Unique Protocol ID: 174052

Brief Title: National Institute for Health Research (NIHR) Health Informatics Collaborative Troponin Study (NHIC-Troponin)

Official Title: National Institute for Health Research (NIHR) Health Informatics Collaborative Cardiovascular Project

DATA ACQUISITION AND ANALYSIS PLAN

Author: Amit Kaura
Approvers: Jamil Mayet / Darrel Francis
Date: 07/03/2017
Updated: 04/07/2024

Introduction

This document outlines the specifications and procedures for the NIHR Health Informatics Collaborative (NIHR HIC) Cardiovascular research database and defines the processes for the collection of the NIHR HIC Cardiovascular research data and onward sharing with researchers. The central research database is held within Imperial College Healthcare NHS Trust (ICHNT).
Data are collected locally at the following trusts data for all patients receiving a troponin test and submitted pseudonymously to ICHNT:

- Imperial College Healthcare NHS Trust
- University College London Hospitals NHS Foundation Trust
- Oxford University Hospitals NHS Foundation Trust
- Kings College Hospital NHS Foundation Trust
- Guys and St Thomas' Hospital NHS Foundation Trust

This document covers the processes for local collection of data at all sites. The main database and systems are hosted at ICHNT as the lead organisation for the Cardiovascular Theme.

NIHR HIC Cardiovascular Database Definitions

Local data store

Each trust will have a local store of NIHR HIC Cardiovascular data; this collects their information in an identifiable form from clinical systems, the data will then be de-identified within the local NHS Trust. These de-identified data will be passed to the central research database.

Research database

The research database will contain only de-identified information. This database combines the data from each site (including ICHNT) and will be used for the TROP-RISK study.

Procedures for local data collection into local stores

Data will be collected automatically from primary clinical systems within each Trust. NHS staff will enter data into clinical systems during routine clinical care of patients, or data will be generated as a result of clinical tests. Data in the primary clinical systems will be processed in accordance with each Trusts clinical guidelines and subject to local quality and governance procedures. Data will be extracted automatically and validated processes for data extraction, transformation and loading (ETL) have been designed to import the data into the local secure data stores.

Access to Local data stores

Access to identified data will be limited to those justified and approved by the local information governance teams and will always be NHS staff in accordance with the duty of confidentiality required by law. Anonymisation procedures will be automated after implementation.

Local data stores will be the only areas that hold any patient identifiers. De-identification is completed at this stage prior to passing data into any research databases and datasets.

De-identification

The data will be pseudonymised locally within each Trust; anonymisation processes will be automated and set up by NHS staff in accordance with:

- advice from local information governance procedures
- the HIC Standard Operating Procedure (SOP) for data sharing and anonymisation
- the Clinical data transfer policy

The data items summarised in Table 1 will be anonymised. Anonymisation will be approved locally by information governance teams before data are sent externally to ICHNT. De-identified data are then shared in accordance with the overarching data sharing agreement.

Each site will hold two versions of the database, one identifiable and one with de-identified, pseudonymised data. The de-identified version is for use in research and shared with the central research database at ICHNT. The identifiable database is held so that if necessary patients can be re-identified if it is of importance to re-contact the patient via their care team (please refer to ‘re-contacting patients’ section.

NHS numbers and hospital numbers are pseudonymised using locally approved procedures. Names are removed from the dataset and date of birth is transformed to year of birth. Date of death is shared, however is converted in to relevant survival rates on provision of data to researchers. Researchers will never see the full date of death or be able to calculate it from other information (all dates are provided as delta for first troponin). The provision of date of death has been agreed by each of the information governance offices at each of the sites in the following de-identification and anonymisation protocol for the TROP-RISK STUDY:

- The exact date of death is required to evaluate mortality after diagnosis.
- Patients presenting with acute coronary syndromes have a high frequency of cardiac events and a high short-term mortality rate. An accurate measure of death is therefore required to fully evaluate this.
- To redact the date of death to year only would misrepresent the survival of these patients, particularly for those who survive for less than one year. The clinical leads at our BRCs have underlined the importance of having the date in full for the TROP-RISK study.

