Study protocol

**UKILD – Long COVID**

**The UK Interstitial Lung Disease Long-COVID19 study (UKILD-Long COVID): understanding the burden of Interstitial Lung Disease in Long COVID.**

V1.3 11/11/2021

MAIN SPONSOR: Imperial College London

FUNDERS: Research Grant, UKRI Grant Ref: MR/W006111/1

STUDY COORDINATION CENTRE: NHLI

IRAS Project ID: 297891

REC reference:

**Protocol authorised by:**

| **Name & Role** | **Date** | **Signature** |
|---|---|---|

**Study Management Group**
Chief Investigator: Professor Gisli Jenkins
NIHR Research Professor and Margaret Turner Warwick Chair of Thoracic Medicine
Head of Margaret Turner Warwick Centre for Fibrosing Lung Disease,
National Heart and Lung Institute,
Imperial College London,

Guy Scadding Building,
Cale Street,
London, SW3 6LY
Email: gisli.jenkins@imperial.ac.uk
Co-investigator organizations:

- University of Nottingham
- University of Edinburgh
- University College London
- University of Sheffield
- University of Oxford
- University of Manchester
- University of Leicester
- University of Liverpool
- University of Southampton
- Royal Brompton Healthcare Trust

Study Management: Dr Mark Weeks (Imperial)

**Study Coordination Centre**
Imperial College London
Respiratory Hub Manager
Room G39
Emmanuel Kaye Building,
1B Manresa Road, London, SW3 6LR

For general queries, supply of study documentation, and collection of data, please contact:

Study Coordinator: Dr M Weeks
Address: Imperial College London **Registration:**
Respiratory Hub Manager
Room G39
Emmanuel Kaye Building,
1B Manresa Road, London, SW3 6LR
Tel: +44 (0)20 7594 7972 E-mail: m.weeks@imperial.ac.uk
Fax: Web address:

**Clinical Queries**
Not Applicable as study will only use anonymized data In the event that any clinical queries arise, these can be addressed to
Chief Investigator: Professor Gisli Jenkins
NIHR Research Professor and Margaret Turner Warwick Chair of Thoracic Medicine
Head of Margaret Turner Warwick Centre for Fibrosing Lung Disease,
National Heart and Lung Institute,
Imperial College London,
Guy Scadding Building,
Cale Street,
London, SW3 6LY
Email: gisli.jenkins@imperial.ac.uk.

**Sponsor**

Imperial College London is the main research Sponsor for this study. For further information regarding the sponsorship conditions, please contact the Head of Regulatory Compliance at:

Research Governance and Integrity team
Imperial College London and Imperial College Healthcare NHS Trust
Room 215, Level 2, Medical School Building
Norfolk Place
London, W2 1PG
**Tel**: **0207 594 9459/ 0207 594 1862**

http://www3.imperial.ac.uk/clinicalresearchgovernanceoffice

**Funder**

This study is funded by a research grant award from UKRI MRC Ref MR/W006111/1.

This protocol describes the UKILD – Long COVID study and provides information for participating partner organizations. Every care was taken in its drafting, but corrections or amendments may be necessary. These will be circulated to investigators in the study. Problems relating to this study should be referred, in the first instance, to the Chief Investigator.

This study will adhere to the principles outlined in the UK Policy Frame Work for Health and Social Care Research. It will be conducted in compliance with the protocol, the Data Protection Act and other regulatory requirements as appropriate.

# Record of Version changes

| Date of version | Version Number | Change |
|---|---|---|
| 15/07/21 to 12/08/21 | V1.1 | Modification to V1.0 to V1.1 in response to REC review. Clarify withdrawal criteria on p12 study protocol and addition of REC committee |
| 26/08/21 | V1.2 | Adaption of inclusion criteria to allow patients who are receiving up to 10mg of prednisolone a day to be recruited, we have found we are having to exclude a lot of patients otherwise. |
| 11/11/21 | V1.3 | Adaption of protocol to include recruitment of data-only participants |
| | | |
| | | |
| | | |
| | | |
| | | |

# 1 INTRODUCTION

## 1.1 Background

The COVID-19 Pandemic has led to over 100 million cases worldwide. In the UK alone, there have been over 4 million cases, over 400 thousand hospital admissions and over 100 thousand deaths. A large number of people diagnosed with COVID-19 suffer from long-term symptoms, predominantly breathlessness and fatigue whether or not they were admitted to hospital. However, long-term symptoms following COVID-19 are more common in people who suffered more severe acute disease [1-4]. There are a number of potential causes of long-term breathlessness following COVID-19 including thromboembolic disease, myo- or pericarditis, and physical deconditioning. However, based on early data from the COVID-19 pandemic, and from other viral infections, a common and potentially severe long-term consequence of COVID-19 is the development of Post COVID related Interstitial Lung Disease (PC-ILD) [5-7].

SARS-CoV-2 leads to pronounced inflammation within the lung and leads to the development of Acute Respiratory Distress Syndrome (ARDS) in a substantial proportion of those infected. Whilst data demonstrate that a short course of corticosteroids can improve survival in patients with hypoxia there is considerable evidence of longer-term inflammation even following short-term corticosteroid therapy [8]. Whilst the inflammatory potential of SARS-CoV-2 is well described, the fibrogenic potential of SARS-CoV-2 is currently unknown but is predicted to be substantial based on the experience of previous coronavirus outbreaks and emerging data from this pandemic [3, 7, 9]. Initial observations have identified substantial levels of post-COVID related ILD (PC-ILD) with up to 24% patients having fibrotic changes, on their initial CT scan, which do not appear to regress at longer time points. Similarly, between 18 - 34% of patients had lung function impairment following COVID-19, consistent with restrictive disease [7]. Risk factors for severe COVID-19 include increasing age, male sex and comorbidities including hypertension and type 2 diabetes mellitus [10, 11], which are also associated with progressive lung fibrosis [10-12].

Pilot data from our consortium indicate that novel functional imaging methods utilising 129Xe and 1H MRI are sensitive to gas transfer limitation, microstructural airway changes and alveolar perfusion deficit in both acute and long COVID patients. These methods also have some ability to dissect both inflammatory and fibrotic pathophysiology and are highly sensitive to disease progression in established ILDs [13-17].

## 1.2 Study Rationale

Given the large number of patients infected with SARS-CoV-2, it is vital that the extent of PC-ILD is determined; its natural history defined particularly whether it is time-limited inflammation and reversible, or develops into persistent, or even progressive, fibrosis. Determining the natural history of PC-ILD, and risk factors as well as biomarkers related to outcome such as disease progression, will enable a precise approach to treatments such as immunomodulation or antifibrotic therapy [6], stratification into clinical trials, prognostication and appropriate service provision. This will facilitate assessment and prioritisation of both conventional and novel therapies used in the treatment of COVID-19 during the acute phase to mitigate the subsequent development of PC-ILD. By exploring the long-term implications of SARS-CoV-2 infection across the full spectrum of COVID-19 disease ranging from non-hospitalised patients managed in the community with mild symptoms to those requiring mechanical ventilation, we will define the risk factors for PC-ILD including disease severity, host genetic factors and the effects of antiviral and immunomodulatory treatment administered during the acute phase of the illness. The UKILD Post COVID study is a prospective multicentre observational cohort study that will be managed through the Imperial College National Heart and Lung Institute and funded by the UKRI Medical Research Council and an NIHR professorship (RGJ).

The PHOSP COVID study (ISRCTN10980107) is a national consortium that provides a platform to study the long-term consequences of COVID19 hospitalisations [18]. An expected 10,000 individuals hospitalised by COVID-19 are to be included in baseline assessments after testing positive on PCR for SARS-C0V-2 and will be recruited through the PHOSP platform. The study described here will aim to recruit a further 2000 individuals, with proven COVID-19, who were not hospitalised but presented to Long-COVID clinics with persistent respiratory symptoms such as breathlessness or cough and are referred for cross-sectional imaging (computer tomography, CT) at baseline (3 months weeks after their first COVID-19 symptoms) [19]. A total of up to 12,000 people will be assessed for longitudinal follow up into the UKILD-Post COVID study. Following assessment of Post COVID patients at baseline (~3-6 months post infection as defined above), those with clinical and radiological features suggestive of ILD will be included into the UKILD-Post COVID study population. Where there are contraindications for CT and if clinically indicated, participants will be eligible for a research-guided 3D ultrashort echo time (UTE) proton MRI as a surrogate for CT. Where individuals meet criteria for initial study inclusion (COVID-19 and clinical indication for CT scanning) but have no clinical, radiological or physiological features of ILD they will be invited to enroll as part of the control cohort for follow up.

This protocol details the research partnership between Imperial (BREATHE Respiratory hub partner), and partner organizations.

## 2 STUDY OBJECTIVES

### 2.1 Primary objective

The primary objective of the study is to determine the prevalence of ILD at 12 months following SARS-CoV-2 infection and whether clinical severity correlates with severity of ILD in survivors.

### 2.2 Secondary objectives

1. to further define the PC-ILD population, in particular, to describe the emerging phenotypes and risk factors of PC-ILD
2. to determine the natural history of PC-ILD phenotypes longitudinally
3. to explore pathomechanisms of PC-ILD for candidate prognostic and theranostic biomarkers.

### 2.3 The primary endpoint

1. The primary endpoint of the study is as radiologically confirmed diagnosis of fibrotic or non-fibrotic ILD in the 12 months following COVID-19. Sub-studies will have a co-primary endpoint of change in the radiological extent of PC-ILD.

### 2.4 The secondary endpoints

1) progressive lung function impairment between 3 and 12 months, defined as ≥10% relative decline in FVC, or ≥10% relative decline in DLco, or increasing radiological extent of PC-ILD using image analysis [20].
2) Resolution of ILD, as defined by ≥10% relative improvement in FVC, DLco, or reduction of radiological extent.

3) Persistence of ILD in those not meeting definition of progression or resolution
4) Presence of interstitial lung abnormalities (ILA) on radiological images that do not meet definition of ILD.
5) A comprehensive series of clinical, molecular, MRI and biochemical parameters will be assessed as biomarkers.

## 3 STUDY DESIGN

Prospective observational study of hospitalised and non-hospitalised patients post- infection with SARS-CoV-2. The study aims to recruit 2000 individuals, with proven COVID-19, who were not hospitalised but presented to Long-COVID clinics with persistent respiratory symptoms such as breathlessness or cough and are referred for cross-sectional imaging (computer tomography, CT) at baseline (3 months weeks after their first COVID-19 symptoms). The study will run for 18 months.

