BMJ Health & Care Informatics

National Institute for Health Research Health Informatics Collaborative: development of a pipeline to collate electronic clinical data for viral hepatitis research

David Anthony Smith ,[1,2] Tingyan Wang,[2,3] Oliver Freeman,[2] Charles Crichton,[2] Hizni Salih,[2] Philippa Clare Matthews,[3,4] Jim Davies,[2,5] Kinga Anna Várnai,[1,2] Kerrie Woods ,[1,2] Christopher R. Jones,[6] Ben Glampson,[7] Abdulrahim Mulla,[7] Luca Mercuri,[7] A. Torm Shaw ,[8] Lydia N Drumright,[9] Luis Romão,[10,11] David Ramlakan,[10,11] Finola Higgins,[12] Alistair Weir,[12] Eleni Nastouli,[10] Kosh Agarwal,[13] William Gelson,[9] Graham S. Cooke,[6] Eleanor Barnes[1,3]

To cite: Smith DA, Wang T, Freeman O, *et al*. National Institute for Health Research Health Informatics Collaborative: development of a pipeline to collate electronic clinical data for viral hepatitis research. *BMJ Health Care Inform* 2020;**27**:e100145. doi:10.1136/bmjhci-2020-100145

DAS, TW and OF contributed equally.

Received 21 April 2020
Revised 31 August 2020
Accepted 16 September 2020

For numbered affiliations see end of article.

Correspondence to
Proffesor Eleanor Barnes;
ellie.barnes@ndm.ox.ac.uk

ABSTRACT

Objective The National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC) is a programme of infrastructure development across NIHR Biomedical Research Centres. The aim of the NIHR HIC is to improve the quality and availability of routinely collected data for collaborative, cross-centre research. This is demonstrated through research collaborations in selected therapeutic areas, one of which is viral hepatitis.

Design The collaboration in viral hepatitis identified a rich set of datapoints, including information on clinical assessment, antiviral treatment, laboratory test results and health outcomes. Clinical data from different centres were standardised and combined to produce a research-ready dataset; this was used to generate insights regarding disease prevalence and treatment response.

Results A comprehensive database has been developed for potential viral hepatitis research interests, with a corresponding data dictionary for researchers across the centres. An initial cohort of 960 patients with chronic hepatitis B infections and 1404 patients with chronic hepatitis C infections has been collected.

Conclusion For the first time, large prospective cohorts are being formed within National Health Service (NHS) secondary care services that will allow research questions to be rapidly addressed using real-world data. Interactions with industry partners will help to shape future research and will inform patient-stratified clinical practice. An emphasis on NHS-wide systems interoperability, and the increased utilisation of structured data solutions for electronic patient records, is improving access to data for research, service improvement and the reduction of clinical data gaps.

Summary

What is already known?

- Electronic patient records in National Health Service (NHS) trusts contain a wealth of routinely collected clinical data useful for translational research. However, these data are not easily accessible to the individual NHS Trust or researchers.
- There is a shortage of detailed clinical data available for patients with viral hepatitis in the UK, in particular for patients infected with the hepatitis B or E viruses.

What does this paper add?

- We present a comprehensive methodology that has been proposed, implemented and validated by the The National Institute for Health Research Health Informatics Collaborative for the development of a new data collection and management pipeline.
- We show that routinely collected clinical data from patients with hepatitis C, B and E infection can be collated, integrated and made available to researchers automatically from five large NHS trusts.
- We describe the initial data collected from 906 patients infected with hepatitis B and 1404 patients infected with hepatitis C.

INTRODUCTION

The National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC)[1] was established in 2014, in response to a challenge by the UK's Chief Medical Officer Dame Sally Davies to make routinely collected clinical data available for translational research across multiple sites. The NIHR HIC is a programme of infrastructure development across the network of 25 National Health Service (NHS) Trusts, supported by their university partners through the NIHR Biomedical Research Centres. The programme was initially made

up of five NHS Trusts hosting comprehensive NIHR Biomedical Research Centres; Cambridge University Hospitals NHS Foundation Trust, Guy's and St Thomas' NHS Foundation Trust, Imperial College Healthcare NHS Trust, Oxford University Hospitals NHS Foundation Trust, and University College London Hospitals NHS Foundation Trust. The aim of the NIHR HIC programme is to improve the quality and availability of routinely collected clinical data, making it available for cross-centre collaborative, translational research. This presents both opportunities and challenges:

Opportunities

- The UK's unified healthcare system (the NHS) generates millions of clinical datapoints each year, which can be leveraged to improve collection of clinical information, address clinical research questions and improve patient care.
- The automated collection of data from electronic patient record systems can dramatically reduce the time and cost of data collection for research and provide opportunities for collaboration with both academic and industry partners.
- Modern machine learning techniques using neural networks require large datasets to be used effectively,[2] the reuse of routinely collected data can provide a cost effective way of collating these datasets.

