












OPEN LETTER

REVISED Connected Bradford: a Whole System Data Linkage**Accelerator [version 2; peer review: 2 approved]**

Kuldeep Sohal¹, Dan Mason ¹, John Birkinshaw¹, Jane West ¹,
 Rosemary R.C. McEachan ¹, Mai Elshehaly ², Duncan Cooper³, Rob Shore ³,
 Michael McCooe¹, Tom Lawton ¹, Mark Mon-Williams ⁴, Trevor Sheldon ⁵,
 Chris Bates⁶, Megan Wood¹, John Wright ¹

¹Bradford Institute for Health Research, Bradford Hospitals National Health Service Trust, Bradford, BD9 6RJ, UK²Department of Computer Science, University of Bradford, Bradford, BD7 1DP, UK³Public Health, Bradford Metropolitan District Council, Bradford, BD1 1HX, UK⁴School of Psychology, University of Leeds, Leeds, LS2 9JT, UK⁵Institute of Population Health Sciences, Queen Mary University of London, London, E1 4NS, UK⁶The Phoenix Partnership, Leeds, LS18 5PX, UK

v2 First published: 27 Jan 2022, 7:26
<https://doi.org/10.12688/wellcomeopenres.17526.1>

Latest published: 07 Nov 2022, 7:26
<https://doi.org/10.12688/wellcomeopenres.17526.2>

Abstract

The richness of linked population data provides exciting opportunities to understand local health needs, identify and predict those in most need of support and evaluate health interventions. There has been extensive investment to unlock the potential of clinical data for health research in the UK. However, most of the determinants of our health are social, economic, education, environmental, housing, food systems and are influenced by local authorities.




The Connected Bradford Whole System Data Linkage Accelerator was set up to link health, education, social care, environmental and other local government data to drive learning health systems, prevention and population health management. Data spanning a period of over forty years has been linked for 800,000 individuals using the pseudonymised NHS number and other data variables. This prospective data collection captures near real time activity.

This paper describes the dataset and our Connected Bradford Whole System Data Accelerator Framework that covers public engagement; practitioner and policy integration; legal and ethical approvals; information governance; technicalities of data linkage; data curation and guardianship; data validity and visualisation.


Keywords

Routine population data; Local Government; population health; prevention

Open Peer Review**Approval Status**  

	1	2
version 2		
(revision)		
07 Nov 2022	view	view
		
version 1		
27 Jan 2022	view	view

1. **Jiang Bian**, University of Florida, Gainesville, USA

2. **Iain E. Buchan** , University of Liverpool, Liverpool, UK

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the [Born in Bradford gateway](#).

Corresponding author: Kuldeep Sohal (kuldeep.sohal@bthft.nhs.uk)

Author roles: **Sohal K:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mason D:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Review & Editing; **Birkinshaw J:** Data Curation, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Writing – Review & Editing; **West J:** Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **McEachan RRC:** Funding Acquisition, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Elshehaly M:** Data Curation, Funding Acquisition, Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Cooper D:** Investigation, Methodology, Supervision, Validation, Writing – Review & Editing; **Shore R:** Data Curation, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **McCooe M:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Lawton T:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mon-Williams M:** Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Sheldon T:** Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Supervision, Writing – Review & Editing; **Bates C:** Conceptualization, Investigation, Methodology, Resources, Software, Writing – Review & Editing; **Wood M:** Project Administration, Visualization, Writing – Review & Editing; **Wright J:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the National Institute for Health Research (NIHR) Public Health Research Programme (NIHR133648) with delivery support provided by the NIHR Clinical Research Network. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work is also supported by the UK Prevention Research Partnership (MR/S037527/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities (including Wellcome).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Sohal K *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Sohal K, Mason D, Birkinshaw J *et al.* **Connected Bradford: a Whole System Data Linkage Accelerator [version 2; peer review: 2 approved]** Wellcome Open Research 2022, 7:26 <https://doi.org/10.12688/wellcomeopenres.17526.2>

First published: 27 Jan 2022, 7:26 <https://doi.org/10.12688/wellcomeopenres.17526.1>

REVISED Amendments from Version 1

We have updated the manuscript in line with reviewer comments and addressed key points for the benefit of our readers. We have highlighted that there are novel and longitudinal international comparisons of linked databases that are used to inform policy and practice as well as other civic datasets in the UK including the Civic Data Cooperative model (<https://civicdatacooperative.com>). Whilst there are studies including the Born in Bradford programme which rely on patient consent as a legal basis to link data at an individual level, there is little information on how to overcome or mitigate legal and governance challenges in the development of research databases where patient consent does not exist.

In the UK, healthcare systems generally record and validate the NHS number which is a unique identifier assigned to an individual's; but there are caveats associated with using administrative data when using routinely recorded data to inform policy and practice.

We have also provided more granularities on the datasets acquired as part of the Connected Bradford database and updated each of the datasets to pseudonymised data linkages in Table 1.

We have provided more information on the legal, governance and technical data linkage of healthcare and education data as there is no single unique identifier embedded in health and education systems. In addition, we will be linking non-healthcare data such as housing and geospatial data using the Unique Property Reference Number (UPRN). The UPRN is a unique number that is allocated to each land or property. The UPRN is routinely recorded by the housing team in the local authority.

Any further responses from the reviewers can be found at the end of the article

Introduction

Routine electronic information about individuals is recorded in large quantities by healthcare professionals across different healthcare settings to identify, investigate, diagnose and treat patients. This information is an important and rich resource that enables healthcare professionals to support patient care. International examples of the benefits of such linkage include the PCORnet in the US¹ and the Danish national wide-ranging register². However, in many examples the datasets are built as pure research databases and are not operationalised to have a more direct impact on patient care. The UK is no different in this regard, information is not always linked for direct care purposes by the clinical systems across the different care settings resulting in a disconnect between healthcare organisations, fragmentation of care and incomplete pathways of care.

Opportunities to link individuals routinely recorded data across healthcare settings including primary care, secondary care, community care, urgent and emergency care and social services could be delivered by a unique reference number (NHS number) that is recorded by healthcare systems.