### 3.1 Study procedures objectives and outcomes

| | **Objectives** | **Outcome Measures** | **Timepoint(s)** |
|---|---|---|---|
| **Primary** | 1) Examine and compare the prevalence of ILD in patients with varying severity of COVID-19. (Non-hospitalised will be compared with hospitalised patients and COVID severity defined as varying levels of respiratory support) | MDT diagnosis of ILD at 12 months (+/- 3 months) after acute infection | 12 months (+/- 3 months) post SARS-CoV-2<br><br>NB: Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| **Secondary** | Determine progressive lung function impairment between 3 and 12 months | Lung function decline:<br>≥10% relative decline in FVC or DLco<br>OR<br>Imaging deterioration:<br>increasing radiological extent of PC-ILD using image analysis or MDT discussion | Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | Resolution of ILD | Any of the following: ≥10% relative improvement in FVC, DLco, or<br>reduction of radiological extent. | 12 months (+/- 3 months) post SARS-CoV-2<br><br>Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | Persistence of ILD in those not meeting definition of progression or resolution | | 12 months (+/- 3 months) post SARS-CoV-2<br><br>2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | Persistence of ILAs | | 12 months (+/- 3 months) post SARS-CoV-2 |

| | | | |
|---|---|---|---|
| | | | Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | In hospitalised patients only: Assess the relationship between initial severity of acute infection and ILD | Markers of severity of acute infection (e.g. CRP) | 12 months (+/- 3 months) post SARS-CoV-2<br><br>Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | In hospitalised patients only<br>Assess the impact of specific medical therapy used in the acute setting and ILD e.g. antivirals, immunomodulation | Medical therapy used for acute infection | 12 months (+/- 3 months) post SARS-CoV-2<br><br>Need 2 time points, 6 months apart, between 3 and 15 months post-acute SARS-CoV-2 |
| | Evaluate breathlessness, fatigue and cognition | Medical research council dyspnoea scale<br>Dyspnoea 12<br>FACIT-F<br>Montreal Cognition Assessment (MOCA) | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |
| | Evaluate the quality of life of patients with COVID-19 | Short Form-36 (SF-36)<br>EQ5D-5L | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |
| | Assess impact of COVID19 on aerobic capacity and endurance. | Incremental shuttle walk test or cardiopulmonary exercise test, if clinically indicated | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |
| **Exploratory** | Exploratory circulating biomarkers will be correlated with ILA/ILD | whole blood RNA sequencing (including peripheral blood mononuclear cells) and analysis of epigenetic modifications and proteomics | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |
| | Compare serological and cellular measures of epithelial and endothelial injury and thrombosis | Serum biomarkers of epithelial and endothelial injury, NETosis and thrombosis<br>Cellular measurements in these same patients.<br>Serum/ Plasma: Nordic biomarkers, LRG-1, MMP-7 etc; NETs | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |
| | Compare genetic risk factors for lung fibrosis and radiological progression, continuously and dichotomised | genome-wide array genotyping Genetic testing from blood/ saliva of putative candidate genes and use of GENOMIC data, and telomere length in a subset | No sooner than 3 month visit |
| | Assess the relationship between routinely collected laboratory parameters markers | Laboratory parameters (including CRP, FBC) and extent of ILD in patients | 3 month, 6 months (+/- 6 weeks) and 12 months (+/- 3 months) post SARS-CoV-2 infection |

### 3.2 Description of study procedure(s) Tier 1 and Tier 2

Initial patient approach

Eligibility assessment by clinical team

Study information and invitation letter provided to eligible patients for data only, Tier 1 participation.

**Description of study procedure(s) Tier 2**

History and Examination (10 mins) : Medical history, allergies, medications and anthropometric measurements including height, weight, and body mass index will be recorded during the study visit (or taken from the hospital notes. Or PHOPS data)

Blood sample collection (10 mins) :

Samples will be taken for plasma/ serum for epithelial and endothelial damage biomarkers, coagulation and genetics

Questionnaires (45 mins)

Participants will be provided with seven questionnaires; 36 Short-Form Survey (SF-36), Clinical Frailty Scale (CFS), Patient Health Questionnaire 9 (PHQ9), Montreal Cognition Assessment (MOCA), Dyspnoea-12 score, EQ5D-5L, FACIT-F at the end of each study visit.

There are no investigations required for participants co-enrolled in the PHOSP-COVID study.

3.2.1 **Baseline and follow up assessments**

Initial patient approach

Eligibility assessment by clinical team

Study information and invitation letter provided to eligible patients

Visit 1 (~3 months) (Post COVID-19 infection patients)

1. Review of eligibility with participant
2. Obtain written informed consent
3. History and anthropometric measurements, e.g. height, weight, body mass index (BMI)
4. Blood sample (6-10 ml) collection
5. Pulmonary function test (20 minutes)
6. *Optional* 6-minute walk test (6MWT) (10 minutes)
7. Quality of Life questionnaire (SF-36) (10 minutes)
8. Clinical Frailty Scale (CFS) (5 minutes)
9. Personal health questionnaire (PHQ) (10 minutes)
10. Montreal Cognitive Assessment MOCA (10 minutes)
11. Dyspnea 12 score (10 minutes)
12. EQ5D-5L (10 minutes)
13. FACIT-F questionnaire (10 minutes)

Optional Visit 2 (~6 months) (Post COVID-19 infection patients and matched controls)

All visit 1 assessments will be repeated.

Visit 3 (~12 months) (Post COVID-19 infection patients)

All visit 1 assessments will be repeated.

Visits will be scheduled to take place over 1 day.

Visit 2 is optional,

Visits 1 and 3 are obligatory.

### 3.3 Monitoring of clinical events and re-analysis of previous scans (clinically indicated)

Clinical outcome data and images from clinical scans for all COVID-19 patients will also be collected from around the time of infection and hospitalisation and follow up. Laboratory analyses and chest imaging results undertaken for clinical reasons will also be collected. Access to participant medical records and any relevant hospital data that is recorded as part of routine standard of care; i.e., CT-Scans, blood results and disease progression data etc. will be obtained with patient consent from hospital electronic patient records and NHS digital

### 3.4 Sample Handling Tier 2

Blood samples for the analysis of serum inflammatory markers, serum biomarkers of endothelial and epithelial injury and coagulation studies, viral serology, viral PCR, and whole genome sequencing, ribonucleic acid sequencing and flow cytometry in blood will be collected. Analysis of venous blood samples will be performed in local NHS laboratories. If the Participant consents approximately 10 ml of blood will be stored in the Division of Cardiovascular Medicine for further analysis as part of collaborations with other groups. Samples will not be collected from participants who are also participating in PHOSP-COVID

## 4.0 PARTICIPANT RECRUITMENT

### 4.1 Pre-recruitment evaluations

Screening and Eligibility Assessment

Patients will be assessed by the clinical teams for eligibility against the inclusion or exclusion criteria for Tier 1 and/or Tier 2.

The clinical teams at the recruiting sites will confirm that patients have been infected with SARS-CoV2 before approaching them. Participants must satisfy all the approved inclusion and exclusion criteria of the protocol. If eligible for Tier 1, a ‘data only’ participation, once criteria is confirmed consent can be taken by phone. Participants will then be invited to join the study and informed consent will be obtained at the start of visit 1. For Tier 2, pre-screening for eligibility will take place prior to consent. It will be done by the direct care team using patient notes. There will be no access to patient identifiable data outside of the direct care team prior to consent.

Screening and eligibility for PHOSP COVID/C-MORE participants

For PHOSP-COVID participants, screening and eligibility assessment will occur as per the REC approved protocol (IRAS ID 285439).

All patients can be directed to other relevant post-COVID studies, including but not restricted to POSTCODE and XMAS.

### 4.2 Inclusion Criteria

Tier 1 criteria 'Data Only' participation

1) Age >18 years old -99 years old
2) Evidence of SARS-CoV-2 infection confirmed by PCR or serology at any point during the duration of the pandemic
3) Breathlessness and/or a clinical indication for a chest CT scan as per clinician judgment
4) Participants who have been identified and consented for tier 1 may not meet the inclusion criteria for tier 2 (below) but consent to researcher access to all clinical data relating to their health care during the pandemic and any subsequent follow up in relation to their COVID recovery.

Tier 2

1) Age >18 years old -99 years old
2) Evidence of SARS-CoV-2 infection confirmed by PCR or serology 3 months (+/- 6 weeks) earlier
3) Clinical indication for a chest CT scan as per clinician judgment
4) Participants who have been identified and consented for the main PHOS-PCOVID study are also eligible to join this study (if PHOSP-COVID is still open to recruit at that time).
5) Allow inclusion of patient who are receiving up to 10mg of prednisolone a day

### 4.3 Exclusion Criteria

1) Life-limiting illness within 12 months
2) Significant pre-existing lung disease prior to March 2020, which in the investigator's judgement could make the chest CT scans difficult to interpret

### 4.4 Withdrawal Criteria

Early Discontinuation/Withdrawal of Participants **:** During the course of the study a participant may choose to withdraw early from the study treatment at any time. This may happen for several reasons, including but not limited to:

- The occurrence of what the participant perceives as an intolerable AE.
- Inability to comply with study procedures
- Participant decision

According to the design of the study, participants may have the following three options for withdrawal;

1) Participants may withdraw from active follow-up and further communication but allow the study team to continue to access their medical records and any relevant hospital data that is recorded as part of routine standard of care; i.e., CT-Scans, blood results and disease progression data etc.
2) Participants can withdraw from the study but data and samples obtained up until the point of withdrawal to be retained for use in the study analysis. No further data or samples would be collected after withdrawal.

3) Participants can withdraw completely from the study and withdraw samples collected up until the point of withdrawal. The samples already collected would not be used in the final study analysis.
4) In addition, the Investigator may discontinue a participant from the study at any time if the Investigator considers it necessary for any reason including, but not limited to:
    - Ineligibility (either arising during the study or retrospectively having been overlooked at screening)

Participants that choose to withdraw from the study will not be replaced. The type of withdrawal and reason for withdrawal will be recorded in the CRF. If the participant is withdrawn due to an adverse event, the Investigator will arrange for follow-up visits or telephone calls until the adverse event has resolved or stabilised.

## 5.0 ASSESMENT AND FOLLOW UP

### 5.1 Optional follow up and assessment

As previously detailed in section 3.3 above, Tier 2 participants may also agree to a final follow up assessment and an optional interim assessment

Optional Visit 2 (~6 months) (Post COVID-19 infection patients and matched controls)

All visit 1 assessments will be repeated. All visit 1 assessments will be repeated.

Visits will be scheduled to take place over 1 day. Visit 2 is optional, visits 1 and 3 are obligatory. For research specific visits, participants will be offered reimbursement for travel expenses. This includes reimbursement for petrol, car parking, taxis.

Clinical outcome data and images from clinical scans for all COVID-19 patients will also be collected from around the time of infection and hospitalisation and follow up. Laboratory analyses and chest imaging results undertaken for clinical reasons will also be collected. Access to participant medical records and any relevant hospital data that is recorded as part of routine standard of care; i.e., CT-Scans, blood results and disease progression data etc. will be obtained with patient consent from hospital electronic patient records and NHS digital. The end of the study will be the point at which the final patient follow-up and assessment have been completed.

A detailed Image Management protocol for the UKILD-Post COVID Study provides a comprehensive outline of the purpose, operation, methods, policies, and governance of Image collection for UKILD. It describes the procedures used to collect and store images. (appendix 1)

---

## 6.0 ADVERSE EVENTS

### 6.1 DEFINITIONS

**Adverse Event (AE):** any untoward medical occurrence in a patient or clinical study subject.

**Serious Adverse Event (SAE):** any untoward and unexpected medical occurrence or effect that:

- **Results in death**
- **Is life-threatening** – *refers to an event in which the subject was at risk of death at the time of the event; it does not refer to an event which hypothetically might have caused death if it were more severe*

- **Requires hospitalisation, or prolongation of existing inpatients' hospitalisation**
- **Results in persistent or significant disability or incapacity**
- **Is a congenital anomaly or birth defect**

Medical judgement should be exercised in deciding whether an AE is serious in other situations. Important AEs that are not immediately life-threatening or do not result in death or hospitalisation but may jeopardise the subject or may require intervention to prevent one of the other outcomes listed in the definition above, should also be considered serious.

6.2 REPORTING PROCEDURES

All adverse events should be reported. Depending on the nature of the event the reporting procedures below should be followed. Any questions concerning adverse event reporting should be directed to the Chief Investigator in the first instance.

**6.3 Non serious AEs**

All such events, whether expected or not, should be recorded

**6.4 Serious AEs**

An SAE form should be completed and emailed to the Chief Investigator within 24 hours. However, relapse and death due to COVID -19, and hospitalisations for elective treatment of a pre-existing condition do not need reporting as SAEs.