Challenges

- All NHS trusts are separate organisations, responsible for the protection of the data of their own patients. To enable data to be shared across these separate organisations for research, a governance framework needed to be established.
- Each NHS trust has its own electronic patient record, and its own set of customisations, extensions and variations in data entry practice. Alongside the primary electronic patient record, each trust will also have an extensive collection of departmental systems, again subject to customisation and variations in practice.
- Data definitions are not all standardised.
- Not all data are collected electronically at all sites.
- Large amounts of important data are stored in free text rather than discrete values.
- Clinical practice can differ between sites.
- Data can be produced and collected differently between sites. (eg, different laboratory methods or platforms used for tests).
- Different trusts have different levels of expertise in clinical informatics.
- Projects such as the NIHR HIC require sustained investment before they start to deliver tangible results.

The NIHR HIC aims to overcome these challenges and demonstrate the value of these data for research in key therapeutic areas; the first five areas considered were viral hepatitis, ovarian cancer, critical care, acute coronary syndromes and renal transplantation. This paper focuses on the viral hepatitis theme, which is led by Oxford University Hospitals NHS Foundation Trust.

Viral hepatitis is a global health problem with an estimated 1.35 million people dying from either end-stage liver disease, hepatocellular carcinoma or other viral hepatitis-related diseases in 2015.[3] The majority of these deaths are as a result of hepatitis B virus (HBV) and hepatitis C virus (HCV) infections; this is greater than tuberculosis, HIV or malaria. Unlike these other infections, the number of viral hepatitis deaths has increased since 1990.[4] International targets arising from the United Nations 'sustainable development goals' have set a challenge for the elimination of viral hepatitis as a public health threat by the year 2030.[5 6] As part of meeting this goal, leveraging existing clinical data are a cost-effective way to answer vital research questions. The NIHR HIC Viral Hepatitis Theme aims to address key research questions (table 1), to demonstrate the utility of the NIHR HIC methodology.

This paper presents a comprehensive methodology that has been proposed, implemented and validated by the NIHR HIC for the development of a new data collection

Table 1 Overview of specific issues to be addressed in viral hepatitis theme using the National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC) Viral Hepatitis Theme Dataset

Virus	Example research questions to be addressed using the NIHR HIC Viral Hepatitis Theme Dataset
Hepatitis B virus (HBV) and hepatitis delta (HDV) coinfection	► Characterising demographics and laboratory parameters for cases of chronic HBV infection. ► Improving use of biomarkers for monitoring and treatment stratification. ► Identifying individuals who control/clear HBV infection. ► Identifying subgroups that develop complications. ► Estimating incidence of HDV coinfection in the UK. ► Identifying determinants of HDV treatment success.
Hepatitis C virus (HCV)	► Establishing sustained viral response rates for HCV therapy in the UK. ► Identifying factors determining treatment response in difficult to treat groups of patients.
Hepatitis E virus (HEV)	► Estimating incidence of HEV infection in the UK. ► Identifying risk factors for HEV infection the UK.

and management pipeline. This development is under a comprehensive governance framework that allows data to be collated across multiple centres for collaborative research on viral hepatitis. Under this governance framework, a data collaboration involves the generation of a research-ready dataset that is broad enough to support a wide range of investigations in a specific clinical area. The dataset is assembled to the same agreed standards at each centre. The data transformations needed to achieve this, starting from patient records, are documented and shared.

METHODOLOGY

Governance framework

The protocol for the collection and management of the data for the viral hepatitis theme has been reviewed and approved by South Central—Oxford C Research Ethics Committee (Reference Number: 15/SC/0523). In addition, an NIHR HIC data sharing framework, covering a wide range of data and research collaborations has been signed by all participating centres. This document, in conjunction with the rest of the governance framework established by the NIHR HIC, addresses common requirements and considerations regarding data sharing between centres, contractual responsibilities, confidentiality, intellectual property and a publications policy. This general agreement will be used to underpin individual agreements for research collaborations with third party academic and industry partners. Any collaboration with industry partners requires additional agreements, with additional governance checks by participating sites. This process has been simplified by the creation of a 'framework industry collaboration agreement' which simplifies the addition of individual participating sites to an industry collaboration. A scientific steering committee has been established which is made up of a representative of each participating site, and reviews and approves data requests from external collaborators.

Development of data model

The development of the NIHR HIC viral hepatitis data model, which outlines the structure of the dataset and the associations between the data fields, began by the clinical leads defining data fields, required to answer the initial academic questions posed by the clinical and scientific leads across centres (table 1). The feasibility of collecting this dataset was tested using the electronic data capture software OpenClinica.[7] Each site was required to manually complete case report forms in OpenClinica for a small subset of their patient cohorts; an assessment of the completeness of these data led to a refinement of the dataset and a second version, which was used to generate the Extensible Markup Language (XML) Schema Definition (XSD) used to define the dataset. This XSD was then further refined to remove errors and to provide a more efficient data structure.

