Minutes of the meeting of the Confidentiality Advisory Group

18 November 2021

Held via Zoom

Present:

Name	
Dr Tony Calland MBE	CAG Chair
Dr Martin Andrew	CAG member
Dr Malcolm Booth	CAG member
Ms Sophie Brannan	CAG member
Dr Patrick Coyle	CAG vice-chair
Mr David Evans	CAG member
Professor Lorna Fraser	CAG member
Dr Katie Harron	CAG member (absent from discussion of 21/CAG/0167)
Dr Pauline Lyseight-Jones	CAG member
Ms Diana Robbins	CAG member
Mr Dan Roulstone	CAG member

Ms Clare Sanderson	CAG alternative vice-chair

Also in attendance:

Name	Position (or reason for attending)
Ms Katy Cassidy	Confidentiality Advisor
Ms Caroline Watchurst	Confidentiality Advisor
Mr Paul Mills	Senior Confidentiality Advisor/Service Manager
Ms Natasha Dunkley	Head of Confidentiality Advice Service

1. Introduction, apologies and declarations of interest

2. Support decisions

Secretary of State for Health & Social Care Decisions

The Department of Health & Social Care senior civil servant on behalf of the Secretary of State for Health & Social Care agreed with the advice provided by the CAG in relation to the **16 September 2021 and 30 September 2021** meeting applications.

Health Research Authority (HRA) Decisions

The Health Research Authority agreed with the advice provided by the CAG in relation to the **16 September 2021 and 30 September 2021** meeting applications.

3. New applications - research

a. 21/CAG/0167 - ECHILD - Education and Child Health Insights from Linked Data

Context

Purpose of application

This application from the Department for Education set out the purpose of creating a research database, ECHILD, in which data will be collected to form a comprehensive view of the journey through childhood to adulthood.

The Department for Education, NHS Digital and University College London have created ECHILD database in order to collect data on how health and education services complement or compensate for each other. The research database will be used to understand what works to improve the design and delivery of policies and systems which better meet the needs of children and young people. The ECHILD dataset will fill this gap in evidence by facilitating research that will inform policy-makers and service commissioners about the associations between education risk factors and health outcomes. The database will be created by linkage of data from the National Pupil Database (NPD) and HES data. The ECHILD database is already in existence, and this application is to update the ECHILD database, and also link with further datasets held by NHS Digital – Community Service Data and Mental health data (MHMDS/MHLDDS/MHSDS) for children in ECHILD, and Maternity Services Data and HES records for mothers of ECHILD participants. This extended ECHILD is Phase 2, and this also requires 's251 support'.

The Department for Education (DfE) will disclose identifiable data from the NPD alongside a pseudonymised study specific ID, to NHS Digital. NHS Digital will then link this list from the NPD to the Personal Demographics Service (PDS) to obtain up to date identifiers for NHS patients. The list from the NPD will then be linked to HES data, and additional datasets. NHS Digital will then supply ONS SRS with the pseudonymised HES extract. Appropriate data minimisation will be applied, such as amending full dates of birth and death to month/year of birth and death. Separately, DfE will provide ONS SRS with de-identified attribute data, via an anonymised Pupil Matching Reference (aPMR). ONS SRS will then combine this NPD extract with the NHS Digital linked dataset to create the ECHILD research database, which will contain de-identified data only. The ECHILD database will be held within ONS SRS and DfE will be provided with a copy.

A recommendation for class 1, 2, 4, 5 and 6 support was requested to cover access to the relevant unconsented activities as described in the application.

Confidential patient information requested

The following sets out a summary of the specified cohort, listed data sources and key identifiers. Where applicable, full datasets and data flows are provided in the application

form and relevant supporting documentation as this letter represents only a summary of the full detail.

Cohort	All children and young people in England: a. Born after 01 September 1984, b. And have a record in HES from 01 April 1997 onwards or a record in NPD from 01 September 2001 onwards
Data sources	1. National Pupil Database – held by Department for Education Phase 1 – ECHILD; 2. HES data (Admitted Patient Care, Outpatient Data, A&E data, adult critical care, mortality data) held by NHS Digital. Phase 2 – extended ECHILD; 3. Community Service Data and Mental health data (MHMDS/MHLDDS/MHSDS) and Maternity Services Data and HES records for mothers of ECHILD participants, held by NHS Digital
Identifiers required for linkage purposes	1. Name 2. NHS number 3. Date of birth 4. Full postcode 5. Sex For linkage within ONS-SRS; 1. Anonymised Pupil Matching Reference (aPMR) 2. HES ID
Identifiers required for analysis purposes	1. Medium Super Output Area 2. Index of Multiple Deprivation 3. Gender 4. Ethnicity 5. Month and year of birth 6. Month and year of death

Additional information	Applicants are intending that disclosures will happen annually, and feed into an ongoing data resource. Applicants expect that the first disclosure will be in early 2022.