All SAEs should be reported to the London Riverside REC where in the opinion of the Chief Investigator, the event was:

- 'related', ie resulted from the administration of any of the research procedures; and
- 'unexpected', ie an event that is not listed in the protocol as an expected occurrence

Reports of related and unexpected SAEs should be submitted within 15 days of the Chief Investigator becoming aware of the event, using the NRES SAE form for non-IMP studies. The Chief Investigator must also notify the Sponsor of all related and unexpected SAEs.

Local investigators should report any SAEs as required by their Local Research Ethics Committee, Sponsor and/or Research & Development Office.

**Contact details for reporting SAEs**
RGIT@imperial.ac.uk
**CI email (and contact details below)**
**Fax: N/A, attention** Professor Gisli Jenkins
**Please send SAE forms to:** gisli.jenkins@imperial.ac.uk
**Tel: :** +44 (0)20 7594 7972 **(Mon to Fri 09.00 – 17.00)**

## 7.0 STATISTICS AND DATA ANALYSIS

The prevalence of MDT-confirmed PC-ILD at both early (up to 6 months) and again at late (10-15 month) time-points and will be assessed within the total study population. The prevalence of broader radiological abnormalities and phenotypic patterns will also be assessed in a descriptive analysis, together with demographics, haematological and biochemical profiles, physiological performance, and patient-reported outcome measures. Analyses will be performed overall and stratified according to hospitalised and non-hospitalised, as well as severity of infection in hospitalised patients defined above.

Baseline and longitudinal changes in biomarkers reflecting PC-ILD evolution, including circulating factors and cell-types from detailed serological and cellular analysis, will be assessed in multilevel models for repeated measures to test associations according to the presence or absence of PC-ILD at late time points (10-15 months), and according to progression, resolution, or persistence of radiological patterns over follow up time points.

Analyses will be performed using standard epidemiological and statistical genetics methodology. This will include cross-sectional and longitudinal studies, and analyses of disease prevalence and incidence.
Analysis design and choice of controls will be dependent upon the precise nature of the research question. Identification of risk factors for a specific COVID-19 sequela would involve controls both from within UKILD-Long COVID (without the sequela) and serology-positive controls with prospective questionnaire and healthcare record linkage from Longitudinal Population Studies (for example, UK Biobank, Coronagenes, EXCEED) and pre-existing disease cohorts. Analyses aiming to characterise and understand the clinical features, subtypes and trajectories of sequelae (for example, sarcopaenia) would evaluate the cross-sectional and longitudinal clinical data and biomarkers of UKILD-Long COVID participants who are presenting with the sequelae being studied.
The participant organisations making up the UKILD-long COVID study group have extensive experience of development of, and collaborative use of, disease-specific and general population cohort studies both nationally and internationally enabling access to control populations and alignment of research strategies for rapid validation and replication of findings.

Statistical significance thresholds will be defined in advance of each analysis and will take into account issues of multiple testing and *a priori* evidence.

Data and all appropriate documentation will be stored for a minimum of 10 years after the completion of the study, including the follow-up period.

Project statistician: Dr Iain Stewart email:

iain.stewart@imperial.ac.uk
National Heart and Lung Institute
Guy Scadding Building, Cale Street,
London, SW3 6LY

## 8.0 REGULATORY ISSUES

### 8.1 Ethics approval

The Study Coordination Centre has obtained approval from the Riverside Research Ethics Committee (REC) and Health Regulator Authority (HRA). The study must also receive confirmation of capacity and capability from each participating NHS Trust before accepting participants into the study or any research activity is carried out. The study will be conducted in accordance with the recommendations for physicians involved in research on human subjects adopted by the 18th World Medical Assembly, Helsinki 1964 and later revisions.

### 8.2 Consent

Consent to enter the study must be sought from each participant only after a full explanation has been given, an information leaflet offered and time allowed for consideration. Signed participant consent should be obtained. The right of the participant to refuse to participate without giving reasons must be respected. After the participant has entered the study the clinician remains free to give alternative treatment to that specified in the protocol at any stage if he/she feels it is in the participant's best interest, but the reasons for doing so should be recorded. In these cases the participants remain within the study for the purposes of follow-up and data analysis. All participants are free to withdraw at any time from the protocol treatment without giving reasons and without prejudicing further treatment. Participants who opt for Tier 1 can be consented by phone and provide a signed consent form by post/email.

### 8.3 Confidentiality

The Chief Investigator will preserve the confidentiality of participants taking part in the study and is registered under the Data Protection Act. Data and all appropriate documentation will be stored for a minimum of 10 years after the completion of the study, including the follow-up period which falls within the defined end date .

### 8.4 Indemnity

Imperial College London holds negligent harm insurance policies which apply to this study.

### 8.5 Sponsor

Imperial College London will act as the main sponsor for this study. Delegated responsibilities will be assigned to the NHS trusts taking part in this study.

### 8.6 Funding

This study is funded by a research grant award from UKRI MRC Ref MR/W006111/1. Individual researchers will not receive any personal payments or any other benefits for taking part in this research.

### 8.7 Audits

The study may be subject to inspection and audit by Imperial College London under their remit as sponsor and other regulatory bodies to ensure adherence to GCP and the UK Policy Frame Work for Health and Social Care Research.

## 9.0 STUDY MANAGEMENT

Professor Gisli Jenkins, Chief Investigator and has overall responsibility for the management of the study, the day-to-day management of the study will be co-ordinated through Imperial College National Heart and Lung Institute.

## 10.0 PUBLICATION POLICY

Final results of the study will be disseminated in the form of a manuscript/s in a peer-reviewed journal scientific journals, presented at national and international conferences and in local meetings. In addition, where relevant, data from potential interim analyses will be presented at (a) relevant congress (es). In addition, the study database will be made available for future research.

## 11.0 REFERENCES


1. Carfi, A., et al., *Persistent Symptoms in Patients After Acute COVID-19.* JAMA, 2020. **324**(6): p. 603-605.
2. Huang, C., et al., *6-month consequences of COVID-19 in patients discharged from hospital: a cohort study.* Lancet, 2021. **397**(10270): p. 220-232.
3. Myall, K.J., et al., *Persistent Post-COVID-19 Inflammatory Interstitial Lung Disease: An Observational Study of Corticosteroid Treatment.* Ann Am Thorac Soc, 2021.
4. Evans, R.A., et al., *Physical, cognitive and mental health impacts of COVID-19 following hospitalisation: a multi-centre prospective cohort study.* medRxiv, 2021: p. 2021.03.22.21254057.
5. George, P.M., et al., *Respiratory follow-up of patients with COVID-19 pneumonia.* Thorax, 2020. **75**(11): p. 1009-1016.
6. George, P.M., A.U. Wells, and R.G. Jenkins, *Pulmonary fibrosis and COVID-19: the potential role for antifibrotic therapy.* Lancet Respir Med, 2020. **8**(8): p. 807-815.
7. Laura Fabbri, S.M., Fasihul Khan, Wenjie Chi, Jun Xia, Karen Robinson, Alan Smyth, Gisli Jenkins, Iain Stewart, *Post-viral parenchymal lung disease of COVID-19 and viral pneumonitis: A systematic review and meta-analysis.* medRxiv, 2021(PrePrint): p. 24.
8. Singh, A.K., et al., *Role of corticosteroid in the management of COVID-19: A systemic review and a Clinician's perspective.* Diabetes Metab Syndr, 2020. **14**(5): p. 971-978.
9. Rendeiro, A.F., et al., *The spatial landscape of lung pathology during COVID-19 progression.* Nature, 2021.
10. Hu, J. and Y. Wang, *The Clinical Characteristics and Risk Factors of Severe COVID-19.* Gerontology, 2021: p. 1-12.
11. Jordan, R.E., P. Adab, and K.K. Cheng, *Covid-19: risk factors for severe disease and death.* BMJ, 2020. **368**: p. m1198.
12. Shah, A.S., et al., *A prospective study of 12-week respiratory outcomes in COVID-19-related hospitalisations.* Thorax, 2020.

13. Chan, H.F., et al., *3D diffusion-weighted (129) Xe MRI for whole lung morphometry.* Magn Reson Med, 2018. **79**(6): p. 2986-2995.
14. Chan, H.F., et al., *Airway Microstructure in Idiopathic Pulmonary Fibrosis: Assessment at Hyperpolarized (3)He Diffusion-weighted MRI.* Radiology, 2019. **291**(1): p. 223-229.
15. Saunders, L.C., et al., *Free breathing lung T1 mapping using image registration in patients with idiopathic pulmonary fibrosis.* Magn Reson Med, 2020. **84**(6): p. 3088-3102.
16. Weatherley, N.D., et al., *Quantification of pulmonary perfusion in idiopathic pulmonary fibrosis with first pass dynamic contrast-enhanced perfusion MRI.* Thorax, 2020.
17. Weatherley, N.D., et al., *Hyperpolarised xenon magnetic resonance spectroscopy for the longitudinal assessment of changes in gas diffusion in IPF.* Thorax, 2019. **74**(5): p. 500-502.
18. Dowling, R. *New national study into the long-term health impacts of COVID-19 launched*. [On-line] 2020 05/06/2020; on line press release]. Available from: https://www.phosp.org/study-news/phosp-covid-launching-press-release/.
19. Greenhalgh, T., et al., *Management of post-acute covid-19 in primary care.* BMJ, 2020. **370**: p. m3026.
20. Jacob, J., et al., *Predicting Outcomes in Idiopathic Pulmonary Fibrosis Using Automated Computed Tomographic Analysis.* Am J Respir Crit Care Med, 2018. **198**(6): p. 767-776.

Appendix 1: Image management Protocol for the UKILD-Post COVID study

# Image Management protocol for the UKILD-Post COVID Study

## VERSION 1.0

## 28th June 2021

# Record of Version changes

| Date of version | Version Number | Change |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

# CONTENTS

## Introduction

This document is intended to provide a comprehensive outline of the purpose, operation, methods, policies, and governance of Image collection for UKILD. It describes the procedures used to collect and store images.

The main data collection involved in this protocol is the collection of CT's from a number of healthcare institutes throughout the UK. The Scientific Computing team from the Royal Surrey NHS Foundation Trust (RSNFT) have developed and implemented processes to collect and store digital mammographic images, as part of the CR-UK funded "OPTIMAM" project (IRAS ID: 145706) and the Breast Cancer Now funded "MeDICI" project (IRAS ID: 260722), which have created a large repository of mammographic image for research and training purposes. In addition, the team have been involved with collecting Chest X-rays and CTs for the National Covid-19 Chest Imaging Database and for the PHOSP-Covid study. These projects involve various image and data collection processes which, given appropriate permissions, enables data managers to remotely centralise relevant images and data, remove identifiers, and produce pseudonymised data for use by study researchers. These automated image and data collection procedures, which are now well established, have recently been altered for image collection within the National Covid-19 Chest Imaging Database (NCCID) in a manner to ensure that there is as little as possible disruption to clinical care at the sites participating. These techniques have been redeployed for use in UKILD.

The focus of this document is on how the images are being collected. Particular attention is given to how the data is de-identified at the point of collection by automated processes. The images collected for UKILD are effectively anonymised to researchers and they have no access to the original data or patient identifiers. The NHS/CHI number is encrypted using an AES encryption algorithm and a complex salt to allow checks with the NHS Opt-out service. This encrypted NHS number is stored at the Royal Surrey NHS Foundation Trust and access is maintained by named data managers only. The final location for the de-identified images is infrastructure maintained by the NCIMI team at the university of Oxford.

The collection processes, transfer and encrypted NHS/CHI numbers are managed and maintained by data managers in the Medical Physics Department at the Royal Surrey NHS Foundation Trust. The data managers are trained in information governance policies of the host Trust.

