BMJ Open

Adult survivors of sickle cell disease, transfusion-dependent beta-thalassaemia and childhood acute leukaemia in England: protocol for a mixed methods data linkage and health-related quality of life survey study

Khalid Ahmed,[1,2] Ivana Holloway,[2] Kate Absolom ,[2] Samantha J Mason ,[2] Ruben Mujica-Mota,[3] Georgios Gkountouras,[3] Adam Martin,[3] Thuvia Flannery,[2] Michael Richards,[4] Emma Astwood,[5] Sam Ackroyd,[6] Brigit Greystoke,[7] Diana M Greenfield,[8] Quentin Hill,[4] Beki James,[4] Michelle Kwok- Williams,[9] Robert D Murray,[10] Clare Samuelson,[11] David Simcox,[12] John A Snowden,[1] Joseph Sharif,[13] Nandini Sadasivam,[13] Richard Feltbower ,[2] Adam Glaser[2]

To cite: Ahmed K, Holloway I, Absolom K, *et al*. Adult survivors of sickle cell disease, transfusion-dependent beta-thalassaemia and childhood acute leukaemia in England: protocol for a mixed methods data linkage and health-related quality of life survey study. *BMJ Open* 2026;**16**:e111664. doi:10.1136/bmjopen-2025-111664

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-111664).

Received 29 September 2025
Accepted 05 January 2026

For numbered affiliations see end of article.

Correspondence to
Dr Khalid Ahmed;
xbhn409@leeds.ac.uk

ABSTRACT

Introduction Recent advances in treatment and care have improved survival rates for children and young adults with severe blood disorders such as sickle cell disease (SCD), transfusion-dependent beta-thalassaemia (TDT) and acute leukaemia. However, their quality of life and reproductive and psychosocial outcomes are not yet well studied. For SCD and TDT, robust survival data are mainly limited to North America. Thus, there is a need to fill these knowledge gaps to guide improvements in care, address unmet clinical needs and rigorously assess the efficacy of emerging novel therapies.

Methods and analysis This is an observational population-based mixed-methods study of individuals diagnosed with SCD, TDT or acute leukaemia when under the age of 18 in England, involving a data linkage component and a patient-reported outcomes measures survey. Data linkage-eligible participants will be identified from national and regional databases, including the Hospital Episode Statistics, Yorkshire Specialist Register of Cancer in Children & Young People and the National Congenital Anomaly and Rare Diseases Registration Service. Data linkage will be processed within the NHS England and the University of Leeds' secure, trusted research environments. Data will be accessed without consent under section 251 and approval by the confidentiality advisory group. It will assess survival rates for SCD and TDT as well as clinical, educational and mental health outcomes for SCD, TDT and acute leukaemia diagnosed in childhood.

Survey-eligible participants for SCD, TDT and acute leukaemia cohorts will be checked for their suitability to participate by the North of England clinical care teams. An NHS-approved survey provider will facilitate data checks with the NHS National Data Opt-Out Service. Consent is required for participation in the survey and for subsequent data linkage to existing databases. Surveys are conducted in various formats (online, paper and phone), with reminders sent after 21 days. The survey will assess quality of life and psychosocial and reproductive outcomes. Participants can withdraw at any time, and support is available via telephone helplines.

Ethics and dissemination The study has received ethical and information governance approval from the Health Research Authority (Reference 24/YH/0186) and the Confidentiality Advisory Group (CAG 24/CAG/0138) to process identifiable data without consent. Study results will be available to patients, physicians, researchers, stakeholders and others through open-access publishing, results sharing via media platforms and presentations at conferences and meetings.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Rigorous mixed-methods design.
⇒ Strong patient and public involvement and engagement in the study design.
⇒ Robust ethical and data governance framework.
⇒ The survey is limited to the North of England and Yorkshire, which may affect the generalisability of the patient-reported outcome measures findings.
⇒ Different types of biases could be encountered in the survey, including selection bias and recall bias.

INTRODUCTION

Context

Advances in treatment have significantly increased survival rates among children and young adults with severe blood disorders.[1 2] In Yorkshire (a region in England), the 5-year survival rate for acute lymphoblastic

leukaemia (ALL) trial participants was approximately 88% in the UKALL2003 trial.[3]

Survivors of childhood acute leukaemia, specifically ALL, experience significantly late morbidity compared with the general population. The most prominent late effects include endocrine disorders (such as growth hormone deficiency and short stature), cardiovascular diseases, musculoskeletal complications (such as osteoporosis and osteonecrosis), neurocognitive disorders and increased risk of secondary malignancies.[4] Knowledge gaps remain regarding the long-term psychosocial, reproductive and socioeconomic outcomes. Recent, specific population-based data on education and employment are also lacking.

