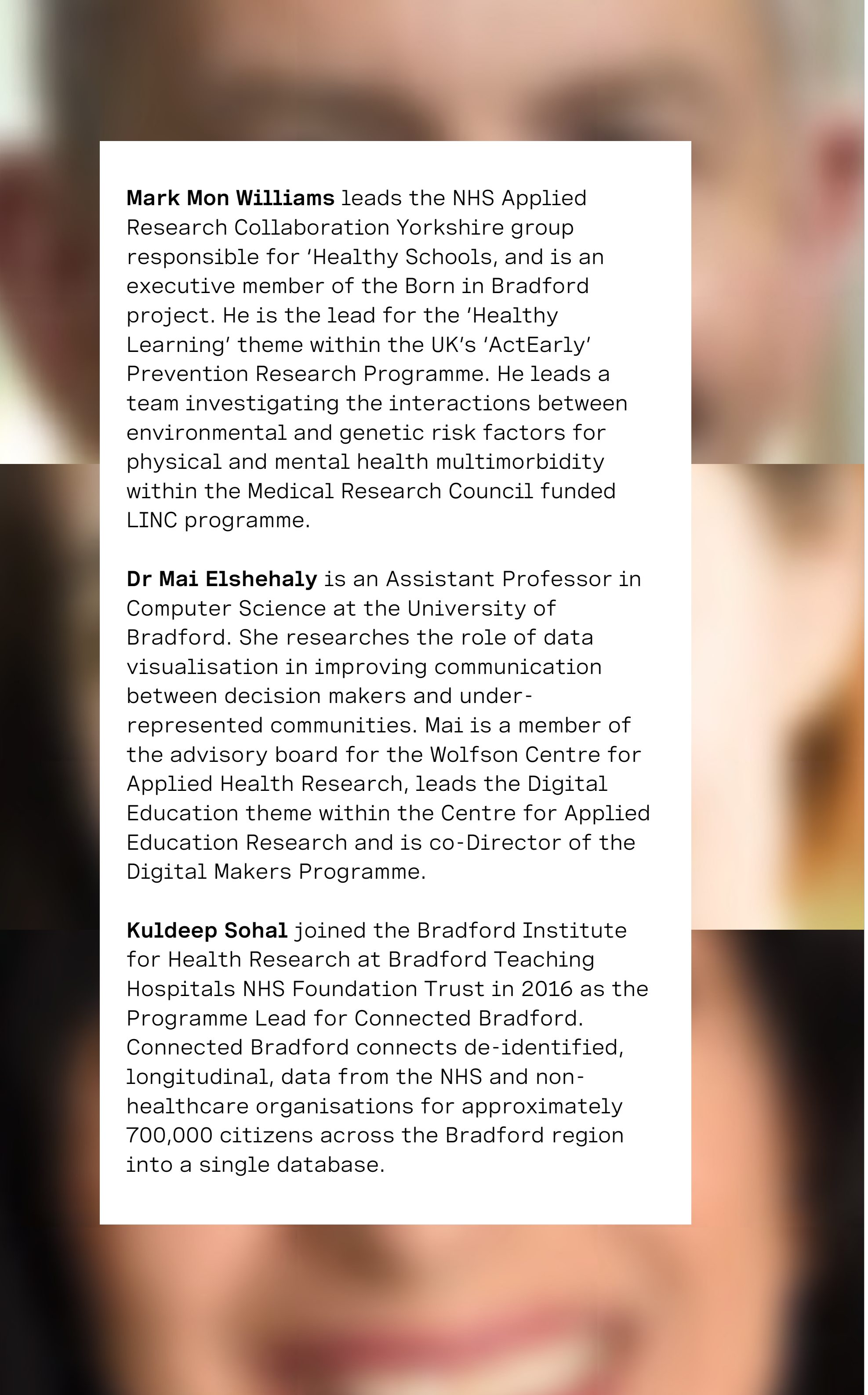

Mark Mon Williams leads the NHS Applied Research Collaboration Yorkshire group responsible for 'Healthy Schools, and is an executive member of the Born in Bradford project. He is the lead for the 'Healthy Learning' theme within the UK's 'ActEarly' Prevention Research Programme. He leads a team investigating the interactions between environmental and genetic risk factors for physical and mental health multimorbidity within the Medical Research Council funded LINC programme.