## Image collection and storage procedures

The first principle of the image collection system is that all data that is transferred from the clinical sites to the central image database is fully de-identified and that at no point do researchers have access to patient identifiable information. The NHS/CHI number is encrypted using an AES encryption algorithm and a complex salt to allow checks with the NHS Opt-out service (See Pseudonyms). This encrypted NHS number is stored at the Royal Surrey NHS Foundation Trust and access is maintained by named data managers only. The final location for the de-identified image and data is a is infrastructure maintained by the NCIMI team at the university of Oxford.

## Pseudonyms

To prevent the need to store Patient Identifiable Data (PID), UKILD utilises a trial ID in the place of Patient IDs. In order to obtain the Trial ID a web portal will be utilised to allow sites to submit trial IDs linked to pseudonym hashes (See SMART Portal Trial ID Registration section). Secondary pseudonyms are created through a one-way encryption process known as 'hashing' which provides the ability to link the images with the Trial ID and link with other datasets that have generated these pseudonyms in the same manner using the same complex salt. The process is described below:

1. The input string (e.g., NHS/CHI number) is combined with a fixed complex salt, and then fed into a hashing algorithm (See Figure 1).
2. The hashing algorithm used is SHA-256 (256-bit Secure Hashing Algorithm), which was developed by the National Security Agency (NSA).
3. The hashed patient IDs are stored within the database and are linked to the registered patient and any images received for that patient.

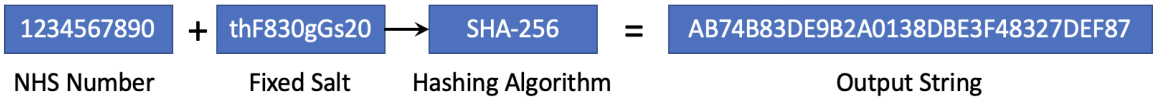



Figure 1: Illustration of hashing algorithm with fixed salt to generate pseudonyms

4. The fixed complex salt is stored on the database server and is applied to all NHS and hospital IDs.

5. It is important to understand that the hashing process is one-way only - it is not possible to reconstruct the original NHS or hospital ID from the hashed string, even if the salt is known.

6. However, we can compare the hashed string to another hashed string which has gone through the same process described in the diagram and thus be able to match up patient identifiers without ever exposing their original data.

The pseudonyms are generated during the image collection process which generates hashed identifiers from the DICOM tags. This process is undertaken before the images are received from IEP and on the client-side of the SMART portal web form.

## Image Collection from Healthcare Institutes

The following section describes the image collection processes that are deployed at numerous healthcare institutes around the UK. Each site is provided with training and guidance documentation to inform the users on the collection processes. Sites are asked for the participating client's identifiers to be registered on the SMART Portal and imaging to be sent to a dedicated Research Node on the Image Exchange Portal. The images and data are de-identified at the point of collection and transferred to NCIMI.

Figure 2 provides a graphical overview of the image collection and storage procedures used

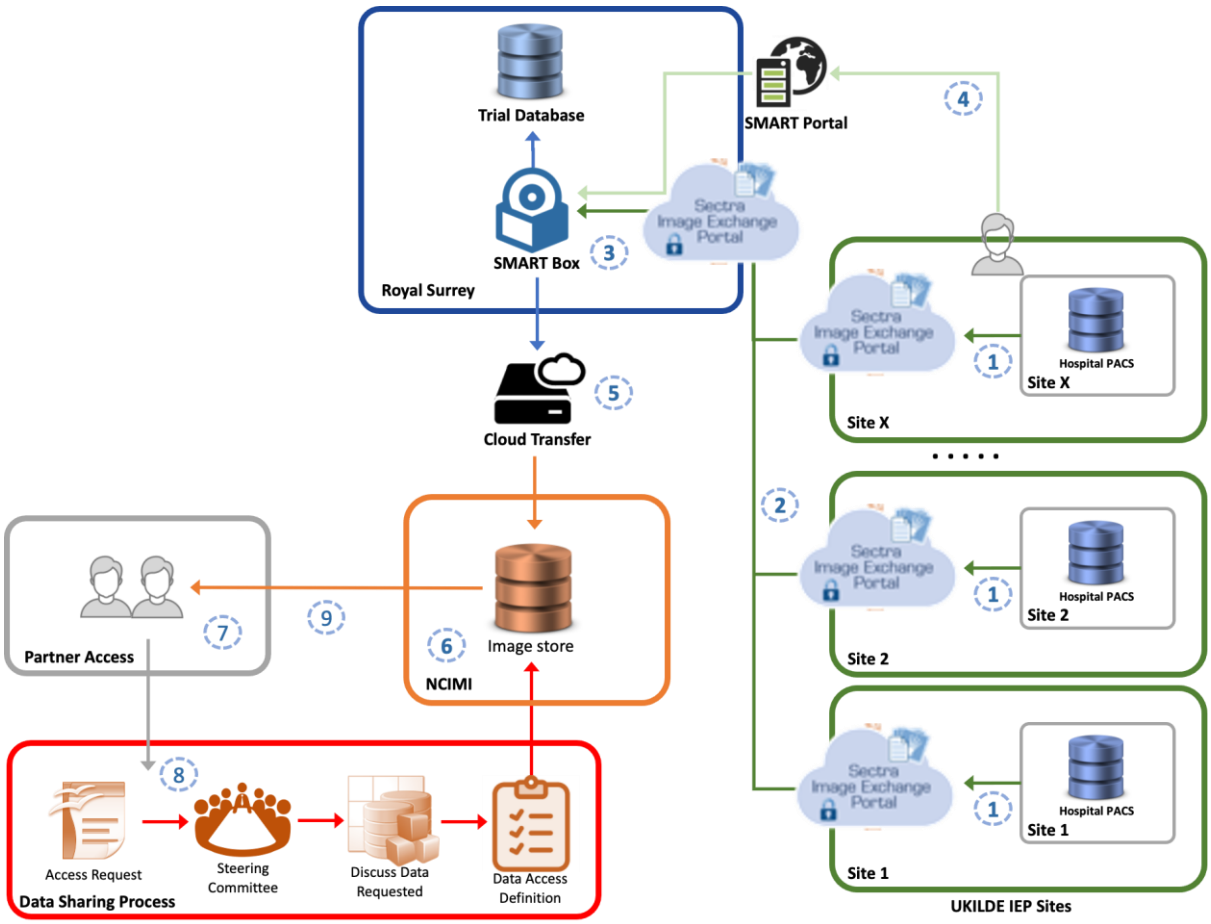



Figure 2: Overview of the image and data collection, transfer and storage processes.

Communications are through HTTPS and data at rest is encrypted with DEKs using AES256. There are two tables in the database which store the data - the other tables are for support only.

a) Clients - This table contains the required information about an individual client (patient). The primary and secondary hashed IDs are stored in this table. There is also an optional field for storing an AES encrypted primary ID (NHS number).

b) Images - This table is populated each time a new image is received by the image receiver software. Each image is written into a new row in the table, and it is linked to the Client, so that we can link images to clients.

The full schema for the collector is shown below.

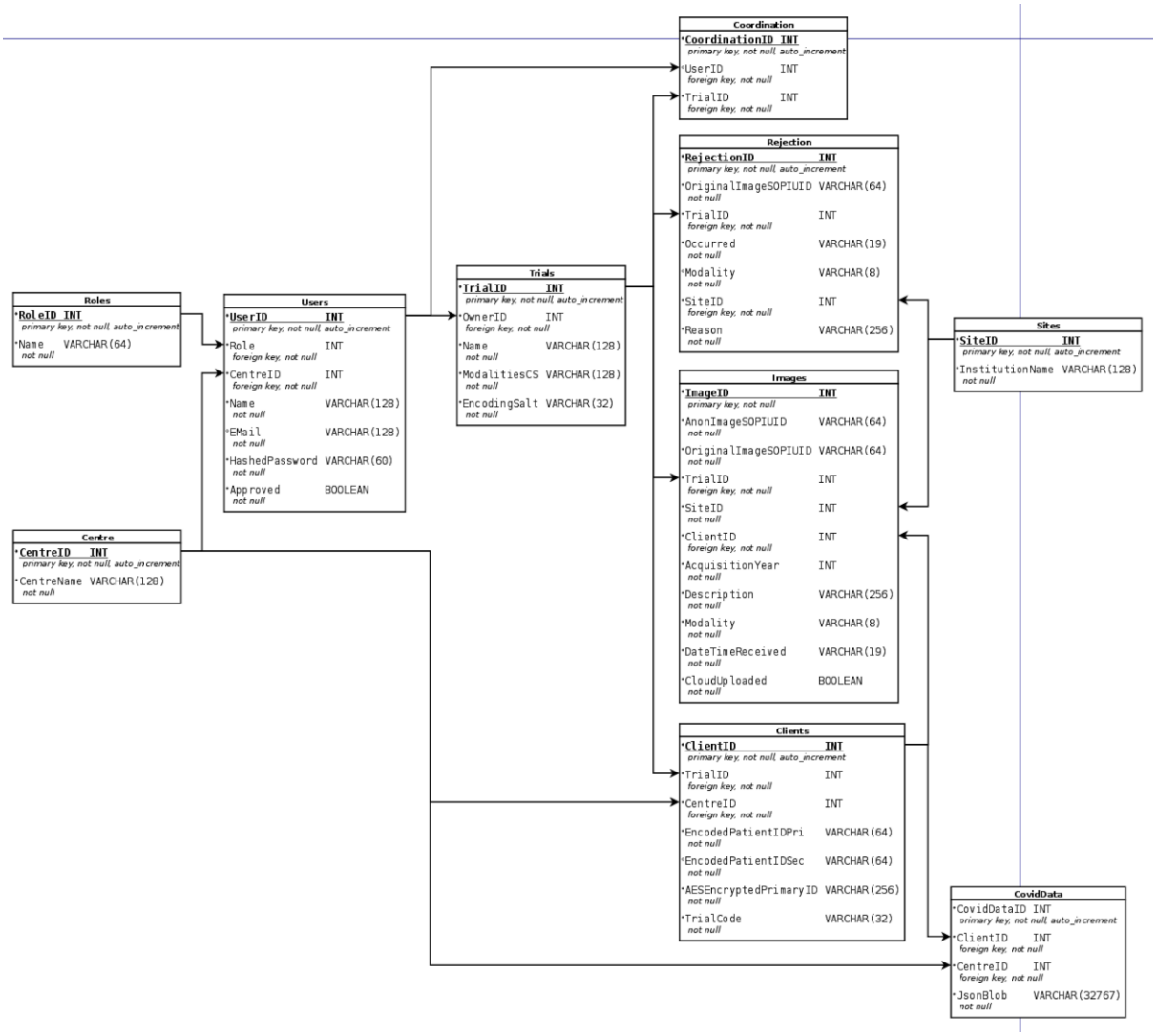



Figure 3: Database schema for the data receiver

## SMART Portal Trial ID Registration

UKILD Study sites will register the participants on the SMART Portal in advance of sending the images. This step is required to ensure that the Image collection system is informed of the Trial ID of the participant before the image is received, allowing the image to be collecting identified as belonging to UKILD and providing the pseudonym to insert into the DICOM image. The Trial ID uploaded and the incoming image will be linked through the use of a common temporary pseudonym in the form of a hashed NHS number (see Pseudonyms Section).

Patient identifiers are de-identified at the point of upload. This means that no patient identifiable information will leave the site. Upon upload, a pseudonym will be assigned to the patient. This will be created by a lossful encoding algorithm and complex salt which will produce an encoded

pseudonym that allows the linking of the patient's trial ID with the images. In addition, the NHS/CHI number will also be encrypted using an AES encryption algorithm and a complex salt to allow linking with national datasets.