Data architecture for collection and integration of data

Each site within the collaboration provided a data product containing the data items outlined in the agreed dataset. While each site had pre-existing data warehouses and systems for collecting, storing and using patient data, these systems were designed to be used for patient care, administrative and financial purposes, and were not always suitable for generating the required data product. The variety of electronic patient record systems and laboratory information management systems used across the participating sites meant that data were often stored in different formats. This meant that at each site a data architecture outlying the flow of data from electronic patient record systems had to be developed to allow the effective flow of data from operational systems to research systems. This required the development of data warehousing or data management infrastructure at all sites.

To allow the compilation of the full dataset, each site generated a data product that was integrated in the NIHR HIC Viral Hepatitis Central Data Repository (figure 1). In some cases, these data were already structured and could be transferred directly into the data repository. Where data were stored in an unstructured (free text) format, the data had to be either manually entered or extracted (figure 2). The data were then anonymised by removing patient identifiers and a data product was created using the XML format. To avoid duplication of records, each site was responsible for maintaining a link between the patient's local clinical identifiers and the identifier used in the database. The data product was securely transferred to the NIHR HIC Viral Hepatitis Central Data Repository, where the data were put into a database for queries and analysis. As each site has different infrastructure in place to produce the data product, updates are submitted on request and there is no fixed schedule for new data submissions from sites. When a dataset is requested by a research group, the request is first reviewed and approved by the scientific steering committee, and following internal governances process, an extract of the integrated dataset is provided to the research group for their analysis.

NIHR HIC Viral Hepatitis Central Data Repository

Anonymised data from the providing centres were transmitted to the lead centre in XML format via secure email or submitted directly to the NIHR HIC Data Acquisition Management (HICDAM) system via a secure web-based front end, again in XML format. The primary service inside HICDAM is the Message Receiving, Curating, and Understanding Repository (MeRCURy), which performs the validation, processing and storing of the submitted data.

The MeRCURy system supports two types of data validation: basic, automatic integrity checks, which must be satisfied before the data are loaded into the database, and more sophisticated, manual checks of data consistency, which are performed after the data have been loaded. The basic integrity checks involve validation against the agreed XSD, confirming that the data are correctly formatted, together with logical checks on the type or