Using the NHS number, fragmented individual patient contacts within different healthcare settings can be linked within a population health management framework³. The framework offers insights from patient interactions at the GP practice, hospital and community settings. Over the last ten years there

has been extensive investment to establish data linkage across health care organisations with the goal of understanding patient pathways and harnessing linked routine data to drive health service improvement and research⁴.

However, it is accepted that the majority of ill-health is caused by wider determinants of health that lie outside the health service⁵. The data and intelligence that describes these factors is typically held by non-health care organisations such as local authorities, schools, housing associations, criminal justice, and environmental agencies. Linking non-health data with health service data at a local level and in ways that ensure a live information system usable by decision makers in real time, would transform our ability to understand the upstream influences on health, to design and test interventions to prevent ill-health, and to influence local decision making and shape policy^{6,7}.

A number of successfully linked data models operating across Bradford provided an ideal environment to use this experience to build a new collaborative solution, integrating research into practice using evidence-based interventions based on epidemiological results. One particularly successful model is the Born in Bradford birth cohort study⁸. Between 2007–2011 the Born in Bradford cohort obtained informed consent for 30,000 participants for linkage of health and education records. While the detailed research and biological data that the cohort collected has led to exciting scientific findings, the most useful intelligence for policy and practice has come from this linked routine data. For example, we were able to demonstrate the associations between air pollution and child health⁹ and use this evidence to design an ambitious clear air zone for the city. Evidence linking green space and mental health has been used to obtain investment to redesign parks. Linkage between birth data and school attainment¹⁰ has led to policy changes in school admissions for children who were premature. Linkage between schools Early Years Foundation Profile and autism diagnoses has been used to completely redesign child autism support¹¹.

Bradford has built on the Born in Bradford experience to establish a new Connected Bradford Whole System Data Linkage Accelerator for a much wider population that is representative of the entire Bradford population. Connected Bradford (cBradford) covers 800,000 citizens, five NHS Trusts, 86 general practitioners and 200 schools, and links pseudonymised health, education, social care, environmental and local government data. After extensive community consultation and citizen juries, the team have worked closely with NHS and local authority agencies to develop safe and secure data linkage and established a trusted research environment as part of the regional Yorkshire Health Care Record Exemplar (YHCR)¹².

Setting

Bradford is a post-industrial city in the North of England with high levels of deprivation and poor health, and a multi-ethnic population including a large Pakistani community and growing communities of East European and Roma people. Bradford is governed locally by Bradford Metropolitan District Council (BMDC)

which is the 4th largest metropolitan council in England. BMDC serves a population of 534,300 and covers an approximate area of 141 square miles¹³. The NHS Bradford District and Craven Clinical Commissioning Group (CCG) came into operation on 1 April 2020 and supports a population of almost 600,000. The CCG commissions hospital services, urgent and emergency care services and supports the Bradford and Craven population with their mental health and wellbeing from Bradford Hospitals NHS Foundation Trust (BTHFT), Airedale NHS Foundation Trust, Bradford District Care Trust, 86 General Practices and other community care organisations.

Connected Bradford explored existing linked datasets across these providers which were truly characteristic of the population of 600,000 and identified early on that this did not exist.

Limitations with existing linked datasets including the Born in Bradford study did not capture all age groups within Bradford's population. Other linked datasets were either anonymised and had no flexibility to link additional datasets including datasets from the wider determinants of health or enable analysts and researchers to explore for secondary use analysis or for research purposes.

These limitations provided the catalyst to develop the Connected Bradford Whole System Data Linkage Accelerator.

Ethical approval

Ethical approval for a research database was sought to provide reassurances to data providers and develop research studies by making use of the linked dataset that covers both the entire Bradford population and the wider Yorkshire region. The Bradford Institute for Health Research applied to the East Midlands – Derby Research Ethics Committee due to the committee's existing experience with research databases and ethical approval as a research database to Connected Bradford was granted on 31 August 2017 (IRAS ref:227117 and REC ref:17/EM/0254). A further amendment was submitted to the East Midlands – Derby Research Ethics Committee to include a) add additional datasets to the database b) allow GPs to opt out of the wider use of the data for all pathway projects identified by the programme c) allow GPs to opt out of the use of Apollo to extract the pseudo data from the GP practice and d) extend the availability of the database to external researchers. Support from the East Midlands – Derby Research Ethics Committee was granted on 20 March 2019. Ethical approval for a further five years was granted on 13 June 2022 and the updated REC reference is 22/EM/0127.

However, to support the programme's vision to link healthcare data with education data, we identified that there would only be a limited number of individuals that would have both their healthcare and education data recorded by the respective organisations. Education records are in existence for those that are born from 1991 to date where individuals have attended schools in the UK and reference a Unique Pupil Number (UPN) which is not made available or recorded routinely by healthcare organisations. Similarly the NHS number is not routinely available

to education departments and thereby data linkage using a unique reference number across healthcare and education is not supported. This necessitated the need to identify additional legal bases to use personal non-unique identifiers and the need to develop a new research database for this linked data.

An application for a new research database was submitted to the Yorkshire & Humber Bradford Leeds Research Ethics Committee and Confidentiality Advisory Group (CAG) for s251 approval for approval under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 to process confidential patient information without consent. The new application referenced a new research database Connected Bradford Health and Education Research Database (HERD) which identified a specific cohort of over 300,000 individuals that reside in the Bradford region who are registered at one of the 86 General Practices and were born from the year of 1991. The Yorkshire & Humber Bradford Leeds Research Ethics Committee favourable approval supported the application on 3 August 2018 and CAG favourable opinion was granted on 3 September 2018 (IRAS ref: 239924, CAG ref: 18/CAG/0091 and REC ref: 18/YH/0200). A further amendment to the HERD Database was submitted to the respective committees to flow personal identifiers to the Department for Education (DfE) to enable healthcare data to be linked with the National Pupil Database (NPD) and the DfE social care data as well as link Geospatial data. Yorkshire & Humber Bradford Leeds Research Ethics Committee approval was received 25 July 2019. The CAG favourable opinion was granted on 16 December 2019 and 30 July 2020 respectively whilst the DfE formally approved the NPD application on 26 May 2020. More recently, to enable the Connected Bradford research database to be updated on an ongoing basis for the existing cohort of 315,693 individuals and include additional children that start school in Bradford at the beginning of the academic year in September, favourable opinion was granted for this amendment by the CAG committee on 16 April 2021. Ethical approval covers the updates to the database.