When a client is registered on the SMART portal the central collection server is queried for an encoding salt, which is transmitted to the local computer. This encoding salt is unique to the trial, and is used to hash the NHS number, as well as the secondary patient identifier (e.g., hospital number) before it is transmitted over the Internet to the central server. The hashing technology used is SHA-256 and this hash cannot be reversed back to the original data. No NHS or patient identifier is ever transmitted to the central server using the SMART portal.

Each site is issued with a user account for the SMART Portal with an initial password from the Royal Surrey Scientific Computing team. Sites will register a client by entering an NHS/CHI number, a secondary ID, which must be the ID used by radiology (unless radiology uses NHS/CHI), as well as the trial code which uniquely identifies the client (See Figure 4).

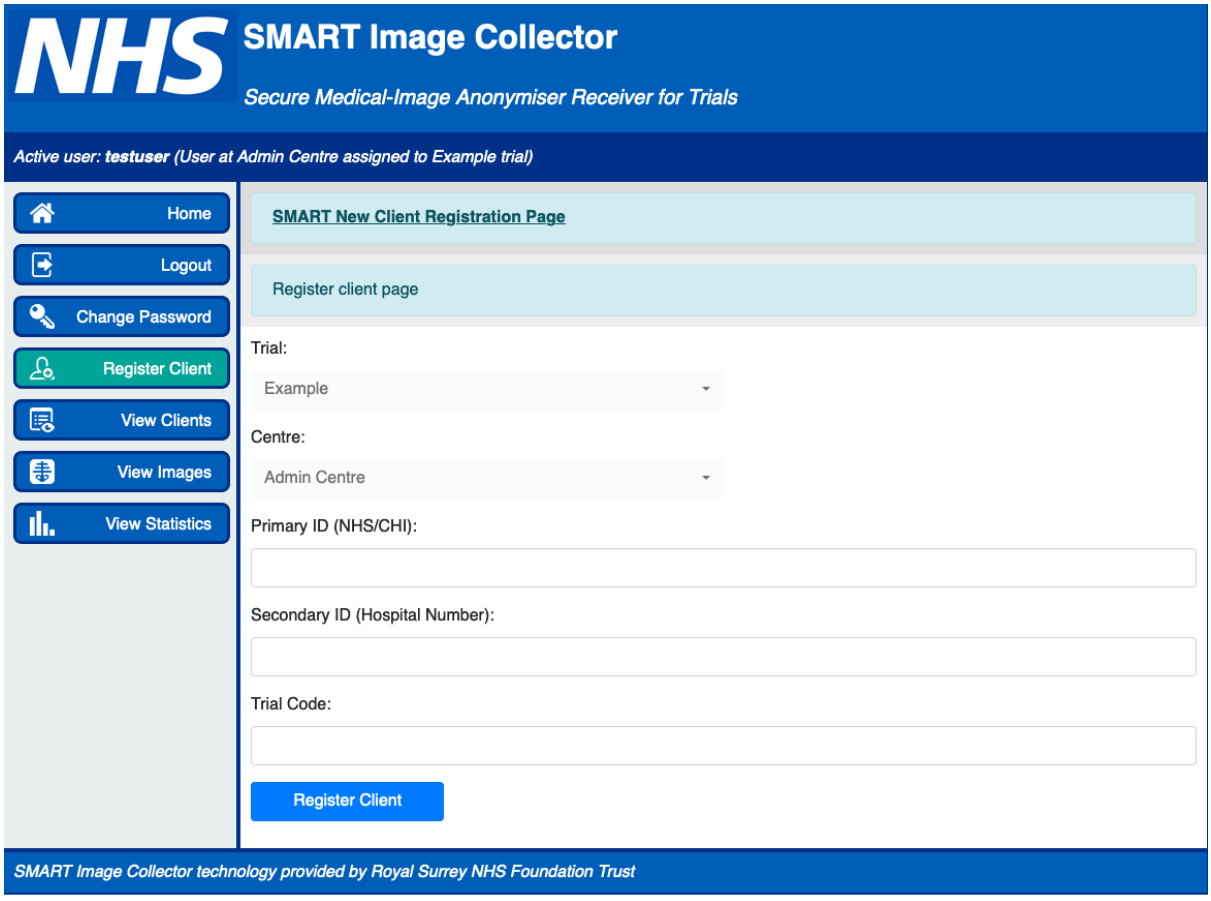



Figure 4: Register Client page on SMART Portal

## Image Collection

UKILD collects images using the Image Exchange Portal (IEP). This is advantageous as the processes for transferring images via IEP are well known and robust. IEP is in use by almost all NHS sites and private hospitals in the UK and is in routine use to facilitate the transfer of medical images from one healthcare site to another. At most clinical sites, a large number of staff are trained in the use of IEP and indeed most sites have a robust image transfer request service in place (usually run by the PACS team). UKILD has a dedicated node on IEP available for use for receiving images that are to be de-identified. This node is provided by the RSNFT. The node has been set up such that the images are de-identified as they transition from IEP, meaning that no patient identifiable information is received outside the IEP network. A pseudonym is generated using the same technique as described in the section above (Pseudonyms). This results in a common pseudonym between the images and the linking with the UKILD Smart portal entry.

Collection sites are requested to submit images through IEP when a new patient is enrolled into UKILD. Many sites have internal processes for requesting image transfers via IEP (often through the PACS office). Guidance is provided to utilise these processes where possible. Guidance is also provided showing a step-by-step guide on how to submit images through IEP.

Once an image has been sent to the UKILD dedicated node, a transfer is initiated to the RSNFT node receiver. The node receiver is a DICOM router (DICOM XR) designed to receive images and route them to different study modules based on sets of rules. As part of the DICOM XR process the received DICOM image is written to a virtual "RAMDISK" to act as the temporary location. The purpose of the RAMDISK is to provide a non-permanent file-storage location for the temporary storage of images, thus avoiding the need of ever writing patient identifiable information to permanent storage (e.g., hard disc). Before progressing further, the incoming images are filtered using the following criteria:

The incoming images are filtered (discarded) by a number of rules:

1. The accepted ClassSOPUID is limited to a small set listed. This filter effectively removes any Secondary Capture Objects, Dose reports and anything that is not a primary image: [class_uid:00080016 :STRING: MATCH_ANY ]{'1.2.840.10008.5.1.4.1.1.1', '1.2.840.10008.5.1.4.1.1.1.1', '1.2.840.10008.5.1.4.1.1.1.1.1', '1.2.840.10008.5.1.4.1.1.2', '1.2.840.10008.5.1.4.1.1.2.1', '1.2.840.10008.5.1.4.1.1.4', '1.2.840.10008.5.1.4.1.1.4.1', '1.2.840.10008.5.1.4.1.1.4.2'}

2. The accepted modality is limited to the following Modality tag [modality:00080060:STRING:MATCH_ANY]{'DX', 'CR', 'MR', 'CT'}.

3. Any CT series that have ImageType (.*)DERIVED(.*) are removed

4. As a final filter, the images are passed through a further filter provided by the MIRC-CTP tool set. This is a curated list of DICOM tags and modality/model specific rules that could

potentially indicate the presence of PID. This tool highlights (provides a log) any such cases found and either removes the images or blackouts known PID potential areas. Chest X-ray DERIVED data have to be added to the whitelist. Logs are reviewed on a daily basis to follow up on any cases indicated.

Once filtered, the accepted images are de-identified. The de-identification of the image takes place in RAMDISK, including the replacement of the PatientID tag with the pseudonym (described above). Details of the DICOM tags that are anonymised and those that are pseudonymised are given in Appendix 1. The acceptability of our pseudonymisation procedures was extensively reviewed by staff at the clinical sites during the original OPTIMAM project (2008--2013) and subsequent OPTIMAM projects (2013-2020), MeDICI (Current), FastMRI (Current) and the NCCID project (Current). We took advice from our Information Governance Officer on the acceptability of our procedures. All de-identification is in compliance with DICOM Supplement 142.

### 2.2.4 Data Types

UKILD collects CT images and DICOM header information (de-identified). We will be collecting all follow up images for patients enrolled in UKILD. The categories of image data collected are:

- Chest X-ray imaging at one or multiple timepoints
- CT imaging of the chest including High-resolution imaging, contrast and non-contrast enhanced imaging, CT pulmonary angiograms, inspiratory and expiratory phase imaging and prone CT imaging. The CT imaging will be at one or more timepoints in each patient.
- The imaging will be acquired from hospitalised and non-hospitalised COVID-19 patients

### 2.2.5 Data Flow

Images transferred via IEP to the UKILD node (hosted at the RSNFT). RSNFT is providing the image collection services for UKILD. The RSNFT stores the incoming images on a fully redundant RAID 5+1 Synology (http://www.synology.com/) NAS server, with 54 TB capacity. This storage is considered temporary storage as the data is only transitioned through the RSNFT to its final destination. While the images reside on the RSNFT storage, for added insurance in the case of hardware failure, the system is replicated each evening to a mirror located in a separate building within the Trust. Both the storage and backup systems are also replicated to redundant systems to ensure continuous availability. Regular database dumps of all databases are made every two hours and archived on the backup systems. Data is transferred from the RSNFT to NCIMI on a regular basis. Once confirmation is received that the data transfer is successful, the images and data are removed from the RSNFT temporary storage and backups.

## 2.3 Transfer of De-Identified Images to NCIMI

On a regular basis, dictated by the frequency of the data uploads, a job running on a virtual environment at RSNFT will transfer images to a shared transient cloud storage. NCIMI will download and remove. All images will be encrypted in flight during the transfer using Secure HTTP (HTTPS) protocol.

## 2.4 Access to Patient Identifiable Information

Images and data are de-identified at the point they are transferred to UKILD. The staff transferring the images via IEP will require access to patient identifiable information. However, these staff will be employees of the collection site. The NHS numbers of the patients whose images are collected are pseudonymised as detailed above. The NHS/CHI number is also encrypted using an AES encryption algorithm and a complex salt in order to allow checking with the NHS Opt-out service. This encrypted NHS number is stored at the Royal Surrey NHS Foundation Trust and access is maintained by named data managers only.

## 2.5 Linking to External Datasets

The ability to link to other data sources is essential. In order to facilitate this there are two options available. The first is linking via the encoded pseudonym and the second is through the use of two-way encrypted NHS numbers.

### 2.5.1 Linking via encoded pseudonyms

It is possible to link UKILD with external datasets via a shared encoded pseudonym. This is relevant for linking to other studies which have access to the original NHS number. RSNFT tools can be utilised by the data managers of the other studies that can be applied to a list of NHS numbers. This would result in a list of encoded pseudonyms that would enable the two datasets to be linked. This process is detailed in Appendix 2

### 2.5.2 Linking via encrypted NHS numbers

The NHS/CHI number is encrypted using an AES encryption algorithm and a complex salt and stored at the RSNFT. This is in order to facilitate linking with national datasets that require a list NHS numbers to be submitted. No researchers will ever have access to this encrypted NHS number, and it is stored physically separate to the central data warehouse. If a data access request for linked data was approved, then the RSNFT staff would be able to utilise automated software tools to decrypt the NHS numbers and submit them to the national dataset retrieval processes.

Imperial College London

Imperial College London

## Appendix 1: Anonymisation details

Each medical image is stored in a standard DICOM format. In addition to the actual image there is a DICOM header with information stored in fields that are known as “tags” which contain a very large amount of information including patient identifiable information. In order to preserve the confidentiality of the patients the following changes are made to each DICOM header at the clinical site before it is transferred to the central warehouse.

All the tags not mentioned below are retained as specified in the DICOM supplement 142 standards for anonymisation as they do not contain patient data or data likely to permit patient data to be accessed.