Sickle cell disease (SCD) and transfusion-dependent beta-thalassaemia (TDT) are among the most severe blood disorders, primarily affecting individuals from minority ethnic groups and carrying significant mortality rates. In the UK, 98.5% of SCD and TDT cases occur in non-white individuals,[5] who are more likely to be at risk of substantial health and social inequalities than those of white ethnic origin.[6]

SCD is a global issue, and approximately 300 000 infants are born annually with it, which is expected to rise to 400 000 by 2050.[7] In the UK, the NHS reports that 1 in 2436 newborns has a severe haemoglobin disorder.[8] Although survival data for TDT have been rarely reported, recent advances mean that >80% of patients with TDT live beyond 40 years.[9] Life expectancy for individuals with SCD in the USA is around two decades shorter than that of the general US population,[10 11] and a recent UK study suggested a median survival age of 67 years, though paediatric cases were excluded.[12]

Little is known about long-term health and social outcomes for young people with SCD and TDT. Lifespan extension in these cases relies on intensive treatments, including haematopoietic cellular therapies such as chemotherapy-conditioned allogeneic haemopoietic stem cell transplant (HSCT). UK data on long-term survival are limited, with most insights from North American studies. Recent research on TDT revealed a significant reduction in health-related quality of life (HRQL) and work productivity, indicating unmet clinical and social needs.[13]

Further research is necessary to assess the impact of these conditions on the UK population, especially among minority ethnic groups. These data will guide NHS service improvements by supporting optimal care to manage late complications, better counselling for patients and families, helping develop risk stratification tools and alert systems to ensure targeted resource allocation for those in greatest need, evaluating current and future curative treatments including haematopoietic cellular therapies against other treatment options and contributing to the design of healthcare delivery programmes at local, regional and national levels.

Aims

The primary aim of this study will be the following:

1. To define the duration of survival of young individuals diagnosed with SCD or TDT in England.
2. To measure the quality of life (QOL) of young individuals diagnosed with acute leukaemia, SCD or TDT in the North of England.

The secondary aims will be as follows:

1. To identify factors and groups of individuals associated with poor survival and adverse outcomes.
2. To provide insights into the long-term health and social impacts of life-threatening haematological conditions in young people.
3. To identify social factors contributing to optimal QOL within these groups.

Objectives (outcomes)

All participants (morbidity measurement)

1. To assess physical morbidity and healthcare costs by linking to Hospital Episode Statistics (HES) (frequency of hospital admission and health resource use).
2. To assess mental health outcomes by linking to NHS England's mental health datasets (incidence of mental health conditions such as depression).
3. To compare educational outcomes with those of peer groups through data linkage with the National Pupil Database (NPD) at the end of key stages 2 (11 years old), 4 (16 years old) and 5 (16–18 years old), as well as determining the prevalence of special educational needs.
4. To evaluate QOL outcomes, including HRQL, employment status, living standards, independent-living capacity and relationship status through a patient-reported outcomes survey.
5. To evaluate reproductive health outcomes based on responses from the patient-reported outcomes survey.
6. To examine the influence of ethnicity and socioeconomic deprivation on the above outcomes.

SCD and TDT (mortality measurement)

1. To determine the survival rates and life expectancies of individuals diagnosed with SCD or TDT across England.

METHODS AND ANALYSIS

Study design and eligibility criteria

This is an observational population-based study of individuals diagnosed with acute leukaemia, SCD or TDT in England. The study is structured in two main parts: data linkage and a survey. The inclusion criteria for each phase are as follows:

1. Data linkage

   Inclusion criteria

   Acute leukaemia: individuals diagnosed with ALL or acute myeloid leukaemia (AML) (see online supplemental material 1(A,B) for ICD-10 codes) before their 18th birthday in Yorkshire (a region in the north of England).

   SCD: individuals diagnosed with SCD consistent with one of the following phenotypes (SS, SC and SβThal),

identified in England (see online supplemental material 1C for ICD-10 codes).
TDT: individuals diagnosed with TDT identified in England (see online supplemental material 1D for ICD-10 codes).
Exclusion criteria
- Individuals with types of blood cancers other than AML or ALL.
- Individuals with AML or ALL diagnosed after their 18th birthday.
- Individuals with SCD phenotypes other than SS, SC or SβThal.
- Individuals with non-TDT.

