/1471-2288-14-36>.
17. Roos LL, Wall-Wieler E, Burchill C, Hamm NC, Hamad AF, Lix LM. Record linkage and big data-enhancing information and improving design. *J Clin Epidemiol* 2022 150:18–24 <https://doi.org/10.1016/j.jclinepi.2022.06.006>.
18. Doney ASF, Bonney W, Jefferson E, Walesby KE, Bittern R, Trucco E, Connelly P, McCrimmon RJ, Palmer CNA. Investigating the relationship between type 2 diabetes and dementia using. *Diabetes Care* 2019 42(10): 1973–80 <https://doi.org/10.2337/dc19-0380>.
19. Wagner SK, Hughes F, Cortina-Borja M, Pontikos N, Struyven R, Liu X, Montgomery H, Alexander DC, Topol E, Petersen SE, Balaskas K, Hindley J, Petzold A, Rahi JS, Denniston AK, Keane PA. AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ open*. 2022;12(3):e058552. <https://doi.org/10.1136/bmjopen-2021-058552>.
20. Karr AF, Taylor MT, West SL, Setoguchi S, Kou TD, Gerhard T, Horton DB. Comparing record linkage software programs and algorithms using real-world data. *PLoS ONE*. 2019;14(9):e0221459. <https://doi.org/10.1371/journal.pone.0221459>.
21. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *J Clin Epidemiol*. 2011;64(5):565–72. <https://doi.org/10.1016/j.jclinepi.2010.05.008>.
22. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *J Biomed Inform*. 2015;56:80–6. <https://doi.org/10.1016/j.jbi.2015.05.012>.
23. Public Health Scotland. National Data Catalogue. [28th September 2022]; Available from: <https://www.ndc.scot.nhs.uk/National-Datasets/>.
24. Nind T, Sutherland J, McAllister G, Hardy D, Hume A, MacLeod R, Caldwell J, Krueger S, Tramma L, Teviotdale R, Abdelatif M, Gillen K, Ward J, Scobbie D, Baillie I, Brooks A, Prodan B, Kerr W, Sloan-Murphy D, Herrera JFR, McManus D, Morris C, Sinclair C, Baxter R, Parsons M, Morris A, Jefferson E. I. An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. *Gigascience* 2020 9(10) <https://doi.org/10.1093/gigascience/giaa095>.
25. Public Health Scotland. Scottish Medical Imaging (SMI) Service. 2022; Available from: <https://www.isds.scotland.org/Products-and-Services/eDRIS/Scottish-Medical-Imaging-Service/>.
26. Bailey M, Cole C, Henderson M, Massey C. How well do automated linking methods perform? Lessons from U.S. Historical Data. *J Economic Literature* 2020 58(4):997–1044 <https://doi.org/10.1257/jel.20191526>.
27. n der Loo MPJ. The stringdist package for approximate string matching. *R J*. 2014;6(1):111–22. <https://doi.org/10.32614/rj-2014-011>.
28. Scottish Government - Digital Health and Care Directorate. Digital health and care strategy. 2021 [updated 2021]; Available from: <https://www.gov.scot/publications/scotlands-digital-health-care-strategy/>.
29. Goldstein H, Harron K. Methodological developments in data linkage. 2016 Chap. 6: record linkage: a missing data problem. 109 – 24. <https://doi.org/10.1002/9781119072454>.
30. Scottish Government Minister for Public Health Women's Health and Sport. Letter of support. 2022 Available from: https://www.ed.ac.uk/sites/default/files/atoms/files/thank_you_letter_to_baljean_dhillon_and_colleagues_-_ref_scone_project_-_dated_21_september_2022.pdf.
31. Scottish Government. 2015 A Charter for Safe Havens in Scotland. Available from: <https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/documents/>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.