### Pseudonymisation of DICOM tags

The following tags are pseudonymised:

**0010,0010: PatientName:** This is replaced with the pseudonym

**0010,0020: PatientID:** This is replaced with the pseudonym

**0010,0030: PatientBirthDate:** This is altered, so that the year remains the same, but the day and month are set to “01”

**Study and Series Description are retained**

**Private Tags are removed**

### Nulling of DICOM tags

The following tags are either removed or nulled:

4008,0111 - Interpretation Approver Sequence
0018,9424 - Acquisition Protocol Description
0040,2010 - Order Callback Phone Number
4008,0300 - Impressions
4008,0118 - Results Distribution List Sequence
0400,0100 - Digital Signature UID
4008,0119 - Distribution Name
0012,0031 - Clinical Trial Site Name
0040,0253 - Performed Procedure Step ID
0012,0030 - Clinical Trial Site ID
0040,0254 - Performed Procedure Step Description
4008,0114 - Physician Approving Interpretation
60XX,4000
4008,0115 - Interpretation Diagnosis Description
0040,2016 - Placer Order Number / Imaging Service Request
6000,4000 - Overlay Comments
0012,0021 - Clinical Trial Protocol Name
0040,2009 - Order Enterer’s Location
4008,011A - Distribution Address

0040,2008 - Order Entered By
0040,0248 - ?
0028,4000 - Image Presentation Comments
0040,A027 - Verifying Organization
4008,0202 - Interpretation ID Issuer
4008,0102 - Interpretation Recorder
0012,0062 - Patient Identity Removed
0012,0060 - Clinical Trial Coordinating Center Name
0012,0020 - Clinical Trial Protocol ID
0012,0010 - Clinical Trial Sponsor Name
0010,0010 - Patient's Name
0010,2299 - Responsible Organization
4008,010B - Interpretation Text
0040,2017 - Filler Order Number / Imaging Service Request
4008,010A - Interpretation Transcriber
0010,2297 - Responsible Person
4008,010C - Interpretation Author
0008,010D - Context Group Extension Creator UID
0040,1400 - Requested Procedure Comments
0070,0001 - Graphic Annotation Sequence
300A,0013 - Dose Reference UID
0040,0275 - Request Attributes Sequence
0012,0051 - Clinical Trial Time Point Description
0012,0050 - Clinical Trial Time Point ID
0010,0020 - Patient ID
0040,1004 - Patient Transport Arrangements
0040,1001 - Requested Procedure ID
0010,0021 - Issuer of Patient ID
0010,2180 - Occupation
0032,0012 - Study ID Issuer
0040,2400 - Imaging Service Request Comments
0040,0280 - Comments on the Performed Procedure Step
0040,000B - Scheduled Performing Physician Identification Sequence
0010,1000 - Other Patient IDs
0010,1001 - Other Patient Names
0020,9161 - Concatenation UID
0020,9164 - Dimension Organization UID
0040,1005 - Requested Procedure Location
0038,1234 - ?
0012,0040 - Clinical Trial Subject ID
0040,0006 - Scheduled Performing Physician's Name
0040,0007 - Scheduled Procedure Step Description
0040,2001 - Reason for the Imaging Service Request
0012,0042 - Clinical Trial Subject Reading ID
0040,0004 - Scheduled Procedure Step End Date
0040,0005 - Scheduled Procedure Step End Time
0032,1021 - Scheduled Study Location AE Title
0040,0012 - Pre-Medication
0040,1010 - Names of Intended Recipients of Results
0040,0011 - Scheduled Procedure Step Location
0032,1020 - Scheduled Study Location
0040,1011 - Intended Recipients of Results Identification Sequence
0010,0032 - Patient's Birth Time
0040,0010 - Scheduled Station Name
0010,0030 - Patient's Birth Date
0020,0052 - Frame of Reference UID
0040,A07C - Custodial Organization Sequence
0040,A07A - Participant Sequence
0008,2112 - Source Image Sequence
0040,A075 - Verifying Observer Name
0008,0096 - Referring Physician Identification Sequence
0008,0094 - Referring Physician's Telephone Numbers
0088,0910 - Topic Author

0040,A078 - Author Observer Sequence
0008,0092 - Referring Physician's Address
0088,0912 - Topic Keywords
0008,1060 - Name of Physician(s) Reading Study
0008,1062 - Physician(s) Reading Study Identification Sequence
0032,1030 - Reason for Study
0040,0001 - Scheduled Station AE Title
0008,2111 - Derivation Description
0032,1032 - Requesting Physician
0040,A073 - Verifying Observer Sequence
0032,1033 - Requesting Service
0040,0002 - Scheduled Procedure Step Start Date
0040,0003 - Scheduled Procedure Step Start Time
0010,1090 - Medical Record Locator
2030,0020 - Text String
0010,0102 - Patient's Primary Language Modifier Code Sequence
0008,0090 - Referring Physician's Name
0010,0101 - Patient's Primary Language Code Sequence
0020,4000 - Image Comments
0038,0400 - Patient's Institution Residence
0088,0906 - Topic Subject
0008,1195 - Transaction UID
0008,0080 - Institution Name
3006,00C2 - Related Frame of Reference UID
0008,0081 - Institution Address
0008,0082 - Institution Code Sequence
0008,1050 - Performing Physician's Name
0008,1052 - Performing Physician Identification Sequence
0020,0200 - Synchronization Frame of Reference UID
0008,0201 - Timezone Offset From UTC
4008,4000 - Results Comments
3006,0024 - Referenced Frame of Reference UID
0020,3406 - Modified Image Description
0020,3404 - Modifying Device Manufacturer
0008,1084 - Admitting Diagnoses Code Sequence
0020,3401 - Modifying Device ID
0040,0245 - Performed Procedure Step Start Time
0008,1080 - Admitting Diagnoses Description
0040,0244 - Performed Procedure Step Start Date
0040,0243 - Performed Location
0040,0242 - Performed Station Name
0040,0241 - Performed Station AE Title
0040,3001 - Confidentiality Constraint on Patient Data Description
0020,0010 - Study ID
0032,4000 - Study Comments
0020,0012 - Acquisition Number
0040,A088 - Verifying Observer Identification Code Sequence
0020,000D - Study Instance UID
0020,000E - Series Instance UID
0008,1072 - Operator Identification Sequence
0032,1060 - Requested Procedure Description
0008,1070 - Operators' Name
FFFA,FFFA - Digital Signatures Sequence
0008,0058 - Failed SOP Instance UID List
0038,0010 - Admission ID
0038,0011 - Issuer of Admission ID
0040,A730 - Content Sequence
0008,0050 - Accession Number
0010,1050 - Insurance Plan Identification
0038,0500 - Patient State
0038,001E - Scheduled Patient Institution Residence
0010,21B0 - Additional Patient History
0018,4000 - Acquisition Comments

0018,1030 - Protocol Name
0008,1155 - Referenced SOP Instance UID
0008,1010 - Station Name
0032,1070 - Requested Contrast Agent
0038,0021 - Admitting Time
0038,0020 - Admitting Date
0400,0403 - Referenced SOP Instance MAC Sequence
0008,4000 - Identifying Comments
0400,0402 - Referenced Digital Signature Sequence
0010,21C0 - Pregnancy Status
0010,1060 - Patient's Mother's Birth Name
0040,A124 - UID
0038,0300 - Current Patient Location
0040,A123 - Person Name
0010,2203 - Patient's Sex Neutered
0010,2110 - Allergies
0008,1040 - Institutional Department Name
0018,1000 - Device Serial Number
0018,1002 - Device UID
0008,1049 - Physician(s) of Record Identification Sequence
0018,1005 - Generator ID
0018,1004 - Plate ID
0018,1007 - Cassette ID
0008,1048 - Physician(s) of Record
0008,1140 - Referenced Image Sequence
0018,1008 - Gantry ID
FFFC,FFFC - Data Set Trailing Padding
4008,0042 - Results ID Issuer
0018,0010 - Contrast/Bolus Agent
0070,0086 - Content Creator's Identification Code Sequence
0070,0084 - Content Creator's Name
0008,1030 - Study Description
0040,DB0D - Template Extension Creator UID
0040,DB0C - Template Extension Organization UID
0038,0040 - Discharge Diagnosis Description
0010,2000 - Medical Alerts
0088,0200 - Icon Image Sequence
0018,700A - Detector ID
0400,0561 - Original Attributes Sequence
0010,21A0 - Smoking Status
0088,0140 - Storage Media File-set UID
0038,4000 - Visit Comments
0010,1080 - Military Rank
0018,1400 - Acquisition Device Processing Description
0010,1081 - Branch of Service
4000,0010 - Arbitrary
0070,031A - Fiducial UID
0008,103E - Series Description
0010,1002 - Other Patient IDs Sequence
300E,0008 - Reviewer Name
0400,0550 - Modified Attributes Sequence
0008,1120 - Referenced Patient Sequence
0010,1005 - Patient's Birth Name
0038,0050 - Special Needs
50XX,XXXX
0008,0018 - SOP Instance UID
0000,1001 - Requested SOP Instance UID
0008,0012 - Instance Creation Date
0008,0013 - Instance Creation Time
0008,0014 - Instance Creator UID
0008,2112 - Source Image Sequence
0020,9158 - Frame Comments
0040,0555 - Acquisition Context Sequence

0010,21F0 - Patient's Religious Preference
0038,0060 - Service Episode ID
0010,1010 - Patient's Age
0008,3010 - Irradiation Event UID
4000,4000 - Text Comments
0008,1110 - Referenced Study Sequence
0028,1199 - Palette Color Lookup Table UID
0038,0062 - Service Episode Description
0002,0003 - Media Storage SOP Instance UID
0038,0061 - Issuer of Service Episode ID
0008,1111 - Referenced Performed Procedure Step Sequence
0010,0050 - Patient's Insurance Plan Code Sequence
0010,2160 - Ethnic Group
0040,4027 - Scheduled Station Geographic Location Code Sequence
0040,4028 - Performed Station Name Code Sequence
0040,0404 - ?
0040,4023 - Referenced General Purpose Scheduled Procedure Step Transaction UID
0040,4025 - Scheduled Station Name Code Sequence
0000,0021 - ?
0010,1020 - Patient's Size
0008,0030 - Study Time
0008,0031 - Series Time
0018,A003 - Contribution Description
0010,2150 - Country of Residence
0008,0033 - Content Time
0008,0034 - Overlay Time
0008,9123 - Creator-Version UID
0010,2152 - Region of Residence
0008,0035 - Curve Time
0010,2154 - Patient's Telephone Numbers
0040,4036 - Human Performer's Organization
0040,4035 - Actual Human Performers Sequence
0040,4034 - Scheduled Human Performers Sequence
0040,4037 - Human Performer's Name
0010,21D0 - Last Menstrual Date
0040,4030 - Performed Station Geographic Location Code Sequence
0010,1030 - Patient's Weight
60XX,3000
0008,0021 - Series Date
0040,1103 - Person's Telephone Numbers
0008,0024 - Overlay Date
0008,0025 - Curve Date
0040,1102 - Person's Address
0040,1101 - Person Identification Code Sequence
0008,0023 - Content Date
0008,1511 - ?
0010,1040 - Patient's Address

Imperial College London

## Appendix 2: Guidelines for Linking to External Datasets

### Guidelines on using NHS Hasher tool to link data sources

This document outlines the use of the NHS Hasher tool to create link-able pseudonyms which would allow disparate datasets to be linked.

## Pseudonyms

1. Pseudonyms are created through a one-way encryption process known as 'hashing'.
2. The input string (e.g. NHS number) is combined with a fixed complex salt, and then fed into a hashing algorithm.
3. The hashing algorithm used is SHA-256 (256-bit Secure Hashing Algorithm), which was developed by the National Security Agency (NSA).
4. The hashed patient IDs are stored and are linked to the registered patient and any images received for that patient.