2. Survey
Inclusion criteria
SCD and TDT: individuals diagnosed with SCD (SS, SC and SβThal) or TDT identified before their 18th birthday from 1 January 1990 onwards registered at one of the North-East, Yorkshire and North-West Haemoglobinopathy Coordinating Centres (HCC) for SCD, or North HCC for TDT (https://www.england.nhs.uk/commissioning/spec-services/npc-crg/blood-and-infection-group-f/haemoglobinopathies/specialised-haemoglobinopathy-services/).
Acute leukaemia: individuals diagnosed with AML or ALL before their 18th birthday from 1 January 1990 onwards and resident in Yorkshire at the time of their cancer diagnosis.
Exclusion criteria
- Individuals diagnosed with AML or ALL after their 18th birthday.
- Individuals who are younger than 18 years old.
- Individuals registered with the national data opt-out.
- Individuals deemed unsuitable to fill a survey by their respective clinical care teams, for example, significant mental illness, risk of psychological trauma or impaired cognitive functions.

Data sources for the study

Table 1 summarises the data sources which will be used in the study.

Sampling and sample size

The sample size is primarily determined by practical considerations, such as data availability due to the observational design of the data linkage component, feasibility and survey costs. This study represents the most extensive data linkage and patient-reported outcome analysis of these haematological diseases in young people. Using a conservative example of a 50% increased risk for a specific health outcome among 400 survivors (equivalent to the number of ALL survivors aged <18 years included in the study), we can achieve relative risk estimates with 99% confidence interval ranging from 1.12 to 1.95 for a simple effect size (RR=1.50). These parameters are comparable to published studies and correspond to a study with 90% statistical power. The estimated numbers of SCD and TDT cases registered with the National Haemoglobinopathy Register (NHR) are 15 000 and 1100, respectively (NHR report 2023–2024).[14] The YSRCCYP holds approximately 700 cases of acute leukaemia.

Study procedures

Data linkage

Eligible patients for the data linkage part of the study will be identified through population-based sources as outlined in the case ascertainment part of table 1. Patient identifiers (eg, NHS number, date of birth, sex, postcode and, where available, full name) will be used to link the data. Before any patients are identified from these databases, the National Data Opt-Out will be applied,[15] ensuring that individuals who have opted out via NHS England are excluded. A master list of each cohort will be created from the remaining eligible patients, which will subsequently be linked to other databases, as illustrated in figure 1. Each cohort's secure linked data extract will then be pseudonymised for the analysis. Data linkage for SCD and TDT will be carried out within the NHS England Data Access Environment. As the leukaemia cohort will be derived from the YSRCCYP (a regional population-based dataset held by the University of Leeds under separate HRA CAG approval (20CAG0133)), the data linkage for this cohort will be done within the ARO TRE at the University of Leeds. Figure 2 illustrates a similar process for data linkage from the supplementary data sources if needed. It is expected to take between 6 and 12 months from January 2026.

Survey

Study design

This is a cross-sectional survey study embedded within a larger population-based observational study.

Survey instrument and development

Questionnaire description: The self-administered survey contains four sections: (1) demographic and clinical characteristics, (2) health-related QOL and psychosocial outcomes, (3) reproductive health outcomes and (4) healthcare resource use and productivity. It incorporates several validated patient-reported outcome measures (PROMS) as detailed in table 2. The impact of evolution and modifications in therapy modalities on the QOL will be examined by comparing groups of different birth cohorts.

Instrument validation: The specific PROMs used have established validity and reliability in relevant populations, as cited in table 2. For disease-specific modules (Adult Sickle Cell Quality of Life Measurement Information System for SCD and Transfusion-Dependent Quality of Life for TDT), scoring will follow published algorithms to generate domain and summary scores.

Pre-testing and cognitive interviewing: the draft survey instrument underwent cognitive pre-testing with members of the haematology lived experiences and