Dr Mai Elshehaly is an Assistant Professor in Computer Science at the University of Bradford. She researches the role of data visualisation in improving communication between decision makers and under-represented communities. Mai is a member of the advisory board for the Wolfson Centre for Applied Health Research, leads the Digital Education theme within the Centre for Applied Education Research and is co-Director of the Digital Makers Programme.

Kuldeep Sohal joined the Bradford Institute for Health Research at Bradford Teaching Hospitals NHS Foundation Trust in 2016 as the Programme Lead for Connected Bradford. Connected Bradford connects de-identified, longitudinal, data from the NHS and non-healthcare organisations for approximately 700,000 citizens across the Bradford region into a single database.

Connected data for connected services that reflect the complexities of childhood

Mark Mon-Williams, Mai Elshehaly and Kuldeep Sohal, University of Leeds

The COVID-19 pandemic has revealed myriad problems with the systems, services and processes that are supposed to protect and support children and young people. We argue that a fundamental problem relates to fragmented public services, including education, health and social care. Data science provides analytical systems that can allow us to address this connection problem and facilitate a holistic approach to meeting the needs of children and young people. In this context, the sharing of data across services is an essential step in creating efficient systems capable of safeguarding children effectively, providing timely support for vulnerable children, removing structural inequalities and enabling a whole-system approach to tackling the wider determinants of physical and mental health, educational outcomes and social mobility.

Here, we present three case studies drawn from the connected routine datasets of over 13,500 children within the Born in Bradford (BiB) project to illustrate the benefits of data sharing. These provide the rationale behind the creation of the West Yorkshire Integrated Data Engine for Analytics (IDEA) centre - dedicated to connecting routine datasets and applying the power of science to develop systems that are fit for purpose.

Context

In December 2021, *The child of the north* report (Pickett et al., 2021) provided a harrowing account of the appalling situation facing huge swathes of children growing up in poverty within the North of England (with the same problems affecting disadvantaged children regardless of where they live). The resulting cost to families is enormous, but so is the financial public burden that arises from our collective failure to support vulnerable children. It is extremely difficult to believe that anyone could digest *The child of the north* report and not be motivated to join the voices demanding radical change to our systems and services. We suggest that connected datasets can allow us to start understanding the root causes of the problems (by revealing the intersections and interactions between different factors such as health, education and social care), and can enable the genuine multiagency responses that are required if the complex needs of disadvantaged children are to be addressed effectively.

The consequences of failing to use information effectively in public service delivery are catastrophic. The Children's Act 2004 requires *a serious case review* after the death of a child where abuse or neglect is known or suspected. It is rare to read a serious case review that does not conclude that failure to share information across professionals and organisations was a contributing factor to the death. In December 2021, the tragic cases of Arthur Labinjo-Hughes and Star Hobson (Child Safeguarding Practice Review Panel, 2022) provided a depressing reminder that our systems and processes are antiquated and do not take advantage of scientific advances in data science. Arthur and Star are the personification of a wide malaise affecting a multitude of children within the UK's most deprived areas. Our central argument is that we could tackle these problems by connecting the information available through the routine datasets held by the different organisations that have a share of responsibility for the wellbeing of children and young people (e.g., health, education, social care, policing etc.).

We start by defining information as the state of the Universe, relative to an observer, measured as the logarithm of the number of its possible states. In order to avoid a tedious scientific exposition of this definition, we consider information in the context of Wordle, a popular game that illustrates the key points we wish to make. Wordle requires players to guess five letters chosen from the 26 letters of the English alphabet. Wordle selects different letters each day and gives the player six opportunities to guess the five letters. In this description of the game, the player is trying to guess which one of 7,893,600 permutations is the target. Fortunately, Wordle (as the name implies) constrains the task by only using words within the English language.

It is clearly helpful for a Wordle player to use their existing knowledge about the English language when attempting to guess the target word, which brings us to another key feature of Wordle – information is provided after each guess is made. A correct letter in the right location is highlighted in green, whereas a correct letter in the wrong location is shown in yellow. It follows that a good Wordle strategy is to use an initial word that will yield maximum information. This means using the most frequent letters and placing those where they most frequently appear. Thus, there are two possible approaches to playing Wordle. One is to use all available information and enjoy a daily average Wordle performance of between three and four guesses. The alternative is to ignore the available information. Our argument is that the current approach to public service delivery is equivalent to this alternative – and frankly bonkers – Wordle strategy.