Demographic	Anonymisation
Local Identifier	Provided if NHS number is missing
NHS Number	Use local pseudonymisation algorithm (key to be retained at source) Rename field to subject ID
Family Name	To be removed – LOCAL use
Given Name	To be removed – LOCAL use
Date of Birth	YYYY
Date of Death	DD-MM-YYYY

Table 1. Anonymisation of data elements

Data validity and quality

Prior to pseudonymisation within the clinical systems, NHS number, Date of Birth and patient names will be automatically checked to remove duplication of patients. Samples of data will be clinically validated by members of the clinical team to ensure that the transformation process is correct and data are attributed to the correct patients prior to pseudonymisation. After clinical validation is complete, the data will be transformed to a standardised XML format and validated (Figure 1).

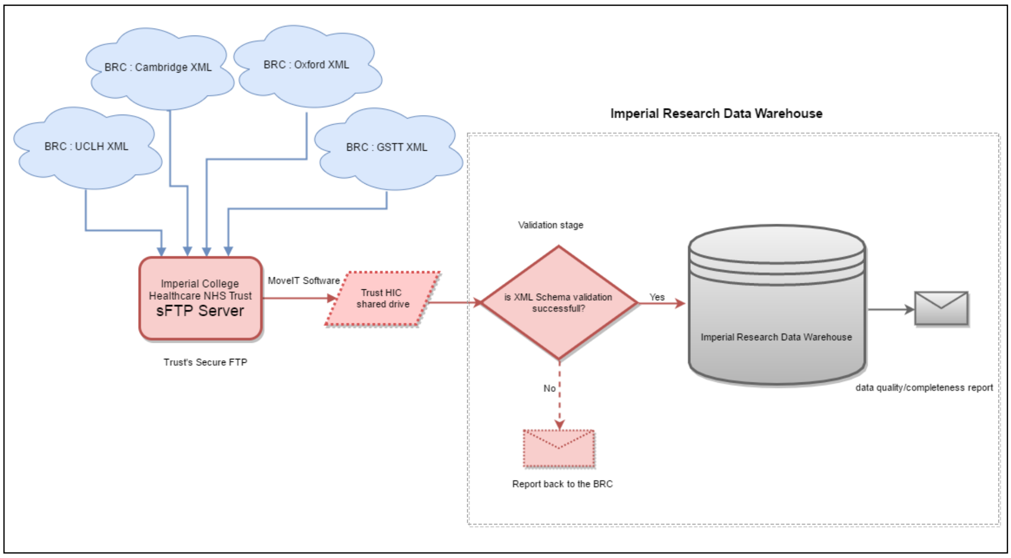
BRC : Cambridge XML
BRC : Oxford XML
BRC : UCLH XML
BRC : GSTT XML
Imperial Research Data Warehouse
Imperial College Healthcare NHS Trust sFTP Server
Trust's Secure FTP
MoveIT Software
Trust HIC shared drive
Validation stage
is XML Schema validation successfull?
Yes
No
Imperial Research Data Warehouse
data quality/completeness report
Report back to the BRC

Figure 1. Data import and validation process.

Data sharing between Trusts

Data are shared in accordance with the SOP for data sharing, as outlined in Section 1.3.2. Data are encrypted in transit via sFTP on the N3 network (Figure 1). The sFTP is set up and hosted by ICHNT.

Procedures for secure research database

Data validity and quality

ICHNT validate the data to ensure that the structure, data items, units and data types are in accordance with the standardised data model. Data will be rejected if validation fails. On rejection, the files will be archived and the data manager will contact the data provider to review the submission and resend once corrected.
Once imported into the secure research database, data are subject to clinical validation by Cardiovascular clinical experts; these validations will be completed by the clinical researchers using de-identified data. Data will be reviewed for data completeness, spread and actual data point values. If data appears invalid it will be rejected.
Data quality reports will be generated and provided to the research team and local data provider after each submission. These will be reviewed after each submission to ensure all areas are populated.

Research Database software

The database is built using Microsoft SQL server 2014, Microsoft's principle database management system software. The installation of the software was carried out by certified technical consultants and tested by the trust ICT team and data warehouse team in accordance with Trust ICT procedures and policies.