Table 1 Data sources for case ascertainment and data linkage

Category	Data source	Purpose	Key features
Case ascertainment	Yorkshire Specialist Register of Cancer in Children & Young People (YSRCCYP) (https://ysrccyp.org.uk/).[30]	To identify cases of AML and ALL	A registry with young individuals diagnosed with cancer in Yorkshire and Humber
	Hospital Episode Statistics (HES) (https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics)	To identify cases of SCD and TDT	Contains details about admissions, outpatient appointments and emergency attendances at NHS hospitals in England
	National Congenital Anomaly and Rare Diseases Registration Service (NCARDRS) (https://digital.nhs.uk/ndrs/about/ncardrs)[31]		Records people with congenital anomalies and rare diseases across England. Includes newborn screening data
	North-East and Yorkshire and North-West Haemoglobinopathy Coordinating Centres (HCC) (NHS commissioning » Specialised haemoglobinopathy services)	Provides access to patients with SCD and TDT for surveys across Northern England	Covers (Leeds, Sheffield, Liverpool, Manchester and Newcastle)
Achieve study objective	Office for National Statistics (ONS) (https://www.ons.gov.uk/)[32]	Assessment of mortality and survival rates for SCD and TDT	Supplies death records, including causes, dates and locations. This will be via the ONS-HES linked dataset accessed via NHS England
	Mental Health Services Data Set (MHSDS) (https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set)	Assessment of mental health outcomes	A patient-level, output-based data set aims to deliver robust, comprehensive, nationally consistent and comparable person-based information for patients in contact with mental health services
	National Pupil Database (NPD) (https://www.data.gov.uk/dataset/9e0a13ef-64c4-4541-a97a-f87cc4032210/national-pupil-database)[33]	Assessment of educational outcomes	Includes test/examination results, progression records, attendance rates and special educational needs information
Supplementary	Education and child health insights from linked data (ECHILD)[34 35]	Gather preliminary data and outcomes while awaiting approval for the data linkage and survey components	Includes hospital activity, educational records, social care and community health services. Links health and education data for individuals born in England since 1984

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukaemia; SCD, Sickle cell disease; TDT, transfusion-dependent beta-thalassaemia.

outcomes (HALO) and patient and public involvement and engagement (PPIE) group, which includes individuals with lived experience of the target conditions. This process assessed question clarity, relevance, comprehensiveness and estimated completion time (approximately 20–25 min). Feedback was used to refine the wording of items, instructions and the survey flow. The final survey versions for acute leukaemia, SCD and TDT are provided in the online supplemental materials 2–4.