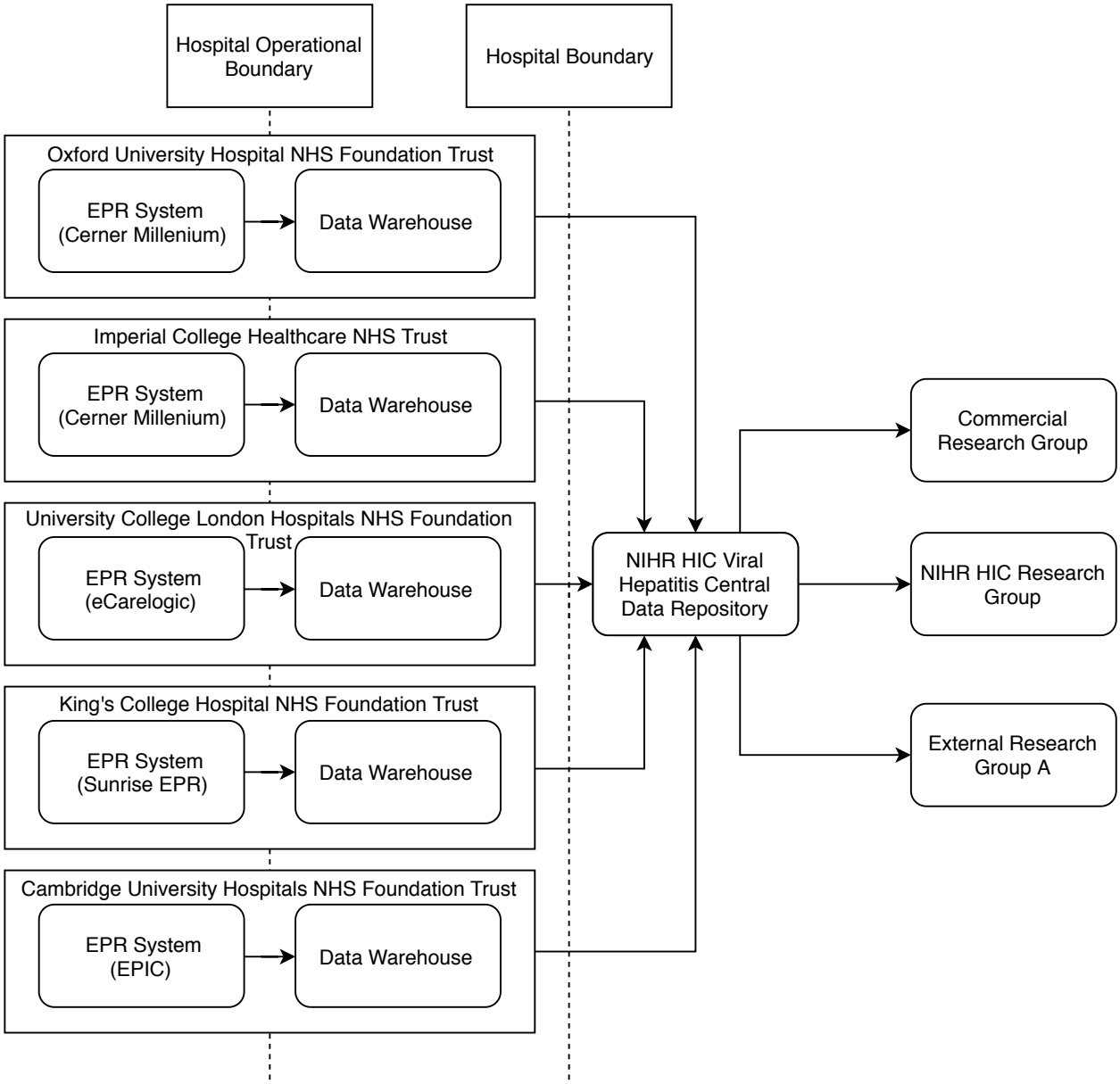

Figure 1 Data flow for the National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC) viral hepatitis clinical exemplar theme. For each participating site, data originating from clinical systems were used to populate a corresponding data warehouse that contains data fields in the agreed dataset for viral hepatitis research. Data in the data warehouse were transferred to the NIHR HIC Viral Hepatitis Central Data Repository. The data stored in the central data repository can be extracted and provided to internal and external research groups according to the governance process. NHS, National Health Service, EPR, Electronic Patient Record.

range of values submitted: for example, a check that any value given for the date of death is strictly later than those given for dates of treatment.

The manual checks reflect working assumptions regarding the relationships between the values of different data items submitted, for example, that a certain combination of treatments would never be used in practice, or where there are clear duplications of data. In each case, some additional information may be needed to determine whether data is incorrect, or whether the assumptions are invalid.

The system will inform data providers of the outcome of any submission. If the data submitted fail a basic integrity check, they are rejected, and a report is generated containing appropriate diagnostic information. If the data are accepted, a confirmation message is sent, and the data may then be reviewed using a secure interface provided by the LabKey[8] application. Data can be then be explored and/or exported from the system in a variety of formats, including .xls, .xlsx, .tsv and .csv.

RESULTS

The NIHR HIC Viral Hepatitis Research Database has been developed and populated with data. Figure 3 provides an overview of the data model of the database