Database management

The database is fully backed up on a daily basis. The back-ups are standardised for all Trust databases within the Trust data warehouse. Backups can be restored at any point by warehouse staff. Data are secondary copies from clinical systems; at any point the participating Trusts can re-extract the data from primary sources.

Each site will submit data on a quarterly basis to the database. Data integrity checks will be completed to ensure correct structure is maintained and duplicates are not present.

Once entered on the system, data will not be changed. All access will be 'read only' except via exception, approved by the research Informatics Programme Manager and Clinical Leads group.

The database will be managed by the data manager (Ben Glampson) and developer (Abdul Mulla). Database changes are controlled by the research informatics Programme manager (Ben Glampson), and are sanctioned by the NIHR HIC Cardiovascular scientific steering committee (Chaired by Jamil Mayet). All staff are substantive NHS employees at ICHNT.

Data extracts taken for research will be stored within the data warehouse and retained for the period specified in the data request. All information pertinent to the request will be retained and tracked by the data manager.

Data access for research

TROP-RISK STUDY dataset

All analyses for the TROP-RISK STRUDY will be completed on fully de-identified data. This includes further de-identification to remove dates.

A designated clinical researcher (Vas) will freeze a copy of the database on 1st April 2017, so a static dataset can be used for analysis. This will be retained separately from the live database to allow reproducible analyses.

The study dataset will comprise all patients who had a troponin measured at each of the five academic centres between 2010 (2008 for University College Hospital) and 2017.

Dates will be converted to delta dates, with date zero being the date of the first troponin test. All further dates are provided as number of days from date zero. Age will be provided in years, at the time of the first troponin test.

Date of death will be converted to the number of days since date zero. All patients will be retrospectively followed up, using routinely collected data on the NHS Spine Application, Summary Care Record, until death or censoring on 1st April 2017.

Data elements

The database model will include 156 data points, grouped into demographics, emergency department attendance and inpatient episodes, biochemistry, diagnosis, angiography, revascularization, echocardiography and mortality. Diagnostic data will be based on International Statistical Classification of Diseases and Related Health Problems (ICD) discharge codes.

Outcomes

Primary outcome

The primary outcome will be all-cause mortality. The nature if the data sources means that this is the outcome that will be available and it will be available with high fidelity.

Secondary outcomes

The secondary outcomes will be:

- Acute coronary syndrome (ACS) diagnosis (unstable angina, non-ST segment elevation myocardial infarction (NSTEMI), ST-segment elevation myocardial infarction (STEMI)
- Non-ACS diagnosis
- Angiography
- Revascularisation (coronary artery bypass grafting CABG), percutaneous coronary intervention (PCI))

Statistical Methods

Baseline data

Baseline and demographic characteristics will be summarised by standard descriptive summaries:

- means (standard deviation) for continuous variables which are normally distributed
- median (interquartile range) for continuous variables which are not normally distributed
- number (percentage) for categorical variables

Age distribution

The age distribution of patients who have a troponin measured will be compared with that of the general population using the World Health Organisation (WHO) population estimates for the UK.

Troponin data

Assay performance

Each Trust measures either troponin I or T using either contemporary or high-sensitivity assays. These tests yield results in different measurable ranges with unique cut-off points for the 99th percentile of the upper limit of normal (ULN). A summary table of the performance of each assay will be assembled.

Handling of troponin data

The study dataset will comprise all patients who have had a troponin measured at each of the five academic centres between 2010 (2008 for University College Hospital) and 1st April 2017.

In clinical practice, troponin levels are frequently dichotomised into "positive" (meaning above ULN) or "negative". Troponin levels may have a progressive relationship with prognosis, too, but the shape of this relationship is not known across the full spectrum of values and making the assumption of a linear relationship of mortality with troponin (or log troponin) may not be secure.

For these reasons, we will treat the data in two ways.

1. We will dichotomise the peak troponin level as being either positive or negative based on the ULN for each troponin assay. This makes no assumption of the shape of the relationship.