Confidentiality Advisory Group advice

The following sets out the Confidentiality Advisory Group advice which formed the basis of the decision by the Health Research Authority.

Public interest

The CAG noted that this activity fell within the definition of medical research and was therefore assured that the application described an appropriate medical purpose within the remit of the section 251 of the NHS Act 2006. The CAG agreed that the application was in the public interest. However, the medical purpose behind the data collection was less clear as the information collected would be used for education, rather than medical purposes.

Scope

The CAG noted that the scope of the support sought was unclear.

The applicant had stated that NHS Digital do not currently have a legal basis for holding identifiable data from DfE in order to establish the research database, and are therefore submitting this CAG application in order to establish a legal basis for NHS Digital to act as a data processer on behalf of the DfE. The CAG agreed that s251 support would not be needed for the transfer of information from the National Pupil Database at the DfE to NHS Digital, as this is not confidential patient information, although a legal basis would be required for such a transfer. Support was also not required for the linkage of the datasets held by NHS Digital, as NHS Digital have an existing legal basis to process these datasets.

Support is also sought for DfE to become data controller for patient identifiers in the linked dataset, however it did not appear that any items of confidential patient information would flow from NHS Digital to DfE.

The CAG asked the applicant to provide further details on why support had been applied for given that data from DfE is outside of the remit of s251and the data it is to be linked with is already held at NHS Digital and therefore there was no breach in confidentiality.

Practicable alternatives

Members considered whether a practicable alternative to the disclosure of confidential patient information without consent existed in accordance with Section 251 (4) of the NHS Act 2006, taking into account the cost and technology available.

- Feasibility of consent

The applicants anticipate that 20 million records will be included. Seeking consent for this number of people would not be feasible. The applicants also noted that obtaining the names and addresses of people would require further disclosures of identifiable data.

The applicants also noted the potential risk of introducing bias, should consent be sought. Those who do not provide consent are also more likely to be those with particular health conditions (such as rare diseases), and these individuals are particularly important to capture in the ECHILD database. The CAG accepted that consent was not feasible.

- Use of anonymised/pseudonymised data

NHS Digital require access to confidential patient information in order to link the NPD data to datasets held by NHS Digital. As noted above, the CAG was unclear whether support was needed, as the only processing of confidential patient information would be undertaken by NHS Digital, who have an existing legal basis to process the patient information they hold.

'Patient Notification' and mechanism for managing dissent

It is part of the CAG responsibility to support public confidence and transparency in the appropriate sharing and use of confidential patient information. Access to patient information without consent is a privilege and it is a general principle of support for reasonable measures to be taken to inform the relevant population of the activity and to provide a right to object and mechanism to respect that objection, where appropriate. This is known as 'patient notification'. This is separate to the local obligation to comply with the principles of the General Data Protection Regulation and Data Protection Act 2018.

The applicants explained that they were taking a layered approach in providing information. A website for the study, hosted by UCL, had been created. Summary information had been developed with lay and journalist support and would be made available on the website. Email, twitter and a link to an online webform for queries is included on the right-hand side of the webpage. The webpage included assurances that individuals can't be identified and information on how the data is kept safely.

Information will also be available on websites for NHS Digital and DfE. DfE are in the process of setting up a Data Improvement Across Government (DIAG) webpage, in which the details of this project, and others within the programme, can be found. This will include information about how to opt out. The DIAG website will provide information about the study, including a list of Q&As for members of the public and for the press. DfE plan to establish a dedicated mailbox for queries and concerns, which will open for a 6-8 week period.

DfE will provide suggested wording for privacy notices to schools and local authorities. This will explain that data will be used by DfE for research purposes and may be shared with researchers and other organisations connected with promoting the education or wellbeing of children in England.

Individuals will be able to opt out of their inclusion in ECHILD. NHS Digital have advised that they currently do not have a mechanism for providing study-specific opt-out for Research Databases. However, to avoid relying entirely on the National Data Opt-Out, NHS Digital are exploring whether a mechanism can be put in place to allow opt out for ECHILD, separately from the National Data Opt-Out. If patients register an Opt-Out, NHS Digital will flag the patient record which will block subsequent extracts of records for that individual, i.e. at annual refreshes, data for that individual will be removed from the database.

The applicants advise that the website contains information on how to opt-out. The privacy notice for ECHILD advises patients to register with the National Data Opt-Out at NHS Digital. The applicant is working on developing a study specific opt out.

Members noted that the database would contain details from pupils, their parents/carers and wider family members. Approximately 20 million pupils would be involved, plus several million sets of data for their mothers, and the data will include sensitive data such as mental health information. The CAG agreed that the online only approach described in the application would be insufficient. Members suggested that the applicants required a detailed communications plan which should include other ways of promoting the activity, such as via social media and the press. Poster and leaflets should also be used. The CAG suggested that the applicants work with relevant NGO's and parent groups to promote the study.