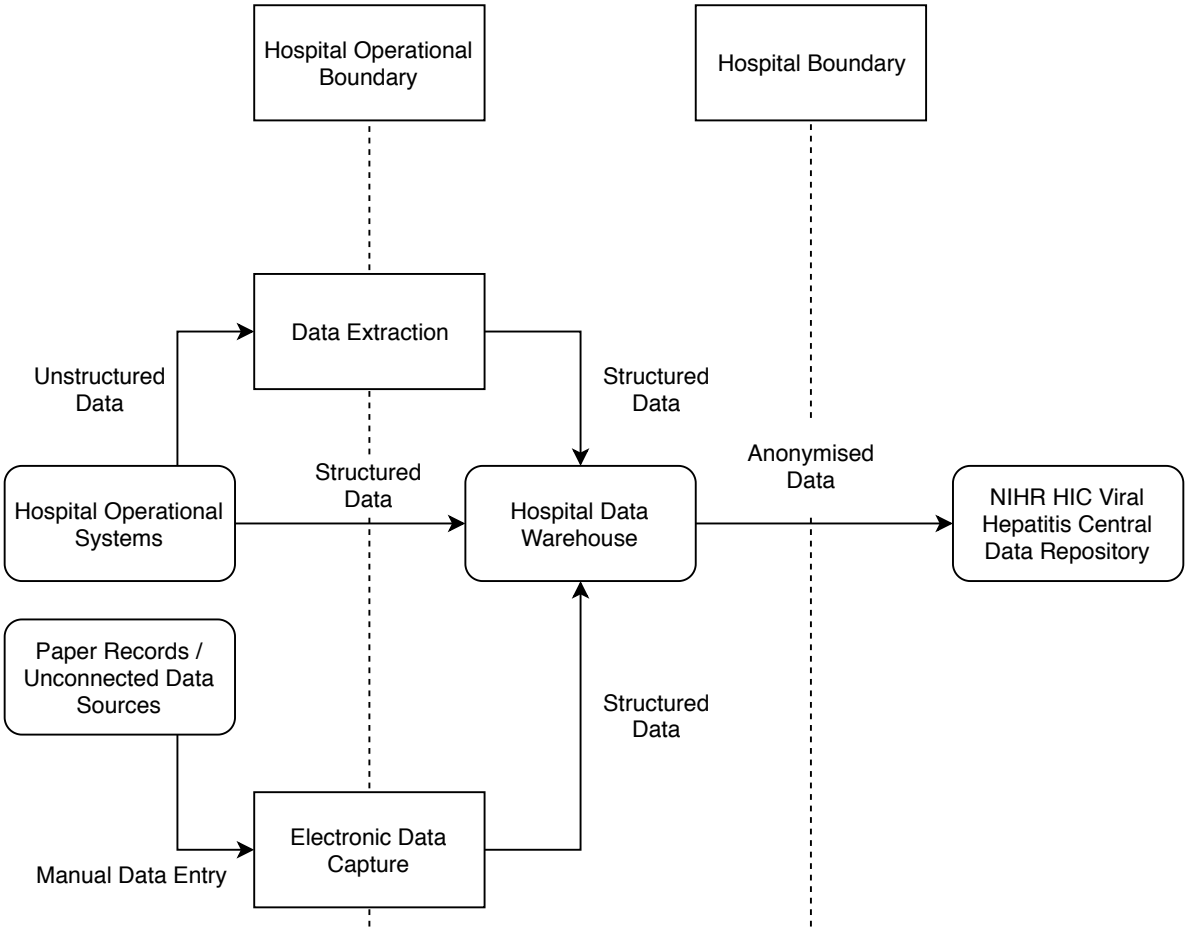

Figure 2 Conceptual data flow for an individual site participating in the National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC) Viral Hepatitis clinical exemplar theme. In an individual site, structured data in the hospital operational systems were directly transferred into the hospital data warehouse, and data stored in unstructured format was automatically or manually transformed to produce structured data either before or after transfer to the data warehouse. In addition, data from paper records or unconnected data sources were manually entered into a structured electronic data capture system and transferred into the data warehouse. Data was then anonymised prior to transfer to the central data repository for viral hepatitis research.

and the relationships among different entities. The data collected falls into several categories: (a) basic information (eg, demographics, hospital visits, death, discharge and study sites), (b) laboratory data, (c) treatments, (d) diagnoses, (e) liver conditions, (f) hepatitis virus genotypes and (g) other clinical information.

The database contains 32 tables for storing collected datapoints, and several tables for data field definitions. The final database consists of 349 data fields, split into 20 different element types. There are 203 data fields that are common to HBV, HDV, HCV and HEV. The remaining data fields are specific to a type of viral hepatitis (HBV and HDV (n=75), HCV (n=47) or HEV (n=24)). Due to the differences between sites, no datapoints were made mandatory. The current dataset was submitted between 7 December 2018 and 14 June 2019. The database contains 3494 patients with associated clinical information. There are 842 676 records regarding laboratory tests, 2824 records regarding imaging data and 8514 records for medications (including antiviral therapy and others), with data on comorbidities, diagnoses, genotype information, liver conditions, treatment side effects and social behaviour also included.

Chronic HBV cohort description

Data have been extracted from the NIHR HIC Viral Hepatitis Research Database by using HBV lab test information. Patients included in the extract meet at least one of the following criteria:

a. Two positive hepatitis B surface antigen tests, at least 6 months apart.
b. Positive hepatitis B surface antigen and positive HBV DNA, at least 6 months apart.

The characteristics of the chronic HBV cohort of 960 patients are described in table 2, including age, gender, race, HCV coinfection, HIV coinfection, comorbidities, the severity of liver disease and treatment information. Among these patients, 254 patients (26.5%) have previously received medications or are currently under treatment. There are 17 patients coinfected with HDV, with data extracted based on detectable HDV viremia or anti-HDV antibody. However, there were 475 patients who