2. To standardise the many troponin assays, we will scale the results using the ratio of the observed troponin value divided by the ULN for that particular troponin assay. For example, a patient with a troponin value of 96 using an assay which has an ULN of 40 would have a scaled result of 96/40 = 2.4 xULN.

   In some hospitals and some assays, troponin values above a certain level are systematically reported as "more than X". According to what the hospitals have reported so far, the threshold at which the system gives results like this is in all cases very high. In no case does his truncation of this results occur below 10,000 xULN. We therefore plan to set the uppermost band to be ">10,000 x ULN". This ensures that the unavailability of an explicit value does not affect the troponin band into which any result is allocated.

   Similarly, no hospital has reported truncating lower values of troponin at a threshold that is higher than ½ xULN. Therefore, we plan for the lowermost band to be "<½ xULN".

   Troponin has a marked positive skew. Therefore, we plan for the bands to be spaced not arithmetically but exponentially, with narrow spacing nearer low values, where there is a higher frequency density of troponin values.

   Adopting this approach for analysis has the following advantages:

   1. It allows the shape of the relationship to be explored without making assumptions about the shape in advance.
   2. It allows values documented in the hospital databases as "less than X" or "more than X" to be used correctly in the analysis rather than omitted or have imputed values.
   3. The resulting display of the mortality relationship should be easily interpreted and applied by clinicians.

All analyses on troponin will be performed using the peak troponin level. For patients who have a single troponin measurement, the peak troponin will be based on this measurement. In the remainder of the patients who have more than one troponin test in the same hospital episode, the peak troponin value will be defined as the highest of all measurements.

Relationship between troponin positivity, age and mortality

To assess the association of dichotomous troponin level (above ULN or not) with all-cause mortality, Cox proportional hazards regression modelling will be used. Hazard ratios will be determined for the following age groups: 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and 90+ years. We will also stratify the hazard ratios for troponin positive versus negative groups across these age groups by troponin assay (contemporary versus high-sensitivity troponin assay).
The Kaplan-Meier method will be used to calculate and display cumulative mortality, with positive and negative troponin bands compared using the log-rank statistic.

Relationship between troponin level and mortality

The main analysis of the association of progressively higher troponins with all-cause mortality will be performed without making assumptions about the shape of the relationship, i.e. with Cox proportional hazards regression modelling.
The Kaplan-Meier method will be used to calculate and display cumulative mortality, with troponin bands compared using the log-rank statistic. The results will be stratified according to ACS diagnosis and revascularisation status.

ACS diagnosis

The ICD-10 codes in Table 2 will be used to indicate an ACS diagnosis.

ICD-10 Code	Category
I20.0	Unstable angina
I21.0	Acute transmural myocardial infarction of anterior wall
I21.1	Acute transmural myocardial infarction of inferior wall
I21.2	Acute transmural myocardial infarction of other sites
I21.3	Acute transmural myocardial infarction of unspecified site
I21.4	Acute subendocardial myocardial infarction
I21.9	Acute myocardial infarction, unspecified
I22.0	Subsequent myocardial infarction of anterior wall
I22.1	Subsequent myocardial infarction of inferior wall
I22.8	Subsequent myocardial infarction of other sites
I22.9	Subsequent myocardial infarction of unspecified site
I24.8	Other forms of acute ischaemic heart disease
I24.9	Acute ischaemic heart disease, unspecified

Table 2. ICD-10 codes used to indicate an ACS diagnosis

Revascularisation status

Acute revascularisation will be defined as having PCI or CABG in the time window between 48 hours before and 3 months after the first troponin measurement. This will account for patients who had revascularisation, in particular PCI, as an emergency prior to their first troponin blood test and to capture revascularisation, in particular CABG, performed as an outpatient following their index admission.

Statistical package

Statistical analyses will be performed using SPSS software version 24 (SPSS Inc., Chicago, Illinois, United States) and the results verified by replicate analysis using R 3.3.2 statistical package (the R Core Team, Vienna, Austria).

Statistical significance

All hypothesis tests will be 2-tailed. A p-value of <0.05 will be considered statistically significant. No correction will be implemented for multiple testing.