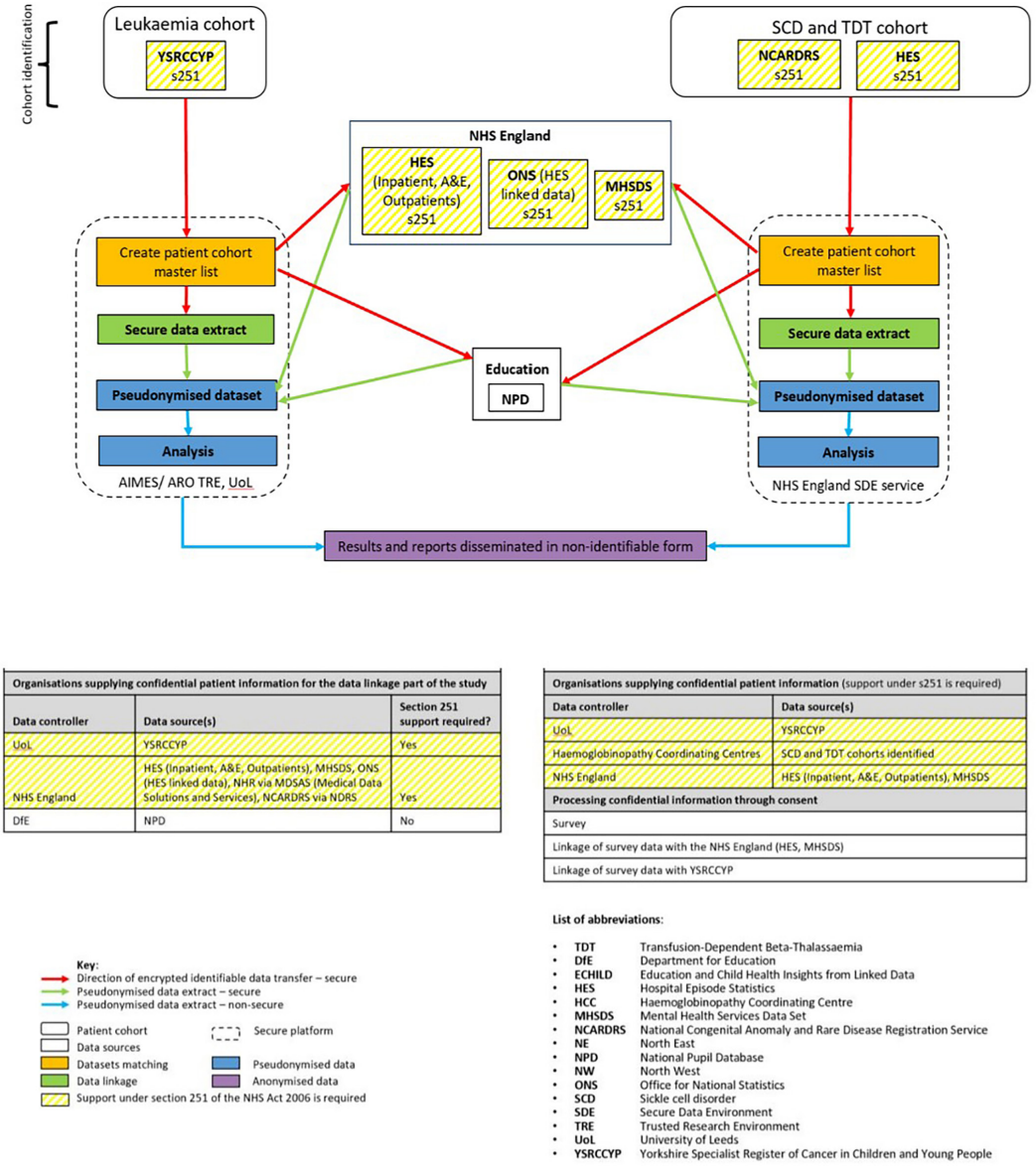
Sample characteristics and selection

The inclusion and exclusion criteria are detailed in the study design and eligibility criteria section. For the leukaemia cohort, eligible individuals will be identified through the YSRCCYP. For the SCD and TDT cohorts, eligible participants will be identified through participating centres within the geographic areas of the HCCs.

Sampling technique: a consecutive (census) sampling approach will be used within the defined geographical and clinical cohorts. All eligible individuals identified through the data sources will be invited to participate. The clinical teams responsible for the participants will review the lists and exclude any participant they deem unsuitable for inclusion in the survey.

Sample size considerations: the survey sample size is determined by the available population within the participating Northern England centres, as detailed in the main protocol. With an estimated eligible pool of approximately 2000 individuals across the three cohorts, a conservative response rate of 30% would yield 600 completed surveys. This provides adequate precision for descriptive summaries of PROMs (eg, margin of error for a proportion of ±4% for n=600) and sufficient power for key comparative analyses outlined in the statistical analysis plan.

Representativeness: As the survey is limited to Northern England, findings may not be fully generalisable to the national population of survivors. Representatives will be assessed by comparing respondent demographics (age, sex, ethnicity and deprivation) with the known



Organisations supplying confidential patient information for the data linkage part of the study		
Data controller	Data source(s)	Section 251 support required?
UoL	YSRCCYP	Yes
NHS England	HES (Inpatient, A&E, Outpatients), MHSDS, ONS (HES linked data), NHR via MDSAS (Medical Data Solutions and Services), NCARDRS via NDRS	Yes
DfE	NPD	No

Organisations supplying confidential patient information (support under s251 is required)	
Data controller	Data source(s)
UoL	YSRCCYP
Haemoglobinopathy Coordinating Centres	SCD and TDT cohorts identified
NHS England	HES (Inpatient, A&E, Outpatients), MHSDS
Processing confidential information through consent	
Survey	
Linkage of survey data with the NHS England (HES, MHSDS)	
Linkage of survey data with YSRCCYP	

Key:
- Direction of encrypted identifiable data transfer – secure
- Pseudonymised data extract – secure
- Pseudonymised data extract – non-secure
- Patient cohort
- Data sources
- Datasets matching
- Data linkage
- Support under section 251 of the NHS Act 2006 is required
- Secure platform
- Pseudonymised data
- Anonymised data

List of abbreviations:

TDT	Transfusion-Dependent Beta-Thalassaemia
DfE	Department for Education
ECHILD	Education and Child Health Insights from Linked Data
HES	Hospital Episode Statistics
HCC	Haemoglobinopathy Coordinating Centre
MHSDS	Mental Health Services Data Set
NCARDRS	National Congenital Anomaly and Rare Disease Registration Service
NE	North East
NPD	National Pupil Database
NW	North West
ONS	Office for National Statistics
SCD	Sickle cell disorder
SDE	Secure Data Environment
TRE	Trusted Research Environment
UoL	University of Leeds
YSRCCYP	Yorkshire Specialist Register of Cancer in Children and Young People

Figure 1 Data linkage flowchart for sickle-cell disease, transfusion-dependent beta-thalassaemia and acute leukaemia cohorts.

characteristics of the full eligible cohort derived from linked databases.

Survey administration and procedures

Mode of administration: the survey will be advertised through social media and posters in the contributing sites. It will be administered in multiple formats to maximise accessibility and response: online (via a secure web link), paper (via post) or by telephone interview on request. IQVIA (an NHS-approved survey provider) will put checks in place to prevent data entry errors and multiple online entries (eg, unique invitation links and dual review).