Data acquisition

Information Governance (IG) is a key challenge in accessing data, sharing data, hosting data, using data and developing a linked dataset. Unfortunately, there is very little publicly available guidance that sets out the information governance framework to establish a dataset that can be used for secondary use analysis and research purposes. Guidance was sought from an independent IG consultant who supported Connected Bradford's commitment to ensure that data from healthcare, local government and other partners is always shared securely and lawfully. With this support, the programme developed a Data Protection Impact Assessment and data sharing agreements that outlined scope, the data linkage pseudonymisation process, mitigations to identified risks and concerns from engagement activities with data providers and the public, legal bases and information security. Further advice was obtained from the Information Commissioner's Office, with subsequent buy-in from the Local Medical Committee and the NHS Bradford District and Craven Clinical Commissioning Group, to secure further reassurances for the regional clinical workforce before agreements and information leaflets were distributed to data providers.

It then took approximately six months to receive all the signed data sharing agreements from all 86 General Practices across Bradford, the three Trusts serving the Bradford district; BTHFT, Bradford District Care Trust and Airedale NHS Foundation Trust and Bradford Metropolitan District Council. Following this, further agreements have been collated from other data providers in Bradford and across the Yorkshire region to expand and scale the Whole Systems Data Accelerator data linkage programme.

Pseudonymisation and linkage

Identifiable information is removed at source by the data providers so that personal information is not available to the study. The NHS number is a unique identifier for UK health records. While errors when inputting the NHS number can lead to gaps in record linkage, duplication is rare. The NHS number is encrypted at source to create a unique but non-identifying linking key. This is obtained using a secure one-way pseudonymisation process to derive an invariant pseudonym from the NHS number to enable data to be linked across multiple organisations at Connected Bradford¹⁴. To ensure that the programme would not have any mechanism by which an individual could be intentionally or unintentionally re-identified, an agreement and process was developed with a trusted organisation to hold the encryption key on behalf of the data providers. The encryption key is only shared with named Caldicott Guardians or designated officers in the data providers that have signed agreements with the programme. De-identified data is then joined to other datasets that used the same process using the non-identifying linking key, which is subsequently removed from production databases.

The data providers extracted and pseudonymised the data at source and using secure methods transferred the pseudonymised data to BTHFT where it is linked using the pseudonymised NHS number. In the case of General Practices, the data processor was Apollo Medical Software Solutions Ltd (Apollo). Apollo extracts the necessary pseudonymised primary care information on an automated basis thus providing a fully managed service. An audit trail enabled general practices to see what data and reports are produced and where they have been transferred. The automated process refreshes the cBradford databases on a monthly basis but has the potential to move to a daily basis thereby realising near to real time activity (Figure 1).

Linkage to non-health sources such as education, employment, benefits data recorded by the local authority, crime and housing, requires matching of non-unique personal identifiers as the unique NHS number is not routinely recorded in non-health datasets. Whilst linkage to aggregate non-health data is supported by matching geographic units, for example using the Lower Super Output Area (LSOA), additional legal bases is required to permit individual data linkage. Education records reference a Unique Pupil Number that identifies each pupil in England and is not available to health care organisations thereby requiring non-unique personal identifiers for data linkage. Confidentiality Advisory Group approval was obtained to permit individual data linkage of cBradford health records to NPD education records and children's social care data held

by the Department for Education using non-unique personal identifiers (CAG ref: 18/CAG/0091 and REC ref: 18/YH/0200).

Future plans for the Connected Bradford Whole System Data Linkage Accelerator include implementation of a process to match individual records to a pseudonymised Unique Property Reference Number (UPRN) to allow individual-level linkage of other non-health datasets such as crime and benefits.

Data collected

As well as a data provider, BTHFT acts as the controller of the data which is held in a secure environment that adheres to the ISO 27001 standard and to the NHS Data Security and Protection Toolkit¹⁵. The database is overseen by the Connected Bradford Research Database Committee (Committee) that provides scientific advice, monitors the secondary use of data for research and service improvement purposes and oversees database development. The Committee includes commissioners, data providers, data governance experts and citizens

The cBradford data linkage model combines primary care data from general practices (including appointment history, prescribing and clinical data), community care data (including mental health, school nurse and health visitor interactions), secondary care data from acute hospitals (including maternity, inpatient, outpatient and emergency services), Yorkshire Ambulance Service 999, Electronic Patient Records, patient transport service and 111 data, palliative care data, adult social care data, children's social care data, children's centres data, education, housing and benefits data from local authorities, crime data from West Yorkshire police, housing data from private housing providers and the National Child Measurement Programme data (Figure 2).

Patient and public involvement

The cBradford communication strategy focused on improving communication with the local population on the use of health data for research and service development. Existing patient and public panels co-produced posters and leaflets and provided insights on the acceptability of secondary use analysis of data for research and service improvement purposes. A new active cBradford patient and public panel was set up to bring together a network of people to share knowledge and information, co-develop research studies, support interpretation of anonymised health data and discuss future collaborations.