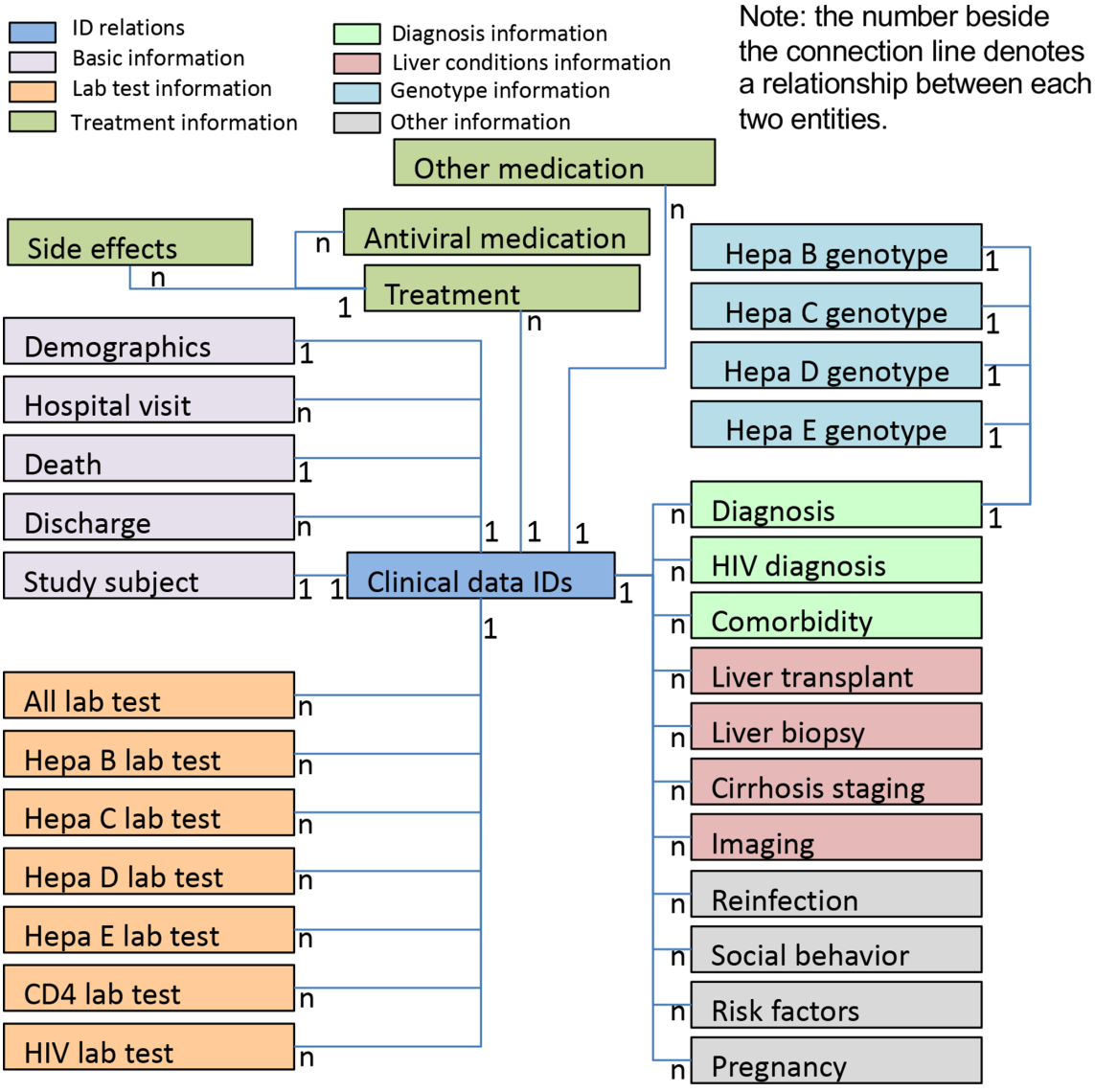

Figure 3 Data model overview of the National Institute for Health Research Health Informatics Collaborative Viral Hepatitis Research Database. The defined dataset for viral hepatitis research includes 32 entities/tables, which can be grouped into different categories, marked in different colours. Each patient has a unique clinical data identifier that is stored in the table 'clinical data IDs'. The relationship between two entities is shown beside their connection line. For example, 1:1 between table 'clinical data IDs' and table 'study subject' indicates that a patient can only have one study subject ID; while 1:N between table 'clinical data IDs' and table 'hospital visit' indicates that a patient could possibly have multiple hospital visits.

had no HDV test data. Missing data for each variable is reported as 'unknown' in table 2.

Chronic HCV cohort description

Data have been extracted for patients with a positive HCV RNA test or HCV genotype from the NIHR HIC Viral Hepatitis Research Database. The characteristics of HCV cohort of 1404 patients are described in table 3, including age, gender, ethnicity, HCV genotype, HBV coinfection, HIV coinfection, comorbidities, the severity of liver disease and treatment information. Among these patients, 914 patients (65.1%) had received interferon-free direct-acting antiviral treatment by February 2018. Similarly, to the HBV cohort, there were various degrees of missing data for each variable, which is reported as 'unknown' in table 3.

HEV cohort description

Data have been extracted from the NIHR HIC Viral Hepatitis Research Database by using HEV IgM and HEV IgG lab test information. HEV usually causes an acute and self-limiting infection but can occasionally cause severe disease and/or chronic infection; our data collection approach sets out to include both categories. Patients included in the extract meet at least one of the following criteria:

a. Patients with anti-HEV IgM and anti-HEV IgG were both positive.
b. patients with HEV RNA detected in serum and/or stool.

There are 14 patients with acute HEV infection in our current database. Among them, 4 patients are male, and