The CAG agreed that the information provided on the website needed to be clearer. It needed to be explained that data would be collected for pupils up until the age of 35 and that this would include mental health related data.

A study specific opt-out mechanism needed to be created, so that patients can opt-out of further updates being added to the database once they leave school. Details of this mechanism provided if a future resubmission is made. The CAG noted that the size of the database presented a risk that public confidence in confidentiality could be undermined, potentially leading to many people registering dissent with the National

Data Opt-Out. Members emphasised that this was why a study specific opt-out was needed.

Patient and Public Involvement and Engagement

Meaningful engagement with patients, service users and the public is considered to be an important factor for the CAG in terms of contributing to public interest considerations as to whether the unconsented activity should go ahead.

University College London have conducted engagement work when preparing the ECHILD database. Participants in this engagement work were asked about the acceptability of processing confidential patient information without consent for the types of research that would use ECHILD data. Initially participants reported feeling discomfort with usage of their data without consent. However, they agreed that this would be accepted if data were used by the NHS or in research to improve health and education. The applicants explained that researchers would not be able to access identifiable data and participants strongly supported linking the data.

As part of the 2021 In2Science Programme, which provides A-level students from low-income and disadvantaged backgrounds with opportunities to gain practical insights into the STEM sector through placement opportunities in universities and industry), the ECHILD project team introduced students to the concepts of administrative data, deidentification and record linkage, using the ECHILD Database as an example. 14 students, whose data will likely be included in the ECHILD database, as they were born after 1984, were asked to complete an anonymous online poll. As part of the whole-group discussion of the poll results, students who were unsure about their data being re-used for research purposes volunteered the information that they supported the ECHILD project research plans but felt uncomfortable with the concept of administrative data in general, as they had been unaware that it was being collected by schools and hospitals. They wanted data owners to give more information about the fact that administrative data is being collected as part of their day-to-day activities, as well as how it is then re-used safely for research.

The CAG agreed that the patient and public involvement activity carried out so far was good, however there were areas that could be explored further. This included discussion on how the study could be promoted and the information that needed to be included, such as clear explanations that mental health related data would be processed and that information about pupils' parents and wider family may be included.

The CAG noted that no lay representation appears to be included on the Project Advisory Team and asked that the applicants consider including lay people on this team.

Exit strategy

The ECHILD database will contain anonymised data only.

The CAG noted that the final linked database would be disclosed from NHS Digital to DfE. Although this would be de-identified before leaving NHS Digital, members queried whether DfE may be able to re-identify pupils by cross-referencing the study database with existing data held by DfE. The CAG asked the applicant to clarify the items of information which would be included in the final database, for example, would the name of the schools the pupils attend be deleted and to provide justification for a copy of such a large database being required at DfE when they could access the same via ONS.

Confidentiality Advisory Group advice conclusion

In line with the considerations above, the CAG agreed that, on the basis of the information provided, they did not have sufficient information to provide a recommendation under the Regulations.

Following advice from the CAG, the Health Research Authority recommended that the application was deferred.

Further information required

1. Provide clarification on why support had been applied for. This needs to include:

    a. Clarification on the exact data flows and processing of confidential patient information which require support under s251.

    b. A written explanation from NHS Digital, advising why support under s251 is needed for them to process the data supplied by DfE, noting that the data from DfE is not confidential patient information and is therefore outside of the remit of s251.

2. Provide justification for DfE's retention of a copy of such a large database, when they could access the same via ONS.

3. Clarify the items of information which would be included in the final database, for example, would the name of the schools the pupils attend be deleted.

4. Revisions are needed to the information on the website so that the scale of the study is clearly explained, including that data will be collected for pupils up until the age of 35, that this would include mental health related data and that information about pupils' parents and wider family may be included

5. Development of a communication strategy including ways of promoting the study, such as a press campaign and use of social media, need to be explored and details of this notification provided if a future resubmission is made.

6. A study specific opt-out mechanism needs to be created at both DfE and NHS Digital and details of this mechanism provided if a future resubmission is made.

7. Further patient and public involvement needs to be conducted. This needs to include discussion on how the study could be promoted and the information that needed to be included, such as clear explanations that mental health related data would be processed and that information about pupils' parents and wider family may be included.

8. Lay representation needs to be included on the Project Advisory Team, or justification provided as to why this is not possible.

b. 21/CAG/0164 - Long title: Recovery, Renewal and Reset of Services to Disabled Children

Context

Purpose of application

This application from Newcastle University set out the purpose of medical research that seeks to establish which reconfiguration of services, practices and strategies for disabled children made during the coronavirus pandemic worked well.