Key collaborative relationships

To develop data requirements and supporting processes, cBradford has fostered key collaborations across a diverse multi-disciplinary stakeholder group that have been instrumental in the development of the Connected Bradford Whole System Data Linkage Accelerator. Across Bradford, key partners include the Local Medical Committee, Public Health, NHS Bradford District and Craven CCG, Bradford Care Alliance as well as the data providers; BTHFT, Bradford District Care Trust, Airedale NHS Foundation Trust, 86 General Practices, Bradford Metropolitan District Council, Incommunities Ltd (social housing provider) and West Yorkshire Police. Outside of

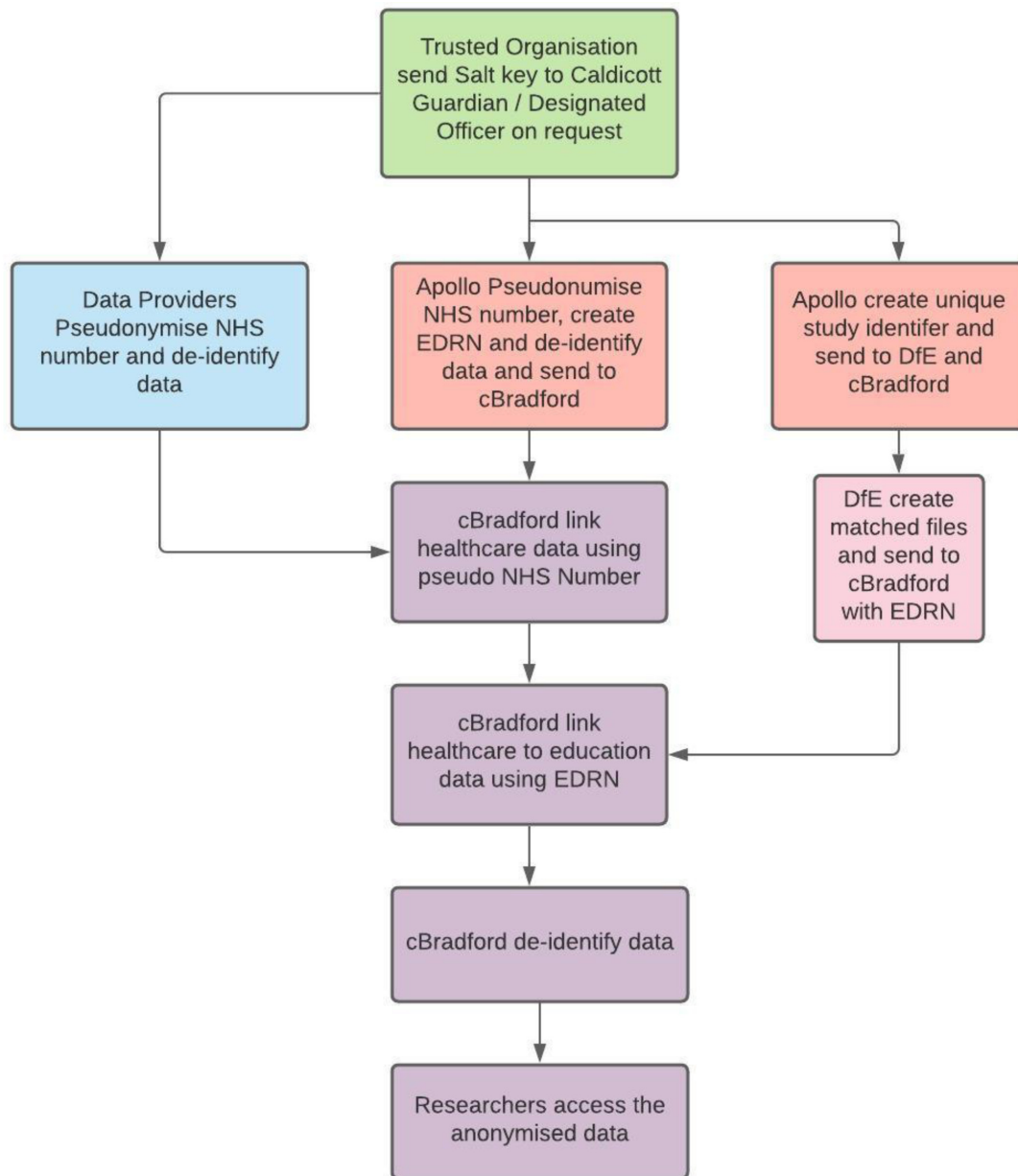


Figure 1. Pseudonymisation process. NHS, National Health Service; EDRN, Education Data Research Number; DfE, Department for Education;

Bradford, we have engaged with Yorkshire Ambulance Service, Calderdale and Huddersfield NHS Foundation Trust, Department for Education who have contributed pseudonymised data towards the Whole System Data Linkage Accelerator as well as Humber NHS Foundation Trust who support the development of the Yorkshire and Humber Care Record¹².

This collaborative enterprise has been instrumental in the resulting signed data sharing agreements and automation of data flows by providing exposure to knowledge, expertise, resource

investment, opportunities and ways of working between different partners and stakeholders.

Dataset content and validity

The scale and content of the data are described in [Table 1](#).

Certain threats to validity arise from the nature of routinely collected healthcare data: variation in code selection, the presence of implausible values, missing data, paper record migration and population stability. Practitioner code selection



Figure 2. Connected Bradford. NHS, National Health Service; CCG, Clinical Commissioning Group.

can be affected by factors such as financial incentives, which distort longitudinal disease prevalence trends, and variation in code selection and coding depth between practitioners. Data entry errors and default values incorrectly reported from non-nullable fields can lead to implausible values such as out of range dates and patient ages. Key missing values such as height, weight, blood pressure and socio-demographics can limit some analyses. Although the longitudinal primary care data spans more than forty years in total, the quality and completeness is substantially lower pre-digitisation due to prior data having been transferred from paper records. As such, most analyses will need to focus on data from the past 10 to 15 years to retain validity. Migration in and out of the city means that the population is not stable, as illustrated by the discrepancy between the 868,000 unique individuals in the GP data, and Bradford population figures produced by the Office for National Statistics in 2017 of 534,300, although the expanding geographic footprint of the data explains some of this¹³.

Data availability

BTHFT and cBradford are committed to making data available for research and service improvement. cBradford is an ethically approved research database (REC ref: 17/EM/0254). Applications for data access are reviewed every two months by the Connected Bradford Research Database Committee. Further details about how to apply for access to cBradford data are available at <https://www.bradfordresearch.nhs.uk/our-research-teams/connected-bradford/>

Discussion

This paper describes the process of linking data from a wider range of health and non-health sources. Connected Bradford has developed from a core of consented Born in Bradford health and education data to cover a whole population system of over 800,000 citizens. There are useful lessons for other districts or cities attempting to establish system-wide data linkage with a number of core themes that we have found to be essential to

Table 1. Data Sources, dates, sample size, key variables.