A simple diagram to explain this hashing process is shown below (these are just example values);


5. The fixed complex salt is stored on the database server and is applied to all NHS and hospital IDs.
6. It is important to understand that the hashing process is one-way only - it is not possible to reconstruct the original NHS or hospital ID from the hashed string, even if the salt is known.
7. However, we can compare the hashed string to another hashed string which has gone through the same process described in the diagram and thus be able to match up patient identifiers without ever exposing their original data.

In either case the process is entirely automated, and the original NHS or hospital identifier is never retained.

Imperial College London

## Creating a Link-able Dataset NHS Hashing Tool

A tool created by the Royal Surrey NHS Foundation Trust called the "NHS Hashing Tool" allows users to quickly generate a list of hashed patient identifiers and assign unique trial identifiers to each patient.

The hashing tool works with Comma Separated Value (CSV) files. These are easily obtained from Excel files by using "Save As…" and selecting CSV as the save type.

See the appendix 2 for a step-by-step guide to using the hashing tool.

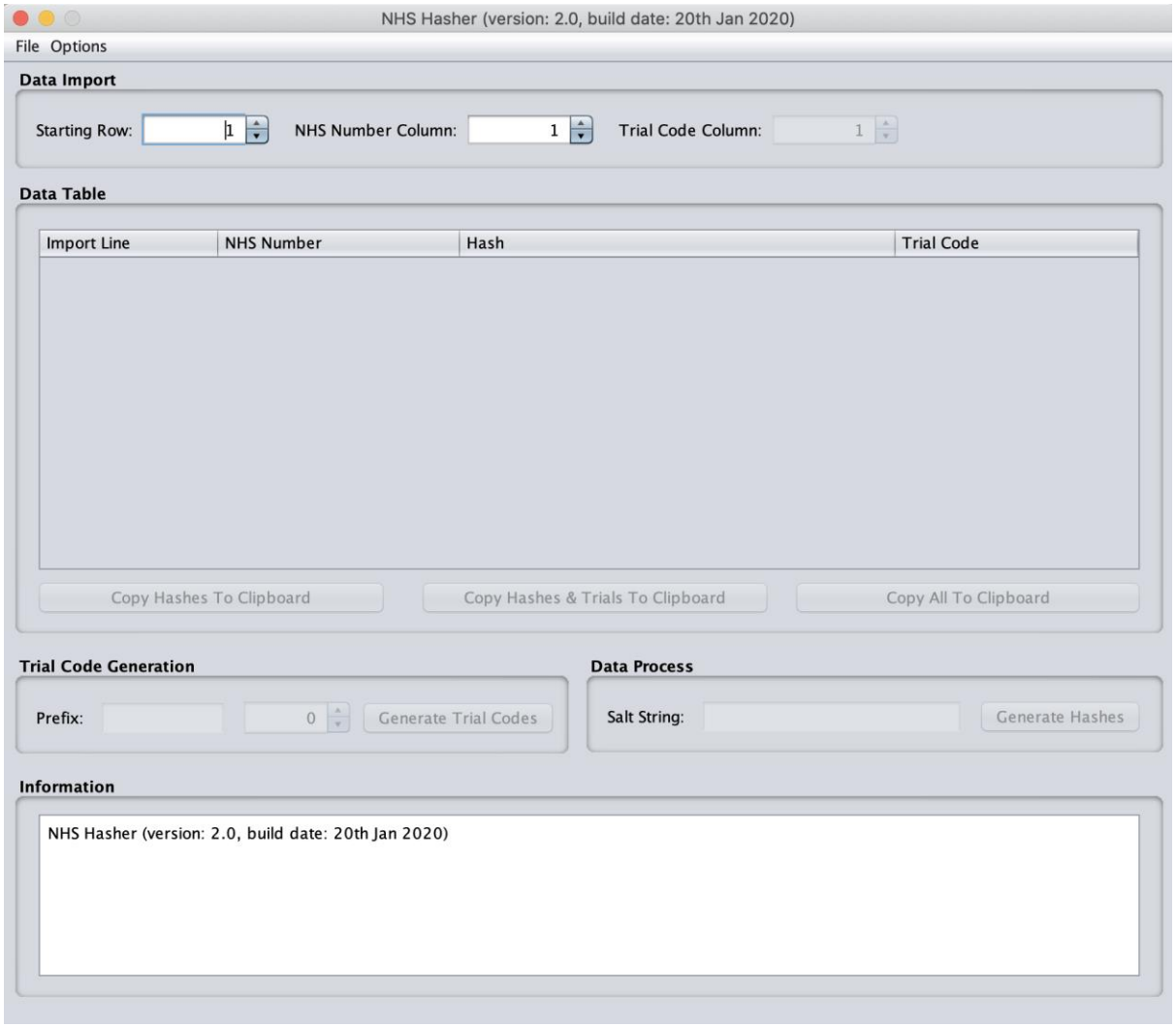
It is important to note that most of these steps can be automated in the case that a regular feed of data is required.

Step 1. Prepare your list of NHS numbers to 'hash'. This can be a simple Excel file with a single column;

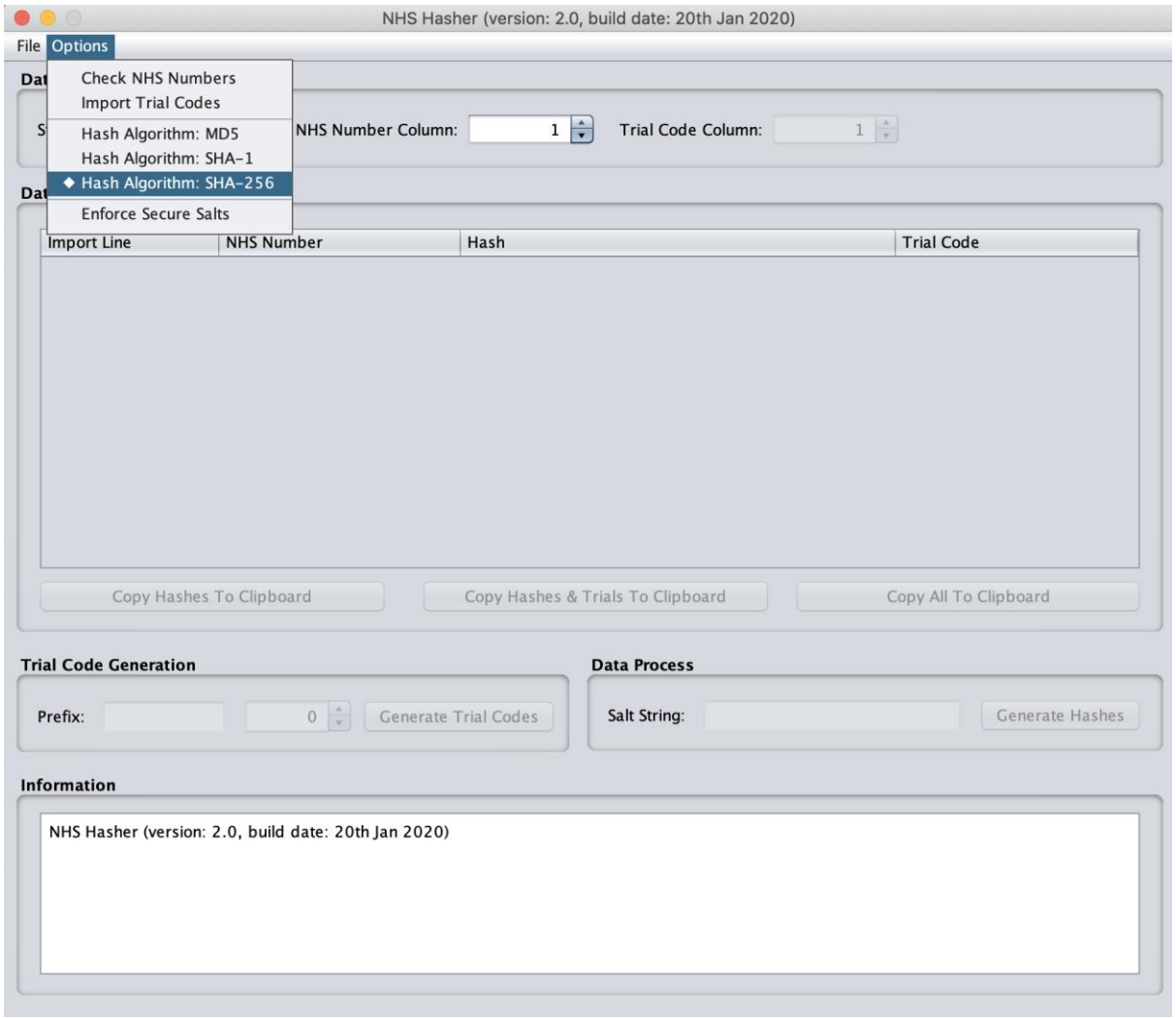
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | NHS Number | | | | |
| 2 | 1111111111 | | | | |
| 3 | 2222222222 | | | | |
| 4 | 3333333333 | | | | |
| 5 | 4444444444 | | | | |
| 6 | 5555555555 | | | | |
| 7 | 6666666666 | | | | |
| 8 | 7777777777 | | | | |
| 9 | | | | | |
| 10 | | | | | |
| 11 | | | | | |

Step 2. The hashing tool only works with "CSV" files. These are easily obtained from Excel, simple select File->Save As… and then select "CSV UTF-8 (Comma Separated)" from the list of options. Save the file to a location you can easily find (e.g. Desktop);

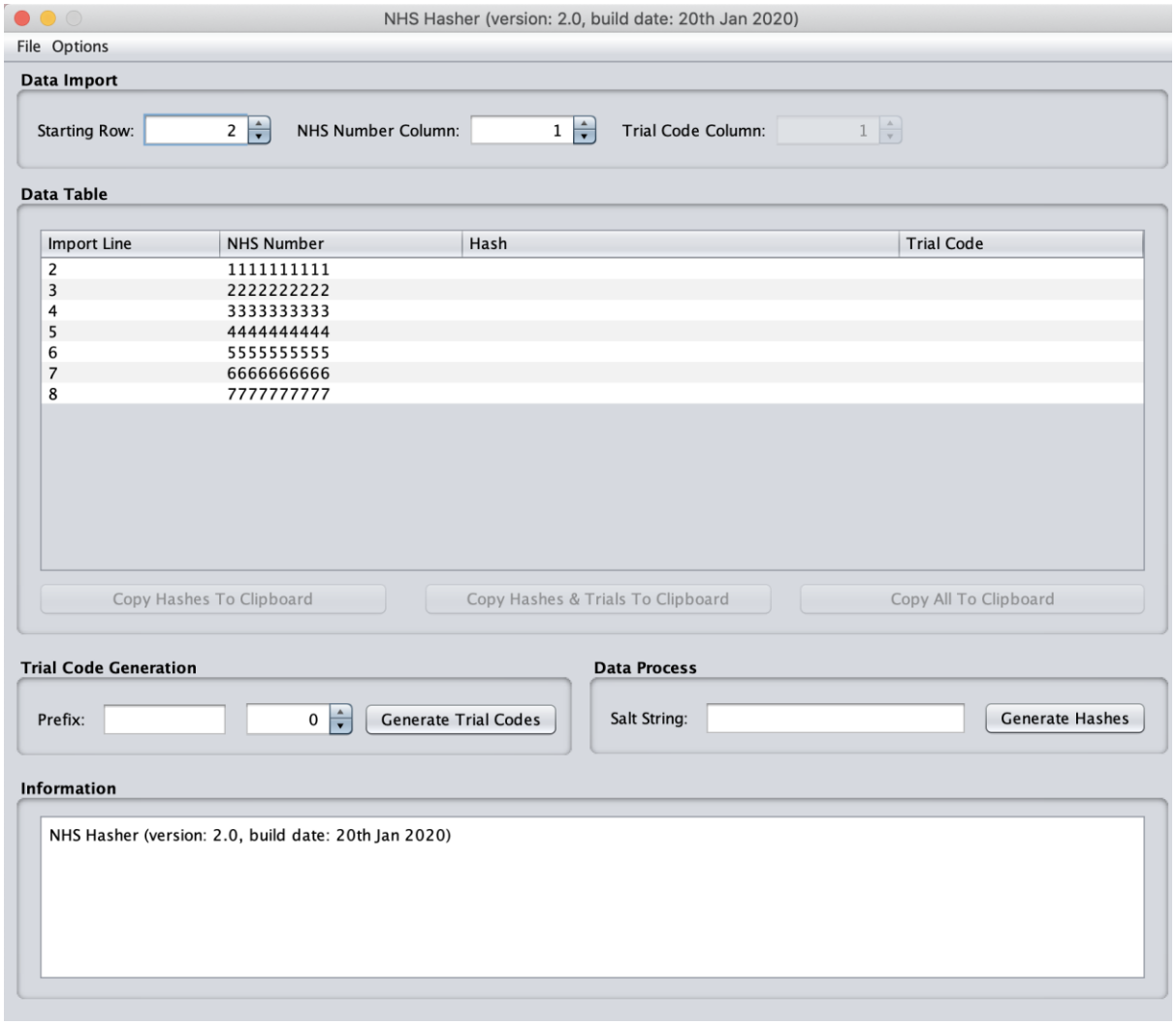
Step 3. Launch the NHS hasher tool. This will be provided in an easy to open format that will not require any special privileges. Simple double-click the "run.bat" file (on Windows) or double-click the "HashingTool.jar" file (on macOS). The program will launch;