The child of the north report establishes what 'bad' looks like when data are not used effectively. Importantly, rich datasets exist, but the information is split across education, health, social care, and so on. The fact that these datasets remain disconnected is deeply concerning as they can describe the intersecting and interacting factors impacting on a child's life. Moreover, the combined data can support the creation of powerful data analytic systems for tackling childhood vulnerabilities. Nevertheless, no organisation within the UK currently uses connected data to address inequality or even to meet its legal responsibilities to protect children and young people. On the basis of first principles, it can be argued that failure to connect information across public services means that we are missing opportunities to better serve children. We do not need to rely on logical arguments alone, however, as we have accumulated a wealth of empirical data within the Bradford district that show the benefits of connecting data.

Connected routine datasets

Bradford is uniquely positioned to show the power of data as routine administrative records have already been connected through the Born in Bradford (BiB) project. BiB is one of the world's largest longitudinal birth cohort studies and has linked routine data for over 30,000 Bradfordians.[†] Frequent engagement with the families and children allows us to collect informed consent for continued routine data linkage (e.g. health, social care and education records).

The success of BiB in using connected data has led to the creation of the 'Connected Bradford' database containing the records of citizens across the Bradford District. Connected Bradford combines a number of records, including primary care (e.g., appointment history, prescribing and clinical data), community care (e.g., mental health, school nurse, health visitor interactions), secondary care (e.g., maternity, outpatient), social care, children's centres data, education, housing and benefits, crime, housing data and data from the National Child Measurement Programme (see Sohal et al., 2022).[‡]

In order to connect the health and education data, we obtained Confidentiality Advisory Group approval for individual data linkage of health records to National Pupil Data education records held by the Department for Education. We used non-unique personal identifiers because the Unique Pupil Number (that identifies each pupil in England) is not available to healthcare organisations.

[†] The BiB cohort comprises 12,453 women recruited at 28 weeks of pregnancy, who gave birth at the Bradford Royal Infirmary to 13,857 children between the period 2007 and 2011 (see https://borninbradford.nhs.uk). Half of all BiB families live within wards classed among the 20% most deprived within England and Wales and 45% of families are of Pakistani origin. The BiB families provided informed consent for their routine electronic records to be linked and used for scientific purposes. These are supplemented with detailed testing (on measures including physical health, cognitive ability, sensorimotor function, household demographics etc.) on a regular basis, providing one of the richest available descriptions of a population's genotype and phenotype.

[‡] For the interested reader, Sohal et al. (2022) provide detailed information on the different datasets, the legal pathways through which linkage occurred, data security and storage, ethical arrangements etc.

We will now use three case studies to demonstrate the usefulness of the resulting connected data (i.e., our examples will focus on linked health and education records).

Glasses in classes

Our first study demonstrates how the existence of a connected dataset enables a proper understanding of the complex factors that contribute towards poor outcomes for children (including poor physical health, mental health, educational attainment and social mobility) – in this instance, poor reading skills.

There are a large number of children in Bradford who fail to learn to read at an acceptable rate (DfE, 2017). The natural response to this situation is to improve school leadership around reading or to provide approaches such as phonics programmes. The BiB data revealed, however, that a fundamental health problem might explain the unsatisfactory levels of reading. The connected data showed that many children identified with an ophthalmic deficit (i.e., they needed a pair of eyeglasses) were not taken to the hospital eye service or the local optometrist despite a letter informing the relevant carer that there was a problem with the child's eyesight. Moreover, the data showed that children with uncorrected eyesight were at increased risk of delayed reading skills (DfE, 2021). These insights were available because the ophthalmic status of the children could be obtained from the health records (i.e., the children's medical records) while the child's reading abilities were available through the connected education data (i.e., information from the Department for Education). This simple example demonstrates the power of connected datasets in flagging important intersections between education and health, showing where we need to address health barriers that impact on education.

Classroom Air Cleaning Technologies (Class-ACT) research

Our second case study relates to the usefulness of connected data when testing interventions targeted at improving outcomes for children. The COVID-19 pandemic highlighted the importance of good ventilation in the prevention of airborne diseases. The pandemic also revealed that a number of classrooms don't have adequate ventilation, which could potentially increase the risk of a child or teacher contracting an airborne illness. One possible solution to poorly ventilated classrooms is the provision of 'air cleaning technologies'. For example, filtration technologies remove particles from the circulating air including the COVID-19 virus and other pathogens. The technologies also remove particulate matter that can cause asthma and the pollen that can cause hay fever. It follows that – *in principle* – these technologies might decrease illness in children (and teaching staff) and thereby reduce school absences (with all of the educational benefits accrued through increased time in school). However, the use of these technologies involves substantial financial investment, and there is currently little evidence available that the potential benefits will actually translate into real world impact.