Data Source	Dates	Sample Size	Key Variables	Linkages	Comments
Primary Care	1950–2022	1,131,462	Appointments, Medication, Clinical history, Reviews data. CTV3 Coded with some SNOMED, IDMultiLexProduct codes.	Pseudonymised data linkage	GP data for 86 practices across Bradford and Airedale.
Secondary Care - BTHFT	2007–2022	1,154,436	Inpatient, Outpatients, Pharmacy, Theatre, Urgent & Emergency Care data, Theatre and Maternity data, ICD-10	Pseudonymised data linkage	Secondary Uses Service, Electronic Patient Records, maternity and other hospital data from BTHFT.
Secondary Care - Airedale	2007–2022	383,893	Inpatient, Outpatients, Urgent & Emergency Care data. Coded as SNOMED, CTV3, ICD-10	Pseudonymised data linkage	Secondary Uses Service data from Airedale Hospital NHS Foundation Trust.
Secondary Care - Calderdale	2007–2022	443,043	Inpatient, Outpatients, Urgent & Emergency Care data. Coded as SNOMED, CTV3, ICD-10	Pseudonymised data linkage	Secondary Uses Service data from Calderdale & Huddersfield NHS Foundation Trust.
Community Care	2016–2020	260,000	Health Visitor, School Nursing, Child and Adolescent Mental Health Services, Child Development, CTV3, ICD-10 Unit data	Pseudonymised data linkage	Community care data from Bradford District Care Trust.
Adult Social Care	2016–2020	47,360	Nursing home, Residential home, Homecare data	Pseudonymised data linkage	Adult Social Care data from BMDC and North Yorkshire County Council.
Children Social Care	2017–2021	315,693	Children in Need, Children Looked After data	Pseudonymised data linkage	Children social care data recorded by the DfE.
DfE education data	1990–2019	315,693	Early Years Foundation Stage Profile data, Phonics, Key Stage 1,2,4 and 5 data, Absences and Exclusions data	Pseudonymised data linkage	Education data recorded by the DfE.
BMDC education data	2015–2019	30,000	Early Years Foundation Stage Profile data, Phonics, Key Stage 1,2 and 4 data, Absences and Exclusions data. Data is coded in list format with descriptor of the event e.g Illness: "2", "3" etc	Pseudonymised data linkage	Education data recorded by BMDC.
Yorkshire Ambulance Service	2019–2022	3,991,879	Urgent and Emergency Care, 999, 111 data . Uncoded by descriptive headers with y/n in data e.g Stroke: "Y"	Pseudonymised data linkage	Urgent and emergency care data recorded by Yorkshire Ambulance Service.
Housing	2020	22,701	Insulation, Type, Rent, Tenancy data, ownership, high level data only	Pseudonymised data linkage	Housing data recorded by BMDC and Incommunities Housing Provider.
Benefits		52,897	Households with income <£15k and >£15k data, numbers in each house (by postcode super output area)	Pseudonymised data linkage	Benefits data recorded by BMDC.
Crime	2017–2022	658,775	Offences, incidents, postcode, lower super output area, date of event. Data is coded in list format with descriptor of the event e.g offence: "Other theft"	Pseudonymised data linkage	Crime data recorded by West Yorkshire Police.
Environmental	2018, 2022, 2027	City of Bradford	Air quality data		Air quality data recorded by BMDC.

GP, General Practitioner; NHS, National Health Service; BMDC, Bradford Metropolitan District Council; DfE, Department for Education.

success (Figure 3). These are aligned with principles such as FAIR¹⁶ and the “five safes,”¹⁷ and with the findings of public perspectives on the use of patient data and data-driven technologies using patient data.

1) Public engagement. Locality-based approaches to routine data linkage have the advantage of being able to recognise the importance of place and communities. Consultation at a local level on the priorities and objectives, participation in data access committees and open communication of how the data is used to improve services are cornerstones in establishing an acceptable, information governance compliant whole system data linkage. A key challenge is how to ensure sustainable and genuine public engagement that is essential to foster trust in how data is used and how this engagement can become a catalyst for public involvement in improving services.

2) Practitioner and policy integration. Data providing organisations must invest time and expertise to setup the data extraction process and help interpret the data; therefore it is important that they see some return on that investment. There is a tendency in public sector organisations to use data purely for performance management rather than for improvement. Engagement with health and local government leaders and practitioners across all sectors is essential in identifying how the linked data can support their priorities and objectives, ensuring that analytical work feeds into quality improvement and learning systems decision and supports the best use of limited resources.

3) Legal and ethical data access frameworks. Information on the Research Database NHS REC and CAG section 251 approvals is described earlier. The Research Database REC approval ensures that data is available for use by academics and researchers as well as providing reassurance for data providers.

4) Information governance. Legal bases, data sharing and processing agreements and the associated transparency and requirements of the Data Protection Act 2018¹⁸, General Data Protection Regulation¹⁹ and Common Law Duty of Confidentiality²⁰ represent a potentially significant obstacle to wider access to health, council and social care datasets. With expertise from IG advisors and the Information Commissioners Office we have developed and implemented model Data Protection Impact Assessments (DPIA) and Data Sharing Agreements (DSA) referencing the legal bases and the data linkage process for system partners. One unintended consequence of the Covid19 pandemic has been the greater recognition and support for data sharing.

5) Technicalities of data linkage. Unique NHS numbers allow deterministic data linkage for NHS records with NHS numbers encrypted at source with the encryption or salt key held by trusted third parties with Caldicott Guardian approval. These methods ensure high quality linkage and reporting²¹. Linkage of health to council datasets that do not have an NHS number requires similar approaches to be developed based on address and other patient identifier matching to safeguard



Figure 3. Framework for developing Whole System Data Linkage.

confidentiality but has the potential for greater challenges with missing and incomplete data sets.

6) Data curation, guardianship and access. Data access for defined purposes is approved by the access committee and enabled using role-based authentication within virtual private networks connections to accredited data safe havens or the trusted research environment. These environments require appropriate analytical tools (e.g. R Studio, STATA) and resource to cover cloud storage. The wide array of databases requires careful curating to promote access and understanding, with clear descriptions, data dictionaries and glossary of codes.