Eight percent of children living in the UK are disabled. Half of these children have neuro disability conditions, affecting their brain and central nervous system. Many have complex physical and mental health needs, which are generally met through multiple services. In response to the coronavirus pandemic, disabled children who are at increased risk due to poor respiratory function were advised to shield. The duty to deliver care as specified in children's individual Education, Health and Care Plans (EHCPs) was relaxed and most community services for children were de-prioritised as efforts centred on those most at risk from the virus. Services were stopped and re-organised, with some resuming via video link and others face-to-face, but the practice varied and has continued to flex and change in response to policy changes. Many parents reported their children's mental health as deteriorating, and parents' own isolation and distress. Impacts on physical health are currently unclear, however virtual consultations may not suit physical examination, diagnosis or interventions requiring touch or instrumentation. There is emerging evidence of delays to diagnosis and treatment of children, increases in abuse, and in regional and socio-economic inequity

in the impacts of coronavirus. The applicants seek to conduct research into changes to service provision and their consequences for children's physical mental health, as well as the impact on their families' wellbeing. The research will be used to inform practical policy solutions for integrated service recovery.

This study is comprised of six work packages. The applicants are seeking support under Regulation 5 for Work Package 2, as this involves analysis of routinely collected data and consent will not be sought for access to this confidential patient information. Children aged 0-19 years with a confirmed or suspected diagnosis of a neuro disability condition, and living in one of five local authority areas in England, will be included in the study. Areas have been selected so that findings will be generalisable and representative UK-wide and internationally relevant. The NHS Trusts in the five areas will provide the pseudonymised NHS number of each child who meets the inclusion criteria to the North of England Commissioning Support Unit (NECS). NHS medical, community and learning disability services, as well as non-medical services, in participating areas will transfer data on contacts with services to NECS. This data will be individual level data and individuals will be identified by their pseudonymised NHS number and diagnostic category. NHS Digital will also disclose pseudonymised HES and MHSDS data relating to identified children's use of services to NECS. NECS will create unique identifiers for all participating children and will apply a second pseudonymisation code to the datasets. NECS will then send datasets in which children are identified by second pseudonymisation code to the research team. Each process will involve data being shared once, e.g. one list of identified children will be shared once with NECS by a lead NHS Trust; NECS will provide NHS Digital pseudonymised data on children from an area once to the research team.

The applicants are seeking support as, while the data disclosed between individual NHS organisations, NECs and NHS Digital will be pseudonymised, NHS Digital will retain the pseudonymisation key and will be required to identify patients before sharing the relevant datasets.

A recommendation for class 4 and 6 support was requested to cover access to the relevant unconsented activities as described in the application.

Confidential patient information requested

Cohort	Children aged between 0 and 19 years of age, diagnosed with a neuro disability (disability arising from maldevelopment or damage to the brain in early development)
Data sources	1. HES and MHSDS datasets at NHS Digital

Identifiers required for linkage purposes	1. Pseudonymised NHS Number NHS Digital will hold the re-identification key and will be able to re-identify patients to conduct the linkage. NHS Digital will link datasets using; 1. NHS Number 2. Date of birth 3. Postcode
Identifiers required for analysis purposes	1. Date of birth (this is modified to month and year of birth for analysis) 2. Gender 3. Ethnicity

Confidentiality Advisory Group advice

The following sets out the Confidentiality Advisory Group advice which formed the basis of the decision by the Health Research Authority.

Public interest

The CAG noted that this activity fell within the definition of medical research and was therefore assured that the application described an appropriate medical purpose within the remit of the section 251 of the NHS Act 2006. The CAG agreed that the application had a medical purpose and was in the public interest.

Practicable alternatives

Members considered whether a practicable alternative to the disclosure of confidential patient information without consent existed in accordance with Section 251 (4) of the NHS Act 2006, taking into account the cost and technology available.

- Minimising flows of identifiable information

The CAG noted that no identifiers had been selected in answer to Q37 on the CAG form and that only the pseudonymised NHS number will be required for linkage across datasets. When the Confidentiality Advice Team queried this prior to the meeting, the applicants explained that participating sites only provide patients NHS number to NHS Digital, but that NHS Digital require date of birth and postcode in addition to NHS number to ensure the correct patient data is selected from the HES and MHSDS data. The CAG asked the applicant to clarify where NHS Digital would obtain the dates of birth and postcodes from, i.e. would this be obtained from records already held by NHS Digital. Clarification on why NHS Digital needed these identifiers also needed to be provided.

- Feasibility of consent

The applicants advised that participants in Patient and Public Involvement carried out determined that use of routine data without consent was appropriate, as researchers would be able to access data without burdening families.

The applicants also noted that families with complex needs and disabilities would been adversely affected by COVID because of the intensive, daily demands of looking after young people with reduced state support. Requiring families to complete an additional task may lead to those families not consenting, potentially biasing the results. The CAG agreed that consent was not feasible.