The fact that Bradford has linked health and education data allowed the creation of the Class-ACT (Classroom Air Cleaning Technologies) project, where we could conduct a randomised trial and use a combination of health and education data to understand fully the impact on children attending schools fitted with these technologies. Class-ACT allows a holistic investigation into the data on childhood infections available from the health records combined with information on school absences available from the education system. In the absence of the connected datasets, it would only be possible to obtain a piecemeal picture of the potential for COVID-19 transmission to be reduced through fitting air-cleaning technologies within schools.

Identifying children with undiagnosed autism through education records

Our third case study highlights the usefulness of connected data in addressing the problem of undiagnosed autism. There is overwhelming evidence to show that identifying autism in the early years of life has great benefits for the child and their family (e.g., French & Kennedy, 2018). Unfortunately, many children do not have their needs identified until the end of primary school, or sometimes not even until they are in secondary school or beyond (Department of Health and Social Care, 2021). The issue of undiagnosed autism places health and education services under great strain, and creates long-term financial costs that could have been avoided through early action. Many areas have lengthy waiting lists for autism assessment, with children often waiting for many years before they receive the support they need.[†] Furthermore, societal inequalities are reflected within the autism assessment process, with children from disadvantaged backgrounds waiting much longer than their more affluent peers (Kelly et al., 2019). Notably, children from disadvantaged backgrounds with undiagnosed autism are far more likely to also have additional needs that will require a holistic response from a number of different organisations (Pickett et al., 2021).

The BiB dataset showed that routine educational data can be used to identify undiagnosed autism in children, and tested novel approaches to address the problems associated with this and other developmental disorders. We were able to show that the Early Years Foundation Stage Profile (EYFSP) scores given by teachers at the end of Reception Year (age 4–5) can be used to help identify neurodevelopmental problems, including autism (Wright et al., 2019). Once again, these insights were only made possible because the children's health records (identifying patients with autism) were linked with their education records (allowing us to explore what 'red flags' in education data might be indicative of children at risk of undiagnosed autism).

The data-driven research then led to a study in 10 Bradford primary schools, involving in-school screening of 600 pupils to identify 'at risk' pupils faster and more accurately (Wright et al., 2021). The study identified children who would benefit from a formal autism assessment. A multiagency team, including Child and Adolescent Mental Health Services (CAMHS) and educational psychology services, then attended the relevant schools to help conduct assessments quickly, share information instantly with teachers, parents and caregivers, and facilitate the development of a single support plan.

[†] For example, NICE Quality Standards for autism stipulate that the wait between referral and first diagnosis appointment should be no more than 13 weeks, but in a government survey just 18% of local authorities in England reported meeting this target (All-Party Parliamentary Group on Autism, 2019).

Building Integrated Data Engine for Analytics (IDEA) centres
These three case studies demonstrate how connected data can allow holistic evidence-based solutions to be implemented when tackling issues related to childhood vulnerability. This is equivalent to using the feedback provided by Wordle when guessing the target word rather than ignoring the information. It is worth emphasising that the current systems for identifying and supporting children with vulnerabilities (e.g., autism) are ignoring the information available across the system because each stakeholder is failing to share data with its partner organisations – despite their shared statutory responsibility for children and young people.

The creation of a connected dataset is essential, but the optimal use of information requires community engagement, intelligent analysis and visualisation of data. This is why we created the West Yorkshire Integrated Data Engine for Analytics (IDEA) centre that brings together experts in data analytics, community engagement, ethics, law, economics and visualisation across Leeds, York and Bradford. Our West Yorkshire IDEA centre is now leading a regional effort to tackle inequality through the Digitally Acting Together As One (DATA 1) programme. This follows the COVID-19 pandemic lifting the lid on the costs of the current fragmented system. For example, the pandemic highlighted the large number of vulnerable children 'under the radar' of organisations with safeguarding responsibilities. The unavailability of connected information during lockdown made coordinating multiagency responses extremely difficult, despite the same families requiring support from multiple organisations. These problems played out against the backdrop of rising inequalities, with service providers finding it increasingly hard to deliver the holistic support needed to address the root causes of many needs (Pickett et al., 2021).