7) Data validity and visualisation. Key threats to data validity include variation in code selection, implausible values, missing data and population mobility and estimation. A wide range of existing open source and commercial tools and techniques exist to enable opportunities for effective data visualisation, a crucial tool for providing actionable responsive insights for local citizens, practitioners, commissioners and policy makers (Figure 4). Ongoing involvement of data-providing organisations is a key requirement rather than just the provision of data, as local intelligence on changes to practice or data

collection are vital in the context of routinely-collected data. Civic data Cooperative²² is one example where public data can be used for service improvement priorities.

Next steps

This paper describes the initial phase of building a whole system linked dataset that we hope will be a useful model for other settings. This is work in motion that will require perseverance and commitment from all the partner organisations and communities if it is to succeed. International technology businesses have gained fabulous wealth from harnessing our private data and the challenge is for public sector organisations to demonstrate how their unique datasets for public good as well as commercial profit. One of the next phases is to demonstrate the utility and added value of how linked data can improve health and wellbeing through improved understanding of needs and better targeting of support. Our goal is to provide impact studies to illustrate this and demonstrate to local communities how their collective data can provide build a learning health system.

We will also explore the potential to add new layers of data from other partners and from environmental data sources. This could include air pollution, green space, built environment

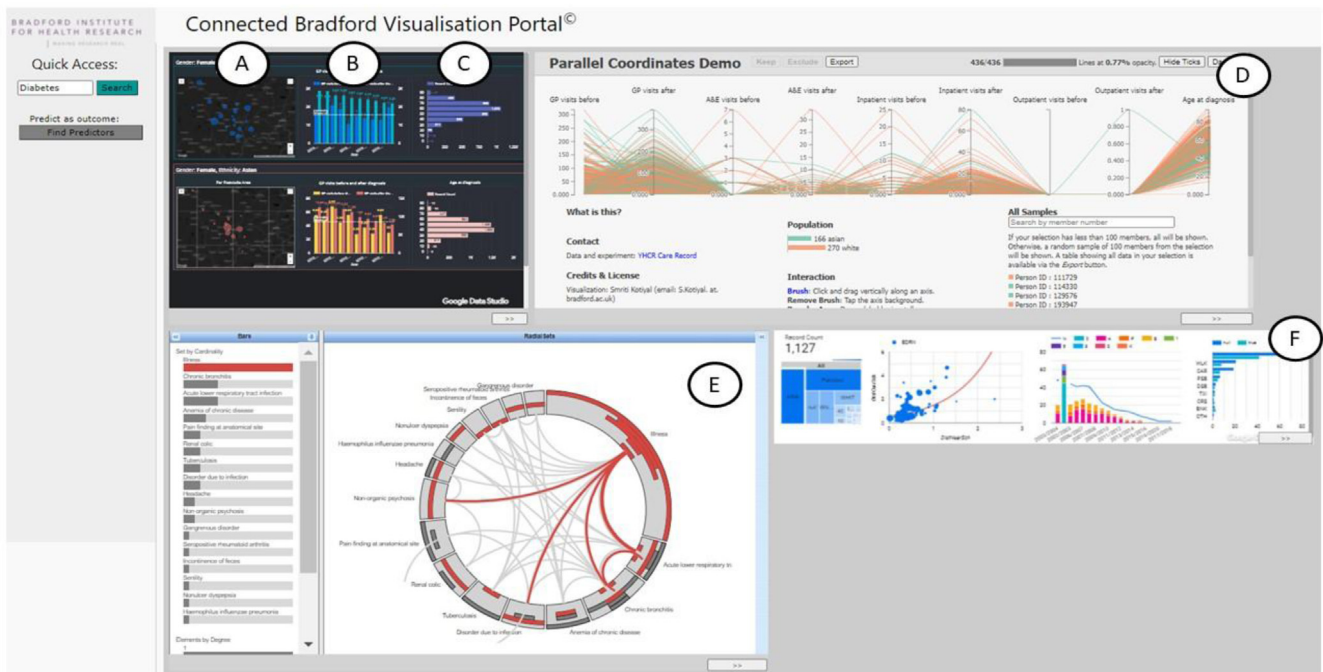


Figure 4. An example workflow connecting Primary Care, Secondary Care and Education data to investigate service demand for women with Type 2 Diabetes. Several visualisation cards have been created in this workflow, starting with a dashboard (top left) showing: (A) the incidence of Type 2 Diabetes for women of White British ethnicity (top row) and Asian / Pakistani ethnicity (bottom row), (B) The bar charts show the number of GP visits before and after being diagnosed broken down by lower-layer super output area (LSOA), and (C) the age distribution at time of diagnosis. (D) A selection of a postcode area (BD4) from the previous card creates a parallel coordinates plot showing the pathways of patients across different services (including GP visits, A&E visits, inpatient and outpatient visits) before and after being diagnosed. (E) Selecting women whose number of GP visits have reduced post diagnosis and who have never been to Accident and Emergency (A&E) before diagnosis creates a new card with a RadialSets visualisation¹ displaying the overlap of reasons why those women visited A&E. (F) A small dashboard is created showing education-related measures for people who suffer from headaches, including ethnic background, commute to school, special education needs and their eligibility for free school meals.

indicators, food outlets and transport to allow us to capture wider influences on health. Pseudonymised UPRNs will be used to link geospatial data to individuals' addresses. Finally we are committed to finding better methods to engage our local communities in the nature and benefits of their own data, such as through data visualization and immersive technologies.