Recruitment, processing and time frame: recruitment will occur over an 8-week period starting from 15 January 2026. The survey provider, IQVIA,[16] utilises the Message Exchange for Social Care and Health service to facilitate data checks and confirm that individuals are not deceased. Through this service, IQVIA will send a list of NHS numbers to NHS England, where the data are cross-checked against the National Data Opt-Out Service.[15] The updated file will be returned with the NHS numbers of individuals who have opted out removed. Invitation packs containing a patient information sheet (PIS), consent form and survey (paper version) will be mailed by the appointed NHS-approved survey provider (IQVIA). A reminder letter will be sent to non-responders after 21 days. If mobile numbers are available, a text message reminder will also be sent. To create a longitudinal leukaemia cohort for the YSRCCYP, data from consented survey participants will be linked to the YSRCCYP database. These data will be retained indefinitely, subject to

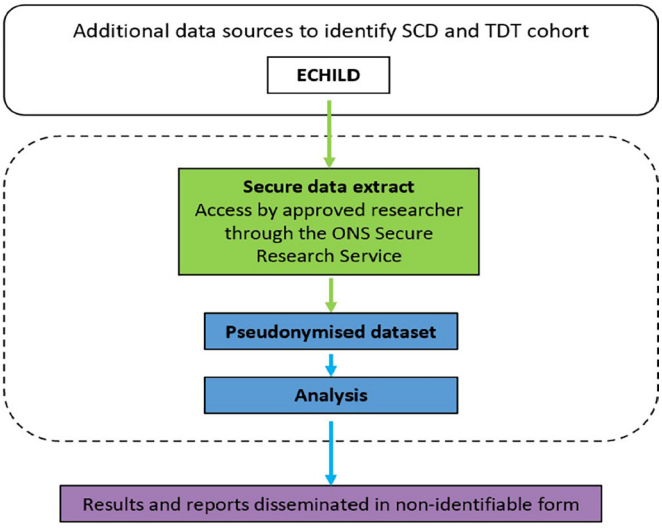


Figure 2 Data linkage from supplementary data sources. ECHILD, Education and Child Health Insights from Linked Data; SCD, sickle-cell disease; TDT, transfusion-dependent beta-thalassaemia.

continued HRA CAG approval and funding. However, if the research funding for the YSRCCYP expires, approval will be sought, and arrangements will be made to archive the data in the TRE within ARO (or another equivalent TRE at the time). The data will be correlated with existing epidemiological information on long-term health and social outcomes and used to monitor changes in patient-reported outcomes over time. Pseudonymised personal data for patients with SCD and TDT, derived from the consented survey participants, will be retained to establish a cohort for studying patient-reported outcomes. See figure 3 for an overview of the survey process.

Participant preparation and support: clinical teams will be briefed on the study to support participant enquiries. A dedicated survey helpline operated by IQVIA will be available. The PIS clearly outlines the study's purpose and provides contact details for the research team. Participants have the right to withdraw from the survey at any time without providing a reason.

Patient and public involvement and engagement

The study team co-developed the HALO PPIE group, which included people with lived experience of the conditions, input from national charities to ensure appropriateness, cognitive testing of the survey, and transparency, as well as adherence to ethical standards and confidentiality in handling patient-identifiable data. Data confidentiality and consent concerns have been addressed through consultations with the HALO PPIE group members, conversations with young people, parents/carers attending local clinics and through patient events. Support was provided by clinical teams, Leeds National Institute for Health and Care Research (NIHR) Biomedical Research Centre PPIE team and the NIHR Ethnic Minority Research Involvement group.

Key steps to reduce participant burden included the following:

1. Designing a concise, essential-item survey that participants can complete at their own pace.

Table 2 Validated questionnaires used in the survey

Questionnaire	Purpose	Key features	Scoring
Generic Quality of Life Measure (EQ-5D-5L and EQ-VAS)	A health utility tool to assess various dimensions of health and provide a patient-driven quantitative measure[36]	► Assesses five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. ► A visual analogue scale (VAS) records self-rated health (0–100)	► EQ-5D-5L (–0.59 to 1): Summary index (0=death, 1=perfect health, negative for worse than death). ► EQ VAS: scale from 0 to 100 (worst to best imaginable health)
Hospital Anxiety and Depression Scale (HADS)	Assesses anxiety and depression in individuals with medical conditions[37]	14 items: 7 for anxiety and 7 for depression	► Subscales range 0–21. ► Severity levels: 0–7 (normal), 8–10 (mild), 11–14 (moderate) and 15–21 (severe)
Adult Sickle Cell Quality of Life Measurement Information System (ASCQ-Me)	Captures patient-reported outcomes in adults with sickle-cell disease[38 39]	30 items across seven domains: pain frequency, pain severity, pain impacts, emotional well-being, social functioning, stiffness and sleep-standardised mean scores	► Scores range from 5 (never) to 1 (always). ► Mean=50, SD=10. ► Higher scores=better HRQoL, except for pain domains
Transfusion-Dependent QoL Questionnaire (TranQoL)	Measures the disease-specific quality of life in individuals with thalassemia major[40]	36 items covering physical health, emotional health, family functioning, school/career and sexual activity	► Provides insights into the unique challenges faced by individuals with thalassemia major ► Score 0 (worst) to 100 (best)

EQ-5D-5L, EuroQol 5-Dimension 5-Level.

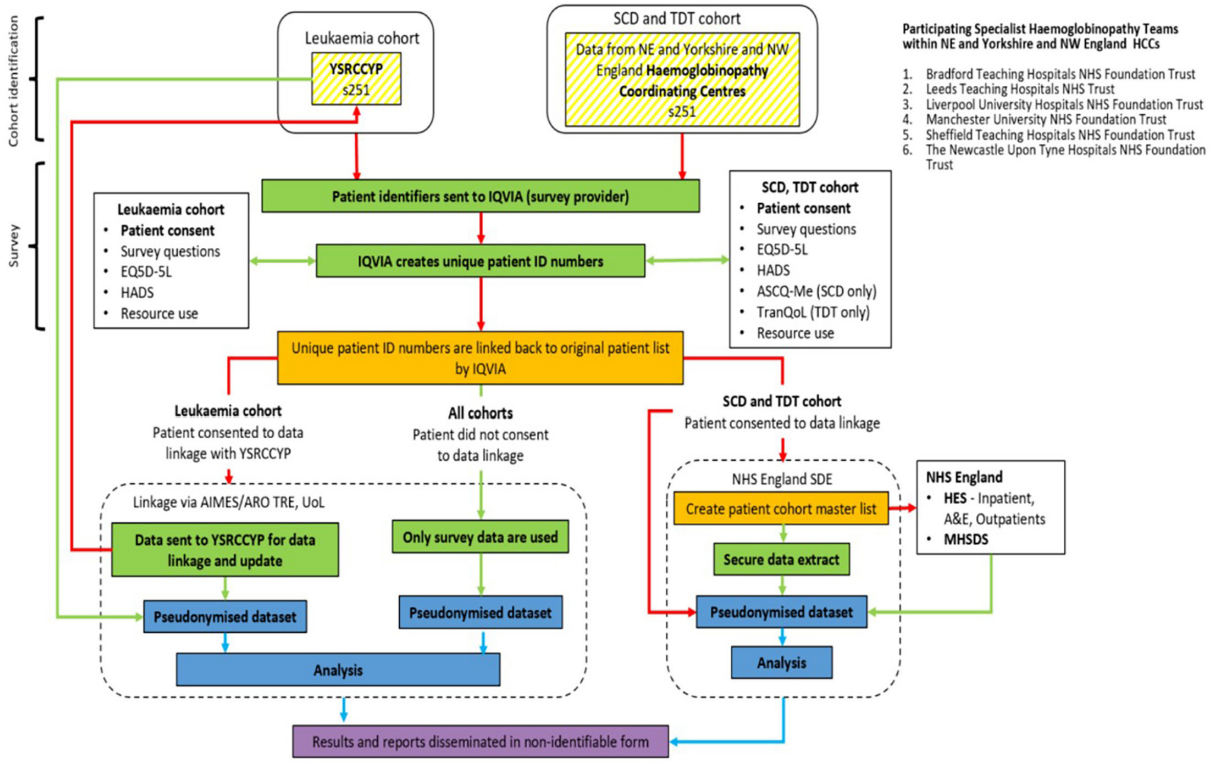


Figure 3 Survey flowchart. AIMS, Adult Sickle Cell Quality of Life Measurement System; ARO, Academic Research Organisation; ASCQ-Me, Adult Sickle Cell Quality-of-Life Measurement Information System; EQ5D-5L, EuroQol 5-Dimension 5-Level questionnaire; HADS, Hospital Anxiety and Depression Scale; HES, Hospital Episode Statistics; MHSDS, Mental Health Services Data Set; SCD, Sickle Cell Disease; TranQoL, Transition Quality of Life; TRE, trusted research environment; TTDT, transfusion-dependent beta-thalassaemia; YSRCCYP, Yorkshire Specialist Register of Cancer in Children & Young People.