- Use of anonymised/pseudonymised data

NECS and the researchers at Newcastle University will only have access to pseudonymised information. The CAG raised no queries in this area.

'Patient Notification' and mechanism for managing dissent

It is part of the CAG responsibility to support public confidence and transparency in the appropriate sharing and use of confidential patient information. Access to patient information without consent is a privilege and it is a general principle of support for reasonable measures to be taken to inform the relevant population of the activity and to provide a right to object and mechanism to respect that objection, where appropriate. This is known as 'patient notification'. This is separate to the local obligation to comply

with the principles of the General Data Protection Regulation and Data Protection Act 2018.

Notices will be displayed as posters in waiting rooms and in newsletters for parents and families of children with neuro disability. These notices will include information on how to opt-out of Work Package 2. The notices included contact details for the Data Protection Office (DPO) at each participating trust, should participants wish to opt-out. If participants contact the DPO to dissent, the DPO will contact NECS to request the removal of the patient from the data. NECS will forward the pseudonymised code of the dissenting participant to the research team, if the data have been forwarded to the researchers. The research team will inform the Sponsor DPO if the data have been analysed and can no longer be removed. The applicants advised that NHS Digital would apply the National Data Opt-Out.

The CAG noted the information provided and asked that the notices to be displayed in waiting rooms and newsletters were revised to include further details on how data would be processed.

Patient and Public Involvement and Engagement

Meaningful engagement with patients, service users and the public is considered to be an important factor for the CAG in terms of contributing to public interest considerations as to whether the unconsented activity should go ahead.

The applicants have consulted with chairs of local Parent Carer Networks in the National Network of Parent Carer Forums, which is a voluntary organisation of parent carers who have come together to support the development of statutory services for children with special educational needs and disabilities. One of the co-investigators is a parent of a disabled young adult and is also a Special Educational Needs and Disabilities (SEND) officer for a local authority. This co-investigator will lead the involvement of parent carers in the Parent Carer Advisory Group. This group will advise on the methods and procedures of the research, the analysis of the findings and their dissemination. Also, young people will be recruited to the projects Young People's Advisory Group. This Group will advise on how best to engage with disabled young people.

When designing the project, the applicants consulted with the Steering Group of the Eastern Region Parent Carer Forum, whose members comprise chairs of eight local forums, who each work with a local steering group across one local authority. The chair of the Eastern Region tabled the research for discussion and provided a verbal summary of the study design. Members of the Eastern Region Forum then cascaded information to their local steering group. In total, up 80 parent-carers across eight local

authorities considered the study design. These parent-carers are of both sexes, are diverse in age (30’s - 60’s), are both working and not working, and are from ethnic backgrounds reflecting the population they serve. Some of the group have additional needs/disabilities. The children of the parent carers consulted range in age from preschool to early twenties and have a wide range of needs. They are educated in a range of settings: mainstream, day and residential special schools, and at home.

The use of routine data without consent was judged by the parent-carers consulted to be appropriate because it would give the study access to data without burdening families. It was also felt that requesting consent would lead to a biased sample. Families of children with complex needs and disabilities have been adversely affected by COVID because of the intensive, daily demands of looking after young people with reduced state support. Requiring families to complete an additional task when they had already given consent to the use of their medical information at source by not opting out was felt to be unnecessarily burdensome.

The CAG commended the applicants for the patient and public work conducted, noting that relevant groups had been involved early on and meaningful engagement conducted.

Exit strategy

The dataset provided to the researchers for analysis will be pseudonymised. The pseudonymisation key will be held by NHS Digital, so the dataset accessed by the researchers is effectively anonymised.

NECS and NHS Digital will delete the confidential patient information provided to facilitate linkage 36 months after the project completion.

Confidentiality Advisory Group advice conclusion

The CAG agreed that the minimum criteria under the Regulations appeared to have been met, and therefore advised recommending support to the Health Research Authority, subject to compliance with the specific and standard conditions of support as set out below.

Specific conditions of support

1. A response to the two first two conditions below need to be provided within one month of the issuing of this letter.

2. Please clarify where NHS Digital will obtain patients dates of birth and postcodes from. Clarification also needs to be provided on why NHS Digital require these identifiers.

3. Please revise the notices to be displayed in waiting rooms and newsletters to include further details on how data would be processed.

4. Favourable opinion from a Research Ethics Committee. **Confirmed 23 November 2021.**

5. Confirmation provided from the IG Delivery Team at NHS Digital to the CAG that the relevant Data Security and Protection Toolkit (DSPT) submission(s) has achieved the 'Standards Met' threshold. See section below titled 'security assurance requirements' for further information. **Confirmed:**

   The NHS Digital 2020/21 DSPT review for North of England Commissioning Support Unit (NECS) was confirmed as 'Standards Met' on the NHS Digital DSPT Tracker (checked 23 November 2021).