Table 2 Characteristics of chronically infected hepatitis B virus (HBV) cohort

Number of patients (total)	960
Age at the beginning of follow-up: median (range)	36 (2–80)
Gender (%)	
Male	502 (52.3)
Female	389 (40.5)
Unknown	69 (7.2)
Ethnicity (%)	
White	106 (11.0)
Mixed	11 (1.2)
Asian	220 (22.9)
Black	161 (16.8)
Other	72 (7.5)
Unknown	390 (40.6)
HDV coinfection	
Yes (%)	17 (1.8)
No (%)	468 (48.7)
Unknown (%)	475 (49.5)
HCV coinfection	
Yes (%)	75 (7.8)
No (%)	347 (36.1)
Unknown (%)	538 (56.1)
HIV coinfection	
Yes (%)	7 (0.7)
No (%)	427 (44.5)
Unknown (%)	526 (54.8)
Documented comorbidities	
Diabetes (%)	30 (3.1)
Depression (%)	28 (2.9)
Chronic kidney disease (%)	16 (1.7)
Coagulation disorder (%)	22 (2.3)
Cryoglobulinaemia (%)	3 (0.3)
Cancer (%)	21 (2.2)
Liver cancer (%)	1 (0.1)
Other cancer (%)	20 (2.1)
Severity of liver disease	
Cirrhosis (%)	27 (2.8)
Decompensated (%)	4 (0.4)
Child–Pugh Score: median (IQR)	NA
Fibroscan stiffness: median (IQR)	5.3 (2.4)
MELD Score: median (IQR)	NA
Patients treated (%)*	254 (26.5)
Lab tests at baseline	
Alanine aminotransferase (IU/L)	
Untreated group: median (IQR)	29 (25)
Treated group: median (IQR)	39 (37)
HBV DNA (log_{10} IU/mL)	
Untreated group: median (IQR)	3.15 (1.58)
Treated group: median (IQR)	4.26 (3.67)
Hepatitis B e antigen positive	
Untreated group (%)	82/706 (11.6)
Treated group (%)	84/254 (33.1)

For a continuous variable, median and the IQR are calculated, for a categorical variable, the number and the percentage of patients is provided.
*Patients who had at least one episode of treatment recorded.
MELD, Model For End-Stage Liver Disease; NA, not available.

gender information was missing for the other 10. The median age at the time of patients first positive HEV test was 61.5 years (range: 21–86).

DISCUSSION

This paper describes the methodology for the NIHR HIC informatics infrastructure (pipeline) development for collating data across multiple sites, and presents initial cohorts created as part of the NIHR HIC Viral Hepatitis Theme. This demonstrates that routinely collected patient data can be aggregated across multiple centres to create datasets for research. Data collected for this collaboration from one site have already been used in an analysis of hepatitis B surface antigen loss[9] and an analysis of data from all sites is currently underway. Further internal analysis is planned and collaborations with industry partners have been established to address specific translational research questions.

While the NIHR HIC Viral Hepatitis Theme has been able to demonstrate that routinely collected patient data can be aggregated across multiple centres to create datasets for research, challenges still remain. For example, data submission completeness differs across sites, as the original data are stored differently and may be, therefore, easier to process at each site. Large amounts of imaging report and biopsy report data remain embedded in free text, which may contain patient identifiers, meaning it cannot be transmitted to the central site for processing, and each individual site has to develop free text anonymisation protocols or perform manual extraction of this data. In addition, as the data are primarily collected for clinical care it is subject to differences in clinical practise between clinicians and sites. These issues can lead to heterogenous patterns of missing data between sites. Missing data are, therefore, clearly flagged to researchers and they are strongly encouraged to investigate and account for patterns of missing data in any analysis performed. With these challenges in mind, optimisation of the data model will be continued, and a comprehensive data dictionary continues to be updated accordingly for researchers across the participating centres and external collaborators. In addition, natural language processing

Table 3 Characteristics of hepatitis C virus (HCV) cohort

Number of patients (total)	1404
Age at the beginning of follow-up: median (range)	47 (16–87)
Gender (%)	
Male	604 (43.0)
Female	281 (20.0)
Unknown	519 (37.0)
Ethnicity (%)	
White	359 (25.6)
Mixed	11 (0.8)
Asian	53 (3.8)
Black	41 (2.9)
Other	49 (3.5)
Unknown	891 (63.4)
HCV genotype	
1a (%)	351 (25.0)
1b (%)	144 (10.3)
1 other (%)	48 (3.4)
2 (%)	58 (4.1)
3a (%)	225 (16.0)
3 other (%)	50 (3.6)
4 (%)	69 (4.9)
5 (%)	2 (0.1)
6 (%)	3 (0.2)
Unknown (%)	454 (32.3)
HBV coinfection	
Yes (%)	75 (5.4)
No (%)	1097 (78.1)
Unknown (%)	232 (16.5)
HIV coinfection	
Yes (%)	60 (4.3)
No (%)	1077 (76.7)
Unknown (%)	267 (19.0)
Documented comorbidities	
Diabetes (%)	111 (7.9)
Depression (%)	253 (18.0)
Chronic kidney disease (%)	18 (1.3)
Coagulation disorder (%)	47 (3.3)
Cryoglobulinaemia (%)	7 (0.5)
Cancer (%)	77 (5.4)
Liver cancer (%)	16 (1.1)
Other cancer (%)	61 (4.3)
Severity of liver disease	
Cirrhosis (%)	408 (29.1)
Decompensated (%)	121 (8.6)
Child–Pugh Score: median (IQR)	6.0 (2.0)
Fibroscan stiffness: median (IQR)	8.4 (11.6)
MELD Score: median (IQR)	9.0 (4.7)
Patients treated (%)*	914 (65.1)

For a continuous variable, median and the IQR are calculated, for a categorical variable, the number and the percentage of patients is provided.
*Patients who had at least one episode of treatment recorded.
MELD, Model For End-Stage Liver Disease.

algorithms for automatically extracting data and information from free text examination reports and clinical notes will be embedded into the data collating process, to eliminate the requirement for manual extraction and reduce amounts of missing data.