Data availability

This open letter does not include any data. It describes a framework for connecting local data, further details about how to apply for access to these cBradford data are available at <https://www.bradfordresearch.nhs.uk/our-research-teams/connected-bradford/>

References

1. **The National Patient-Centered Clinical Research Network.**
[Reference Source](#)
2. **Danish National Health Registers.**
[Reference Source](#)
3. **Population Health Management Flatpack - A guide to starting Population Health Management Version 1.0 (September 2018).**
[Reference Source](#)
4. **OECD: Health in the 21st Century: Putting Data to Work for Stronger Health Systems.** OECD Health Policy Studies; OECD Publishing, Paris. 2019.
[Publisher Full Text](#)
5. Wickham S, Anwar E, Barr B, *et al.*: **Poverty and child health in the UK: using evidence for action.** *Arch Dis Child.* 2016; **101**(8): 759–766.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Lyons RA, Ford DV, Moore L, *et al.*: **Use of data linkage to measure the population health effect of non-health-care interventions.** *Lancet.* 2014; **383**(9927): 1517–1519.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Friedman CP, Wong AK, Blumenthal D: **Achieving a Nationwide Learning Health System.** *Sci Transl Med.* 2010; **2**(57): 57cm29.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Wright J, Small N, Raynor P, *et al.*: **Cohort Profile: The Born in Bradford multi-ethnic family cohort study.** *Int J Epidemiol.* 2013; **42**(4): 978–991.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Khreis H, Ramani T, de Hoogh K, *et al.*: **Traffic-related air pollution and the local burden of childhood asthma in Bradford, UK.** *Int J Transportation Sci Technol.* 2019; **8**(2): 116–128.
[Publisher Full Text](#)
10. Norris T, Johnson W, Petherick E, *et al.*: **Investigating the relationship between foetal growth and academic attainment: secondary analysis of the Born in Bradford (BiB) cohort.** *Int J Epidemiol.* 2018; **47**(5): 1475–1484.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Wright B, Mon-Williams M, Kelly B, *et al.*: **Investigating the association between early years foundation stage profile scores and subsequent diagnosis of an autism spectrum disorder: a retrospective study of linked healthcare and education data.** *BMJ Paediatr Open.* 2019; **3**(1): e000483.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. **Yorkshire Humber Care Record.**
[Reference Source](#)
13. **Office for National Statistics.**
[Reference Source](#)
14. **Open Pseudonymiser Tool.** University of Nottingham.
[Reference Source](#)
15. **NHS Digital Data Protection and Security Toolkit.**
[Reference Source](#)
16. Wilkinson M, Dumontier M, Aalbersberg I, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. **UK Data Service Regulating Data Access.** accessed online 16th December 2020.
[Reference Source](#)
18. **Data Protection Act 2018.**
[Reference Source](#)
19. **Guide to the UK General Data Protection Regulation.**
[Reference Source](#)
20. **Common Law Duty of Confidentiality.**
[Reference Source](#)
21. Gilbert R, Lafferty R, Hagger-Johnson G, *et al.*: **GUILD: GUIDance for Information about Linking Data sets.** *J Public Health (Oxf).* 2018; **40**(1): 191–198.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. **Civic Data Cooperative.**
[Reference Source](#)

Open Peer Review

Current Peer Review Status:  

Version 2

Reviewer Report 22 November 2022

<https://doi.org/10.21956/wellcomeopenres.20258.r53194>

© 2022 **Bian J.** This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jiang Bian

Department of Health Outcomes & Policy, University of Florida, Gainesville, FL, USA

The authors have revised the manuscript and addressed my concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: informatics; ehr; real-world data; data science; AI

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 16 November 2022

<https://doi.org/10.21956/wellcomeopenres.20258.r53195>

© 2022 **Buchan I.** This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Iain E. Buchan 

Chair in Public Health and Clinical Informatics, Department of Public Health, Policy & Systems, Institute of Population Health, University of Liverpool, Liverpool, UK

The responses are fine.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: public health; health informatics; epidemiology; health services research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 03 May 2022

<https://doi.org/10.21956/wellcomeopenres.19378.r48337>

© 2022 Buchan I. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Iain E. Buchan 

Chair in Public Health and Clinical Informatics, Department of Public Health, Policy & Systems, Institute of Population Health, University of Liverpool, Liverpool, UK

The letter describes a population-wide data-linkage platform for healthcare, social care, and other civic functions such as education to combine their intelligence efforts, including inputs from university-based researchers, and co-learn as a civic system.

An impressive array of data sources is assembled around research (e.g. a birth cohort) and service-development (e.g. educational adjustments) use-cases. The nature of the underlying data construct as a persistent longitudinal multi-agency, multi-purpose civic record and intelligence system could be made more explicit for the international reader unfamiliar with the UK setting. International comparisons might include the PCORnet in the US (<https://pcornet.org/data/>) and the Danish national wide-ranging register(s): https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers.

The data-linkage, information governance and analysis (pooled and federated) arrangements required to initiate the 'flywheel' of a 'learning system' are described in general terms. An international comparison might bring out the key component required to pull through timely, frequent analyses (and up-stream data linkage and curation improvements) from service development actions – namely the coupling of population health management with care/service workflow (e.g. <https://www.cipha.nhs.uk>). The governance required to achieve this must cover analytics sufficiently to manage their quality and trust the results to steer services. The distinction between loosely coupled local policymaking and tightly coupled tactical intelligence might be expanded upon.

The authors usefully describe the non-unique identifier challenges of linking education and health data. For the international reader unfamiliar with the use of NHS Number as a globally unique person identifier in social care records from non-NHS providers, the authors might explain how service-users are identified, and the consequences of not identifying them properly.

The authors usefully draw out the distinction between data uses for performance management vs

service improvement, and the context of practitioner trust. In addition, the foundation of citizens' juries in generating the social contract for these data uses is well made. The authors could move beyond public involvement to "foster trust" to active involvement of public members in setting data processing targets such as service improvement intelligence priorities.

The granularity of analysis-ready data, particularly from secondary care, could be described more for an international audience – detailing the subset of structured, curated hospital episode statistics vs. headroom for more information extraction from clinical narrative etc.

The geographical granularity of data marked as "anonymised" might be discussed more, for example the small area level at which crime data are available, and their utility in public health intelligence for example.

The value of the linked data resource is discussed in terms of "public good" or "profit" as a binary distinction, which oversimplifies civic data contexts. For example, the Civic Data Cooperative (<https://civicdatacooperative.com>) model may employ social impact enterprise vehicles such as a UK Company Limited by Guarantee specific data uses. The UK civic context of being able to involve data subjects (of records from public services) in closing the 'data-action gap' to improve services or enable research through the organised efforts of society is unusual internationally and could be expanded upon.