4. COPI Notice Transition Application – research

a. 21/CAG/0172 - THIN - Symptoms of COVID-19 in primary care: a cohort study using the THIN database

Context

Purpose of application

This application from the University College London Institute of Health Informatics sets out the purpose of medical research that seeks to describe the long-term profile of symptoms and complications as recorded in general practice for patients with a history of COVID-19 infection.

Long-term symptoms of Covid-19 are an increasingly recognised consequence of Covid-19 infection. Research in this area is needed to provide appropriate ongoing care for these patients. Many patients will present to their GP with ongoing symptoms, but little information is available on what these symptoms are, and the consequent burden on primary care of treating long-term symptoms and complications of Covid-19. Primary care data is often useful in understanding population risk, morbidity and

mortality due to Covid-19. However, symptoms are not typically recorded in a structured way and may only be present in the free text clinical notes. The applicants therefore seek to utilise natural language processing to analyse free text at scale. As well as investigating the long-term symptoms of Covid-19, the applicants also aim to investigate early symptoms, and why some patients have ongoing symptoms, while others recover quickly.

Structured and free text data from contributing practices on the patient cohort (cases and controls) will be transferred from practices to THIN. Upon receipt at THIN free text data will be automatically deidentified before being run through a natural language processing algorithm to transform the free text data into coded data. Coded data only will be transferred to UCL for analysis. Some free text sample will also be transferred to UCL to validate and train the algorithm, but prior to this each sample will be manually checked to ensure they are deidentified.

A recommendation for class 1, 2, 5 and 6 support was requested to cover access to the relevant unconsented activities as described in the application.

Confidential patient information requested

The following sets out a summary of the specified cohort, listed data sources and key identifiers. Where applicable, full datasets and data flows are provided in the application form and relevant supporting documentation as this letter represents only a summary of the full detail.

Cohort	Covid-19 cohort - Patients aged 18 years and over who are registered with a GP practice that contributes to THIN, with a coded diagnosis of Covid-19 or a diagnosis of a viral or respiratory infection and a mention of Covid-19 in the text of the clinical notes Control cohort - Patients aged 18 years and over who are registered with a GP practice that contributes to THIN and who do not have a history of Covid-19. This equates to approximately 95,000 patients in England and Wales.

Data sources	4. Electronic medical records held at GP practices providing data to THIN
Identifiers required for linkage purposes	6. Free text data is transferred to THIN that may include identifying information about the patient.
Identifiers required for analysis purposes	7. Date of death 8. Gender 9. Ethnicity
Additional information	Free text data will be automatically deidentified using natural language processing before in turn being converted into SNOMED structured data using NLP. The deidentification procedure is 98% effective so there is a risk that some free text data will contain identifiers after the automatic deidentification procedure. Free text records were collected from December 2019 onwards.

Confidentiality Advisory Group advice

The following sets out the Confidentiality Advisory Group advice which formed the basis of the decision by the Health Research Authority to transition the study to support under Regulation 5.

Public interest

The CAG noted that this activity fell within the definition of medical research and was therefore assured that the application described an appropriate medical purpose within the remit of the section 251 of the NHS Act 2006.

Previous applications from THIN to transfer free text data of patients within THIN-contributing practices as part of the wider research database have been rejected after advice by CAG. Whilst the Group were aware of these applications and the reasons for rejection, they were also mindful that this application is for a specific ongoing research project into Long-term implications of COVID-19.

As such the CAG felt that the application does have merits and were supportive of the overall intentions of the research. However, there also was strong uncertainty on a number of areas of the application (detailed below) that the CAG could not justify that

the public interest of undertaking the research outweighed the risks of disclosure of highly sensitive information from the free text records. This was the primary reason for deferral.

The specific points of uncertainty are detailed in the sections below and the applicants are asked to consider these points. The CAG are content to review a revised submission, where full information on the below points can be provided.

Free Text Deidentification Process

Whilst the applicants state that experience to data from a manual check of free text samples indicates "*the free text deidentification procedure removes 98% of identifiers from the text*", the CAG remains unsure on what this means in practice. Does this mean that 98% patients have 100% of identifiers removed, 98% of free text records of an x% of patients have 100% of identifiers removed, 98% of identifiers are removed from x% of free text records of x% patients, 98% of identifiable data items across all records collected, or something else? Absolute clarity on how 98% is calculated and what this means in terms of identifiability of each patient is requested.

Also, whilst noting that 98% has been validated with existing data, the CAG were unsure whether this percentage varies depending on each identifier i.e. are there certain identifiable fields that, from experience, are harder to automatically deidentify than others. A commentary on this aspect would be valuable to understanding what potential identifiers may remain after the automatic deidentification process.