Through the new pipeline, electronic clinical data collected for the routine care of individuals with hepatitis B, C, D, and E infection have been collated over the last 3 years, across five NHS trusts, and reused for viral hepatitis research. The datasets created can be requested by internal and external research groups for analysis; these requests are reviewed and approved according to the agreed governance procedures. Data collection for the NIHR HIC Viral Hepatitis Theme is an ongoing process, and other NHS trusts that have signed up to the NIHR HIC governance framework have been invited to join the collaboration. Collection of large datasets on viral hepatitis via the NIHR HIC programme is not only a cost-effective method of data collection but also allows novel analyses to be performed, giving further insight into viral hepatitis in the UK.

Author affiliations
[1]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[2]NIHR Oxford Biomedical Research Centre, Big Data Institute, University of Oxford, Oxford, UK
[3]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[4]Deptartment of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[5]Department of Computer Science, University of Oxford, Oxford, UK
[6]Department of Infectious Disease, St Mary's Campus, Imperial College London, London, UK
[7]Research Informatics Team, Imperial College Healthcare NHS Trust, London, UK
[8]Clinical Trials Centre, Winston Churchill Wing, St Mary's Campus, Imperial College London, London, UK
[9]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[10]National Institute for Health Research Biomedical Research Centre, University College London Hospitals, London, UK
[11]Institute of Health Informatics, University College London, London, UK
[12]Guy's and Saint Thomas' Hospitals NHS Trust, London, UK
[13]Institute of Liver Studies, King's College London, London, UK

Acknowledgements The authors would like to thank all the research nurses and research admin staff at the contributing sites for their help in developing this project and Theresa Noble for her work supporting the National Institute for Health Research Health Informatics Collaborative.

Contributors DAS, TW and OF contributed equally to this work. DAS, TW and OF performed the data analysis and wrote the manuscript and were supervised by EB, PCM and JD. All authors were significantly involved in the development of the National Institute for Health Research Health Informatics Collaborative Infrastructure and data collection. Clinical leadership was provided by EN, AK GW, GC and EB.

Funding This research has supported by National Institute for Health Research (NIHR) Biomedical Research Centres at Cambridge, Guy's and St Thomas', Imperial, Oxford and University College London Hospitals and funded by the NIHR Health Informatics Collaborative. GC is an NIHR research professor, EB is an NIHR senior investigator. PCM is funded by the Wellcome Trust and holds an NIHR Senior Fellowship award.

Competing interests LND reports grants from National Institute for Health Research, during the conduct of the study. KA reports personal fees from Vir, Gilead, Springbank, Roche, Janssen, Biotest, Aligos, Arbutus, Assembly and Shinoigi, grants from Merck Sharp & Dohme, outside the submitted work. GC reports personal fees from Gilead and Merck Sharp & Dohme outside the submitted work. BG reports other from Imperial National Institute for Health Research (NIHR) Biomedical Research Centres (BRC), during the conduct of the study. ATS reports grants from NIHR Health Informatics Collaborative, other from NIHR BRC, during the conduct of the study. EN reports grants from ViiV healthcare, grants from GlaxoSmithKline, grants from Gilead, outside the submitted work.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data are available to researchers on request, further details available at https://hic.nihr.ac.uk.

ORCID iDs
David Anthony Smith http://orcid.org/0000-0001-7778-7137
Kerrie Woods http://orcid.org/0000-0001-8106-8290
A. Torm Shaw http://orcid.org/0000-0001-5860-0146

REFERENCES

1 NIHR. Health informatics collaborative. Available: https://hic.nihr.ac.uk/[Accessed 17 Sep 2019].
2 Miotto R, Wang F, Wang S, *et al*. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018;19:1236–46.
3 WHO. Hepatitis B. Available: http://who.int/mediacentre/factsheets/fs204/en/ [Accessed 27 Feb 2015].
4 Stanaway JD, Flaxman AD, Naghavi M, *et al*. The global burden of viral hepatitis from 1990 to 2013: findings from the global burden of disease study 2013. *Lancet* 2016;388:1081–8.
5 Griggs D, Stafford-Smith M, Gaffney O, *et al*. Policy: sustainable development goals for people and planet. *Nature* 2013;495:305–7.
6 World Health Organization. *Global hepatitis report, 2017 (ISBN 978-92-4-156545-5)*. World Health Organization, 2017.
7 OpenClinica open source software. Available: www.OpenClinica.com
8 Nelson EK, Piehler B, Eckels J, *et al*. LabKey server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics* 2011;12:71.
9 Downs LO, Smith DA, Lumley SF, *et al*. Electronic health informatics data to describe clearance dynamics of hepatitis B surface antigen (HBsAg) and E antigen (HBeAg) in chronic hepatitis B virus infection. *MBio* 2019;10.