This is a big topic that is hard to cover in one letter, the authors raise some important issues and make a valuable contribution.

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: public health; health informatics; epidemiology; health services research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Author Response 16 Aug 2022

Jane West, Bradford Hospitals National Health Service Trust, Bradford, UK

We thank the reviewer for their very helpful comments and have responded below:

1. We are grateful for the suggestion and suggestion and have added this to the introduction.
2. We recognise the importance of valid and reliable data to inform practice and policy. This is core to delivery of any prediction model and ensuring caution and interpretation.

We have added this to page 2/3.
3. We thank the reviewer for the suggestion and have added this to page 4.
4. This has been added.
5. We agree and have expanded granularity in Table 1.
6. We agree and have expanded granularity in Table 1. We have a data sharing agreement with West Yorkshire Police to provide Connected Bradford anonymised crime data. We are exploring linking this information at a geographical level with other datasets.
7. We have expanded on the Data Action Gap and an amendment has been made in the 'public engagement' section.

Competing Interests: None

Reviewer Report 07 March 2022

<https://doi.org/10.21956/wellcomeopenres.19378.r48336>

© 2022 **Bian J.** This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jiang Bian

Department of Health Outcomes & Policy, University of Florida, Gainesville, FL, USA

Connected Bradford: a Whole System Data Linkage Accelerator

This paper describes a large scale effort of linking health care data across different health care systems and then with their non-healthcare data including education, employment, and benefits data, for a large population in Bradford, UK.

- It is unclear what and how geospatial data are linked to patient data (if by residential history, then need to consider whether the linkage is meaningful; e.g., especially consider that people move around; exposures to poor neighborhood characteristics are only valid, when the neighborhood was poor when the individual was living there). The temporal component is important for research.
- Based on the description in the Ethical approval section, it's not clear updates on the database (e.g., adding new births) need to be re-approved? (a number of amendments were mentioned for adding additional patients to the database).
- "Unfortunately, there is very little publicly available guidance that sets out the information governance framework to establish a dataset that can be used for secondary use analysis and research purposes." Maybe not for this particular context, but information governance does exist (maybe limited), e.g., UK Biobank and various US-based clinical data research networks such as the PCORnet and OHDSI; the All of Us program at US, has pretty complete and well-thought-out data governance guidance.
- It's not clear what the quality of using the NHS number for linkage; as in the US, the Social Security Number has poor quality (e.g., duplicated assignment exist); and it will be interesting to know (although maybe hard to do) how many data entry errors exist across systems and how will affect the linkage. At minimum, it's interesting to know whether quality checks exist (e.g., double checking names of those who are linked by NHS numbers).
- From the description, it seems that individual-level linkage to non-healthcare data does not yet exist, because other data sources do not carry NHS numbers?

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: informatics; ehr

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Aug 2022

Jane West, Bradford Hospitals National Health Service Trust, Bradford, UK

We thank the reviewer for the very helpful comments and have responded below:

1. We are in the process of working with a geospatial data analyst who will be deriving the Unique Property Reference Number (UPRN). The UPRN is a unique number that is allocated to each land or property. The UPRN is routinely recorded by the housing team in the local authority and large private housing associations in Bradford including Incommunities Ltd. In addition to health information, an individual's address history is available in primary care records. Therefore, we are hoping to have a history of an individual's pseudonymised UPRN information. The pseudonymised UPRN will be made available to the Connected Bradford programme and will facilitate the data linkage with geospatial data and other data types. This has been added to the 'Next Steps' section.
2. We have ethical approval to maintain up to date research databases. Therefore we do not need to seek further approvals to add additional individuals including new-borns to the database. For the following research database, new born babies registered with their general practice or recorded in hospital information will be included. At the moment, the primary care information is received on a monthly basis and the hospital data is received on a daily basis. (IRAS ref:227117 and REC ref:17/EM/0254) The ethics approval is superseded by the following ethical opinion from August 2022: (IRAS ref:314829 and REC ref:22/EM/0127). For the following research database, we have approval from both the NHS and the DfE to maintain an up to date database which includes all new school entrants. This means that we expect to have the education and health data for all new school entrants children that reach four years old (IRAS ref: 239924, CAG ref: 18/CAG/0091 and REC ref: 18/YH/0200).
3. When we began the development of this dataset in 2016, we identified a number of potential options to develop this database in the UK including an application to NHS Digital for the linked data but this process was potentially lengthy, limited to specific users for specific purposes, costly; this was therefore not going to be sustainable. We also looked for publicly available information on how other organisations or bodies had developed their research databases, but other than information from consented studies such as Born in Bradford, we struggled to get this information. In addition, we noted that the time to develop a large scale consented study was not going to be feasible within the timeframe. We noted that commissioning organisations that had access to linked databases were again limited to whom and for what purpose they could share this data and also had liaised with NHS Digital for this. As a result, we worked with an independent information governance consultant in 2016 who helped navigate the steps, issues, identify mitigations to risks involved, technical processes

and other processes involved that arose in the development of the Connected Bradford. These steps have been simplified in the form of the Whole System Data Accelerator framework.

There has been a drive in the UK over the last few years to record information for individuals using their NHS number. The NHS number is a 10 digit number and there is a degree of validation involved when this information is entered into clinical systems. This has been added to page 4.

4. The Connected Bradford Health and Education Research Database links healthcare data to education data on an individual level. The education data is a non-healthcare dataset. (IRAS ref: 239924, CAG ref: 18/CAG/0091 and REC ref: 18/YH/0200). An individual's education data does not record NHS numbers and therefore we had to explore an additional technical process which involved additional approvals from the NHS and the Department for Education to support this data linkage. This is described in the third paragraph of the section 'Pseudonymisation and linkage'

The above responses can also be found in a point-by-point table format here: https://wellcomeopenresearch.s3.eu-west-1.amazonaws.com/linked/269958.Response_to_Reviewer_1_Comments_-_Connected_Bradford_0v1_2022-08.docx

Competing Interests: None