Given that the symptoms of long-COVID are unclear the applicants are collecting all free text records. Whilst the CAG are relatively content on the clinical reasons for collecting this data, significant concerns remain about personal sensitive data being collected which would not have any bearing on this research (for example information on domestic violence, extramarital affairs etc.). The CAG requests detail on experience around deidentification from the project so far and approaches that the deidentification procedure takes to limit the collection of this data and its identifiability, and clarification (with examples) of what remains of socially sensitive text after deidentification. This of particular concern to CAG, given that the 2019 UCL patient workshop highlighted patient concerns of a "*Fear of discrimination if free text data on sensitive conditions was more widely available*".

The Group noted that the effectiveness of the automated deidentification procedure has been a continuing concern across all THIN free text applications. Whilst outside the remit of CAG, it is suggested that THIN consider publishing a peer-reviewed paper on the automated deidentification procedures used by THIN and its effectiveness to add to the evidence.

Coding of free text data

Following deidentification it is understood that the free text is run though a natural language processing algorithm to convert it into coded data. The Group were unsure whether the algorithm is limited to identify and convert clinical data only into codes, or whether it would code information irrelevant to the research as well (for example domestic violence). Where irrelevant information is coded, the Group would like to understand what subsequently happens to this data, for example whether it is deleted or kept (with justification if it is kept).

The group is also aware that the standard structured data collection by THIN (outside the scope of this support) does not collect certain sensitive coded data. Reassurance that the algorithm does not code/collect this data from free text is sought.

Progress of research to date, its demonstrable value, and future plans

Given the research started approximately one year ago the CAG would like to understand more about the value that the free text records have provided to date, and why this is the only route to gain the data necessary to answer the research question. The CAG requests a summary of the work and outputs to date, with particular emphasis on the impact that using free text records has had in increasing the understanding of Long-COVID.

It was also noted that, were support to be given, further data will be extracted following expiry of the COPI notice. The CAG requested justification why further data collection is necessary to answer the research question, and why data collected until expiry of the COPI notice is not sufficient to answer the research question.

The Group also requested confirmation on:

- How many further extractions would be undertaken and the latest date that extraction would be completed were any support was to be given
- Whether further extraction will comprise only of new free text entries for existing patients, or whether new patients would be included if meeting the inclusion criteria. If new patients will be included will this be in addition to the already stated number of patients of 141,912, of which 47,421 are from Wales, and 48,348 from England.
- How many samples of free text will be manually viewed and whether this is continuing for the duration of the research. It is noted that the protocol indicates approximately 300 samples will be assessed per data item, totalling a few thousand. Details on how many have been viewed to date and plans to continue this in the future are requested.

Scope

The CAG noted that the application is currently relying on an alternative legal basis to process confidential patient information without consent, under the 'COPI notice' and that this will continue for its duration. The group therefore considered the elements of the project that are expected to be continuing following expiry of the 'COPI notice', and which require support under regulation 5.

It was noted that the applicants also intend to collect and store the full date of death, without minimising this date. The CAG is aware that there can be different views on whether date or death is or is not an identifier, depending on the context. In light of the associated clinical data, an explanation to be provided as to why in totality this dataset is not considered identifiable and thus outside the scope of this support.

Practicable alternatives

Members considered whether a practicable alternative to the disclosure of confidential patient information without consent existed in accordance with Section 251 (4) of the NHS Act 2006, taking into account the cost and technology available.

The CAG felt that in general consent is not practicable, though also noted the 2018 workshop on the use of free text indicated that patients preferred an opt in rather than opt out method to use free text records (further information in the patient and public involvement section).

Concerns around minimising the use of identifiable free text information are discussed above.

'Patient Notification' and mechanism for managing dissent

It is part of the CAG responsibility to support public confidence and transparency in the appropriate sharing and use of confidential patient information. Access to patient information without consent is a privilege and it is a general principle of support for reasonable measures to be taken to inform the relevant population of the activity and to provide a right to object and mechanism to respect that objection, where appropriate. This is known as 'patient notification'. This is separate to the local obligation to comply with the principles of the General Data Protection Regulation and Data Protection Act 2018.

The applicants provided two patient notification posters to be displayed in contributing practices. One was a generic THIN poster to cover all its activities, the second was a poster specific for the activities in this application. The Group considered both communications and noted the patient involvement undertaken on these.

It was noted that the generic THIN poster refers to the collection of free text, but this does not mention that the free text data may be identifiable when transferred. Other aspects of the poster indicate the anonymous use of data and as such patients may assume that no identifiable information is transferred. CAG requested that the generic THIN poster be updated to include reference that the free text information may be identifiable when transferred.

The research specific poster did not contain any reference to the use of free text information, nor what happens to the free text (i.e. potentially identifiable data is transferred and then processed into coded data). CAG requested that explicit clarity is added to the study specific poster around the use of free text and the identifiable nature of this use.