DATA 1 aims to create data analytics tools that can improve service delivery across different providers, including health, education and social care. These tools will: (a) allow early identification of need and (b) enable frontline practitioners to organise efficient and effective multiagency responses to children who would benefit from support. The creation of data analytics tools capable of connecting practitioners will transform public service delivery and improve the support offered to hundreds of thousands of people. Moreover, they will connect policymakers, communities and practitioners, and empower them to tackle the numerous problems that currently plague our society.

The creation of data analytics tools requires us to find solutions to the following challenges: technical issues associated with connecting and visualising data; ethical and legal issues of data protection; engagement with the communities served by the tools; and the imperative of producing tools that can be readily used by practitioners from a range of different organisations.

The technical issues in developing data analytics tools requires us to tackle several challenges, including: integrating heterogeneous data sources; understanding the decision-making tasks and priorities of practitioners from a range of different organisations; and measuring the impact that these decisions can have from the perspective of the communities. This dictates a high level of continued engagement to ensure that the designed technologies benefit and respect these diverse stakeholders and their lived experiences.

Instead of trying to resolve these challenges across a range of different domains, our strategy is to focus on creating data analytics systems that can help clear the queue of children on waiting lists for autism assessment, allow earlier identification of undiagnosed autism and enable children with autism to receive multiagency support as soon as their needs are recognised. Our rationale is that focusing on one specific problem will ensure that we can work through the technical, ethical, legal and practical issues in a manageable manner. It also enables us to coproduce our data analytics solutions with the relevant communities, and allows us to communicate why we are connecting data.

In our opinion, the processes of coproduction and community engagement require the same level of attention as the technical aspects of data science. While we are aware that many people are worried about the misuse of education data information (e.g., Kolkman, 2022), we share their concerns and are passionate about using our connected datasets to help children understand their data rights. Coproduction and engagement lie at the heart of BiB, and we believe this explains why our wonderful families are content for their data to be linked and used in the manner described here. We have teams who lead work around ethics and legal pathways, and we have developed specific programmes of work that directly support coproduction and engagement. For example, our Digital Makers programme involves the 30,000 young people involved in the next phase of BiB (known as 'Age of Wonder'). Digital Makers is working with the young people's schools to provide digital upskilling and help young people understand their data rights. We capture their voice through a 'Youth Summit' that directly feeds into our research endeavours.

Scaling up

In the peer review process, we were asked how the Bradford approach could be replicated elsewhere. The answer is that any region within the UK can *choose* to commit to using data science to tackle the dreadful inequalities affecting our most disadvantaged communities. The question is whether there is the political will to tackle this source of inequality. West et al. (2022) have identified four key aspects needed to develop and sustain such approaches: leadership, resource and capacity, culture, and partnerships. Effective partnerships (between the police, local authorities, health systems, schools, universities etc.) are founded on strong, shared principles, which shape decisions and interactions through planning and delivery.

The connection of services through linked data requires an unprecedented breadth of collaboration across organisations, commitment to community engagement (see Islam et al., 2022) and an openness to change both culture and practice. We would urge every area across the UK to commit to the necessary partnership working and explore - at pace - how their routine datasets can be connected to improve outcomes for families. Our experience is that successful implementation needs to be driven at a regional level (allowing community engagement), with coordination and support provided through central government.

Conclusion

We have deliberately focused on examples of connecting education with health data to make the case for the use of education data in improving outcomes for children. In an alternative approach, we could have shown the benefits of linking education data with other datasets. We could also have shown the power of using routine educational data *per se* in learning how we can better support children. For example, we used the educational records from 8,130 participants in the BiB study to explore the predictive utility of the EYFSP. We found that the school readiness measure ('good level of development') predicted performance in reading, writing, maths and science at the end of Key Stage 1 (age 6-7) and later special educational needs (SEN) status (Atkinson et al., 2022). This means that the EYFSP could be used as a screening tool to identify children at risk of poor academic achievement and/or requiring SEN support. Thus, the EYFSP has the potential to be an effective trigger for early identification (and support) of SEN, and provides an exemplar for the possible use of improved provision through the use of education data.

We hope this essay shows why it is important to use education data effectively in efforts to tackle inequality and provide support to our most vulnerable families. The work in Bradford has demonstrated unequivocally the immense potential of data analytics to generate societal benefit. The connected data have also revealed the huge need for transformation across all of our organisations, systems and processes. Our firm conclusion is that connecting education data with other routine datasets is an essential first step towards reducing inequality and improving life outcomes for children and young people.