NHS Hasher (version: 2.0, build date: 20th Jan 2020)
File Options
Data Import
Starting Row: 1
NHS Number Column: 1
Trial Code Column: 1
Data Table
Import Line
NHS Number
Hash
Trial Code
Copy Hashes To Clipboard
Copy Hashes & Trials To Clipboard
Copy All To Clipboard
Trial Code Generation
Prefix:
0
Generate Trial Codes
Data Process
Salt String:
Generate Hashes
Information
NHS Hasher (version: 2.0, build date: 20th Jan 2020)

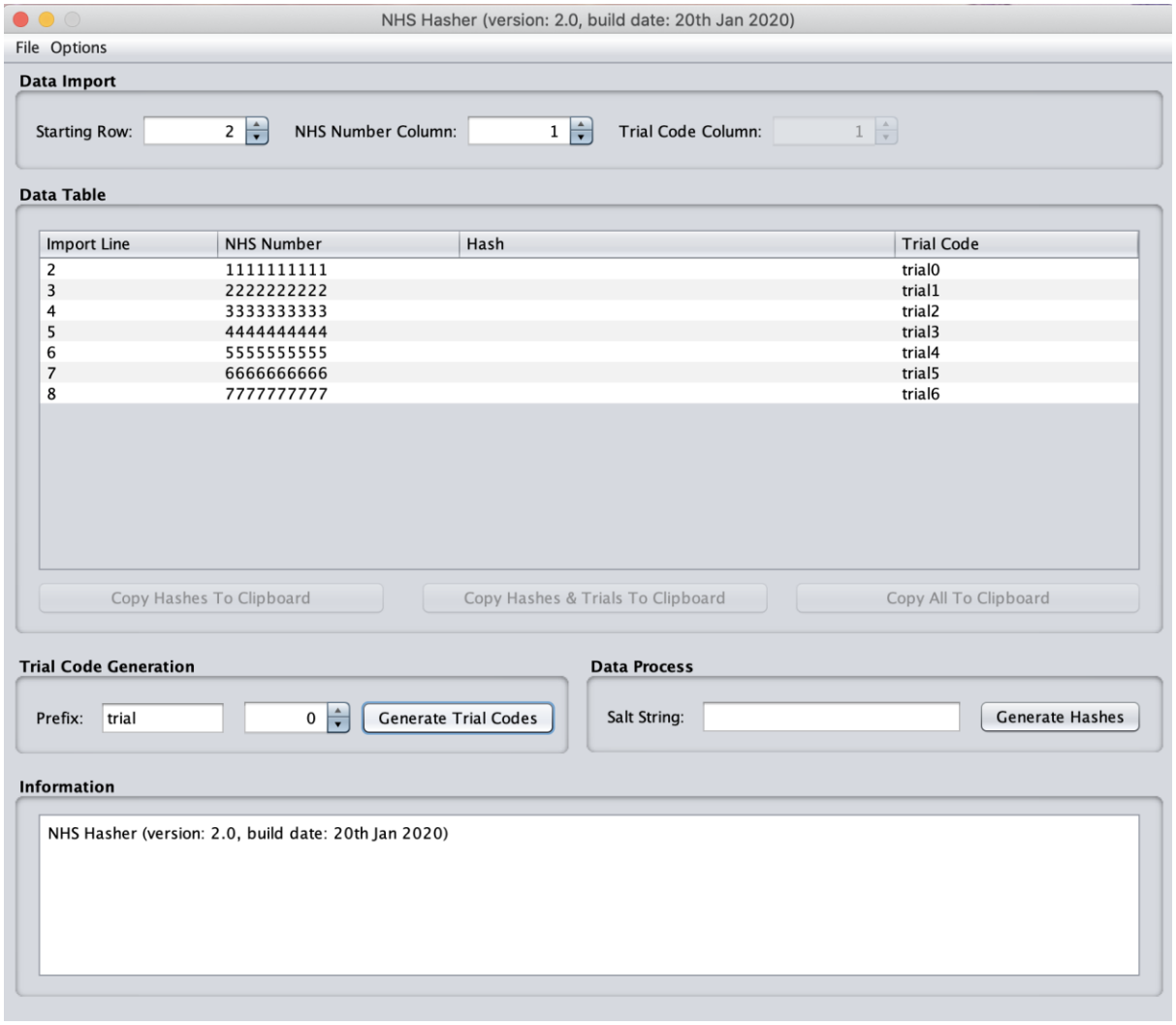
Step 4. It is necessary to change the default options. Click “Options” and make sure that “Enforce Secure Salts” and “Check NHS Numbers” are both turned off. The default hashing algorithm should be set to “SHA-256”. Do not change this;



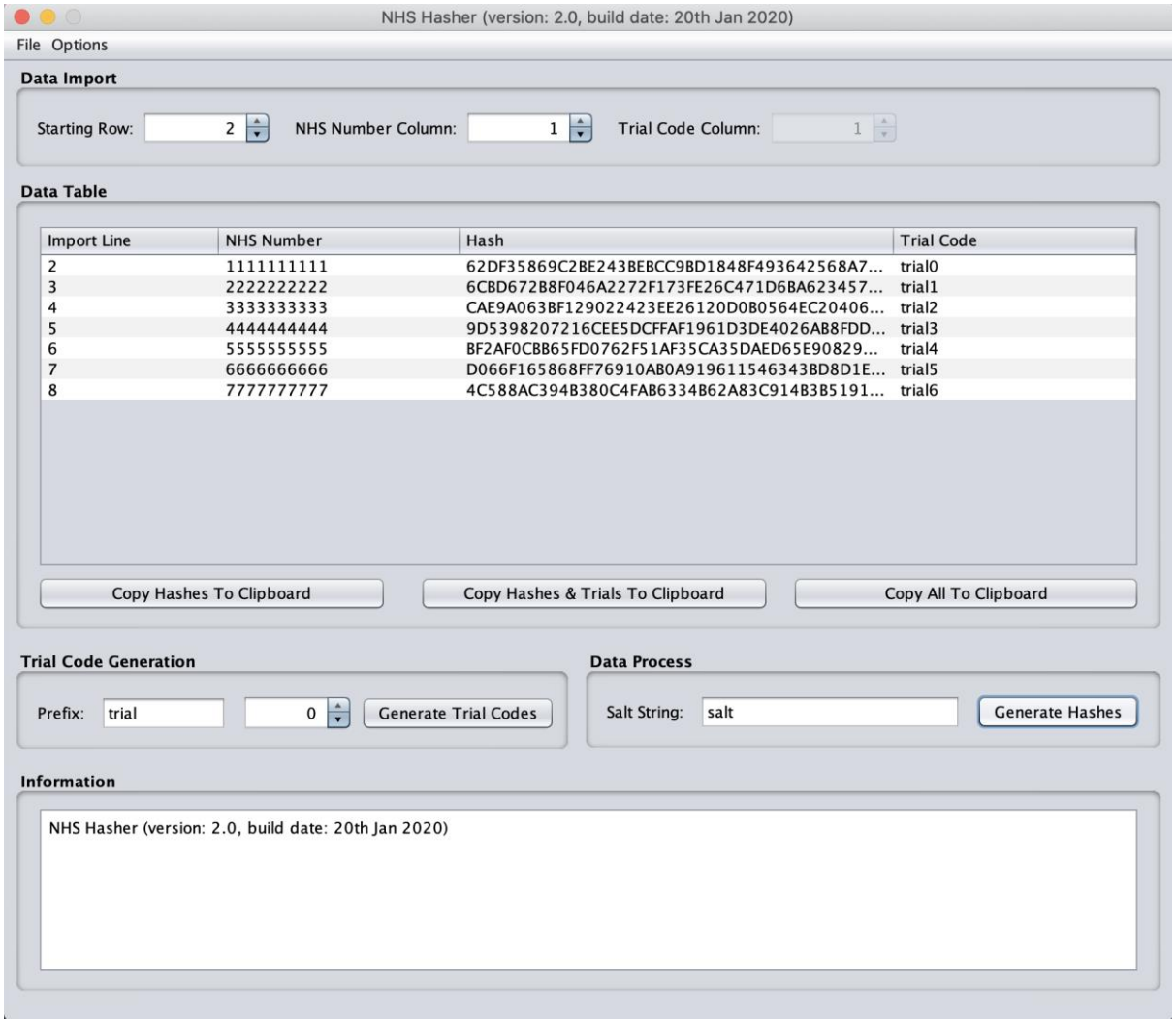
Step 5. First you need to set the “Starting Row” in the “Data Import” section of the program. This is the row number where the first NHS/CHI number is located. In the example given in Step 1 this is row 2. Once you have set this, select “File” from the menu and choose “Import CSV File”. This will give you a file selection window where you can choose a file to import into the program. Please choose the CSV file you just saved from Excel. In the example file given in Step 1 this would look as follows;



Step 6. The NHS Hashing tool was designed to be used with research trials, therefore it expects a "trial code" to be generated against every NHS number before hashing can proceed. This trial code is used as a way to link back to the original NHS numbers should they ever be required. You can simply enter "trial" or something similar in the "Prefix" box, and then click "Generate Trial Codes". The result will look like this;



Step 7. The next step is to generate the hashed values for the NHS numbers. You will be given a "Salt" to enter into the "Salt String" box. This must be copied exactly as given, otherwise the hashes will be wrong. You can use the "Ctrl-V" command in Windows to paste the salt into this box. You then click "Generate Hashes" and the program will create the hashed NHS numbers for you;



Step 8. The final step is to select “File” and choose “Export CSV File”. Select a suitable name, and then save the output file somewhere. As a confirmation that the file has been saved you can double-click it and view the contents in Excel.

Step 9. You can now use the lookup between the NHS number and the Hashed number to provide the Hashed number in any data transfers between datasets allowing linking on the matching Hashed numbers.