The group noted that using the existing local THIN mechanism will opt patients out of THIN in its entirety and suggested that the applicants consider using a study specific opt out mechanism to avoid unintentional consequences for the wider THIN database.

Patient and Public Involvement

Meaningful engagement with patients, service users and the public is considered to be an important factor for the CAG in terms of contributing to public interest considerations as to whether the unconsented activity should go ahead.

The group noted the patient and public involvement that has been undertaken since 2018 to support the use of free text, as well as more recent project specific involvement which focussed on the patient notification materials.

The group reviewed these reports and were concerned whether participants had a full understanding of the full implications of what they were discussing. The CAG were unclear whether full examples of what free text samples may look like. For example, and linked to the earlier sections, whether participants were aware these free text entries may include sensitive identifiable social entries (such as domestic violence, extramarital affairs) along with the clinical information.

As such, CAG requested further patient and public involvement into the acceptability in using free text information, where participants are provided full examples of the types of free text entries that may be shared before

deidentification (including identifiable socially sensitive information), as well as understanding that the deidentification process is not 100% effective. Undertaking this is of particular importance to provide further evidence of the public views in sharing of this type of information in an identifiable form for the purposes of the research, and help support the public interest of undertaking the research.

Confidentiality Advisory Group advice conclusion

In line with the considerations above, the CAG agreed that, on the basis of the information provided, they did not have sufficient information to provide a recommendation under the Regulations.

Following advice from the CAG, the Health Research Authority recommended that the application was deferred.

Further information required

1. Provide clarity, from project experience thus far, on what the phrase “the free text deidentification procedure removes 98% of identifiers from the text” means in practice, in relation to how many patients have identifiable information stored within the free text after deidentification.

2. Provide information on whether the effectiveness of the deidentification process varies by identifiable data item, including experience and refinements from the project so far, and whether any items are of greater risk at remaining following deidentification.

3. Give commentary on approaches that the deidentification procedure takes to limit the collection of this data and its identifiability, and clarification (with examples) of what remains of socially sensitive text after deidentification.

4. Clarify whether the NLP algorithm only converts clinical data into coded form, or whether social information irrelevant to the research (e.g. domestic violence) will also be converted to coded form.
    a. Where irrelevant data is also coded, clarify what subsequently happens to this data, with justification for this approach.

5. Provide reassurance that the algorithm does not code data from free text that THIN would not ordinarily standardly collect through structured data.

6. Submit a summary of the work and outputs to date, with particular emphasis on the impact that using free text records has had in increasing the understanding of Long-COVID.

7. Justify why further data collection after expiry of the COPI notice is necessary to answer the research question, and why data collected until expiry of the COPI notice is not sufficient to answer the research question.

8. Detail how many further extractions would be undertaken and the latest date that extraction would be completed were any support was to be given

9. Clarify whether further extraction will comprise only of new free text entries for existing patients, or whether new patients would be included if meeting the inclusion criteria.

10. Detail how many samples of free text will be manually viewed and whether this is continuing for the duration of the research.

11. Provide an explanation as to why in totality this dataset is not considered identifiable, given it includes full date of death, and thus outside the scope of this support.

12. Update the generic THIN poster to include reference that the free text information may be identifiable when transferred.

13. Add explicit clarity to the study specific poster around the use of free text and the identifiable nature of this use.

14. Consider using a study specific opt out mechanism to avoid unintentional consequences for the wider THIN database by relying on the local THIN opt out mechanism. Note this is a suggestion for consideration only.

15. Undertake further patient and public involvement into the acceptability of using free text information, where participants are provided full examples of the types of free text entries that may be shared with THIN before deidentification (including identifiable socially sensitive information), as well as understanding that the THIN deidentification process is not 100% effective

16. Favourable opinion from REC **Received 06 October 2020**

17. Continual achievement of ‘Standards Met’ in relation to the relevant DSPT submission (or any future security assurance changes) for the duration of support. Evidence to be provided (through NHS Digital confirmation they have reviewed and confirmed the DSPT submission as standards met’ for the duration of support, and at time of each annual review. **The applicant must ensure that NHS Digital confirmation of ‘standards met’ for The Health Improvement Network (THIN) is in place once support under Regulation 5 is active.** See below for further details.
    a. It is noted that NHS Digital have assured the 20/21 DSPT submission for The Health Improvement Network (THIN).

5. Minutes of the meeting held on 16 September 2021 and 30 September 2021

The minutes of the meeting held on 16 September 2021 and 30 September 2021 were not reviewed as an outcome is pending.

6. Any other business

No other business was raised.

The Chair thanked Members for their attendance and the meeting was closed.

Signed – Chair	Date
Signed – Confidentiality Advice Team	Date