Acknowledgements

Born in Bradford (BiB) is only possible because of the enthusiasm and commitment of the children, parents and caregivers in BiB. We are grateful to all the participants, practitioners and researchers who make BiB happen. The work was conducted within infrastructure provided by the Centre for Applied Education Research (funded by the Department for Education through the Bradford Opportunity Area) and ActEarly, a City Collaboratory approach to early promotion of good health and wellbeing funded by the Medical Research

Council (grant reference MR/S037527). Mark Mon-Williams was supported by a fellowship from the Alan Turing Institute. The Centre for Applied Education Research receives core infrastructure funding from the Wellcome Trust (WT101597MA) and the National Institute for Health Research Yorkshire and Humber ARC (reference: NIHR20016). Linkages with education data are further supported by joint grants from the UK Medical Research Council and UK Economic and Social Science Research Council (MR/N024397/1) and the Strategic Priority Fund 'Tackling multimorbidity at scale' programme (MR/W014416/1) delivered by the Medical Research Council and the National Institute for Health and Care Research in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council.

All-Party Parliamentary Group on Autism (2019). *The Autism Act 10 years on: A report from the All-Party Parliamentary Group on Autism on understanding, services and support for autistic people and their families in England*

Atkinson, A. L., Hill, L. J. B., Pettinger, K. J., Wright, J., Hart, A. R., Dickerson, J., & Mon-Williams, M. (2022). Can holistic school readiness evaluations predict academic achievement and special educational needs status? Evidence from the Early Years Foundation Stage Profile. *Learning and Instruction, 77*, 101537

Child Safeguarding Practice Review Panel (2022). *Child protection in England: National review into the murders of Arthur Labinjo-Hughes and Star Hobson*

Department of Health and Social Care (2021). *National strategy for autistic children, young people and adults: 2021 to 2026*

DfE (Department for Education) (2017). *Bradford opportunity area delivery plan*

DfE (2021). Opportunity areas insight guide: Health and education

French, L., & Kennedy, E. M. M. (2018). Annual research review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: A systematic review. *Journal of Child Psychology and Psychiatry, 59*, 444-456

Islam, S., Albert, A., Haklay, M., & McEachan, R. (2022). *Co-production in ActEarly: nothing about us without us*. Bradford Institute for Health Research & University College London

Kelly, B., Williams, S., Collins, S., Mushtaq, F., Mon-Williams, M., Wright, B., Mason, D., & Wright, J. (2019). The association between socioeconomic status and autism diagnosis in the United Kingdom for children aged 5-8 years of age: Findings from the Born in Bradford cohort. *Autism, 23*, 131-140

Kolkman, D. (2022). The (in)credibility of algorithmic models to non-experts. *Information, Communication & Society, 25*(1), 93-109

Pickett, K., Taylor-Robinson, D., et al. (2021). *The child of the north: Building a fairer future after COVID-19*. NHSA (Northern Health Science Alliance) and N8 Research Partnership

Sohal, K., Mason, D., Birkinshaw, J., West, J., McEachan, R. R. C., Elshehaly, M., Cooper, D., Shore, R., McCooe, M., Lawton, T., Mon-Williams, M., Sheldon, T., Bates, C., Wood, M., & Wright, J. (2022). Connected Bradford: A whole system data linkage accelerator. *Wellcome Open Research, 7*(26)

West, J., Wright, J., Bridges, S., Cartwright, C., Ciesla, K., Pickett, K. E., Shore, R., Witcherley, P., Flinders, M., McEachan, R. R. C., Mon-Williams, M., Bird, P., Lennon, L., Cooper, D., Muckle, S., England, K., & Sheldon, T. (2022). Developing a model for health determinants research within local government: Lessons from a large, urban local authority. *Wellcome Open Research, 6*, 276

Wright, B., Mon-Williams, M., Kelly, B., Williams, S., Sims, D., Mushtaq, F., Sohal, K., Blackwell, J. E., & Wright, J. (2019). 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 Paediatrics Open, 3*, e000483

Wright, B., Konstantopoulou, K., Sohal, K., Kelly, B., Morgan, G., Hulin, C., Mansoor, S., & Mon-Williams, M. (2021). Systematic approach to school-based assessments for autism spectrum disorders to reduce inequalities: A feasibility study in 10 primary schools. *BMJ Open, 11*, e041960