2. Providing multiple completion formats (online, paper and telephone) with optional responses and support for language needs.
3. Offering alternatives for participants with cognitive challenges due to fatigue, such as pausing the survey or skipping questions.

The survey was refined with input from the PPIE group to ensure validity. Participants will be informed of support services (eg, the research team, IQVIA helpline and patient advocacy liaison service) in case of concerns. The patient information sheet (PIS) promotes communication and guides participants throughout the study.

Health and social care resource use

Data linkage with HES, Emergency Care Datasets and Civil Death Registrations will track health service utilisation over participants' lifetimes. Costs for primary care visits will be valued using unit costs from the Personal Social Services Research Unit, whereas hospital service costs will be calculated using healthcare resource (HRG) codes and the National Schedule of NHS costs.[17 18] Additionally, the survey will incorporate items related to personal healthcare costs and productivity losses (eg, days off work) to complement health economics analysis, leveraging insights from prior research on the financial impacts of such illnesses.

Statistics and analysis

Data completeness, missing data and data linkage techniques

The extent of missing data will be reported, and patterns or discrepancies in ethnicity and social deprivation reporting across data sources will be analysed. Ethnicity data, as noted in the latest ONS report, tends to be incomplete for certain populations.[19] This may lead to unintentional bias, particularly since SCD and TDT primarily affect individuals from minority ethnic backgrounds. Ensuring high-quality ethnicity recording is critical; therefore, sensitivity analysis will explore the effects of missing data and potential biases, using multiple imputation methods where appropriate to address these gaps.

The statistical analysis plan (SAP) will detail linkage variables and quality assessments for each data source. Deterministic matching will be used for most datasets, with the NHS number as the primary identifier. For datasets without NHS numbers, such as NPD, probabilistic matching will be applied based on full name, sex, date of birth and postcode, following established protocols from YSRCCYP data linkage studies.[20–22]

General statistical approaches

Study outcomes will adhere to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for reporting observational studies.[23] Where applicable, continuous variables will be described

using mean, SD, median, minimum, maximum and IQR. Categorical variables will be summarised as frequencies and percentages. Demographic data will be stratified for each cohort.

Key exposure variables, such as ethnicity and social deprivation, will be modelled to examine their impact on health outcomes within each disease group/cohort. Ethnicity will be categorised following 2021 census groupings,[24] while deprivation scores will be calculated using the Townsend score, derived from the residential address and postcode at diagnosis through geo-coding.[25]

To analyse count data, Poisson regression will be employed to estimate the effects of ethnicity and deprivation, adjusting for relevant confounders identified using causal inference techniques.[26] Continuous outcomes will be analysed using linear regression, and binary or ordinal outcomes will be assessed through logistic/binomial regression or ordered logistic regression, respectively. Survival analysis will use the last known date of contact (or death) and be modelled in relation to deprivation and ethnicity. Survival analysis will use Kaplan–Meier methods to calculate life expectancies and survival probabilities for patients with SCD and TDT by birth cohort and before and after the introduction of practice-changing interventions, including blood safety screening, introduction of hydroxycarbamide, transcranial Doppler screening and iron chelation. Flexible parametric methods will be employed to examine the impact of ethnicity and deprivation on survival outcomes, while adjusting for confounding factors.

Survey and PROMs data will be reported separately for all participants who returned surveys. The response rate, including non-responders, will also be summarised. Likert scale responses will be analysed using ordinal logistic regression, while validated PROMs will be scored according to their respective manuals.

Health economics

The health economics analysis will include the following:

1. *Cost of illness analysis:*
   - NHSE secondary care perspective using HES data: the analysis will assess resource utilisation over time from diagnosis. HES data resource use will be costed by attaching unit costs from the latest NHS reference cost schedule. Healthcare costs will be analysed using appropriate generalised linear regression models, dependent on the distribution of each outcome variable in the dataset. Two-part models may be used for truncated distributions[27 28] and to address censoring in outcomes. Where appropriate, some variables will be log-transformed, e.g. household income.
   - Patient out-of-pocket costs perspective using survey data: we will provide descriptive statistics of the costs reported by participants in the survey data.
2. *HRQL analysis*: HRQL EQ-5D-5L utility scores will be analysed using Gamma distributions.
3. *Inequality analysis*: Differences in survival, HRQL and costs will be analysed across key socioeconomic groups, genders, ethnicities and area-level deprivation (IMD quintiles) to identify unmet needs, potentially using Lorenz and concentration curves where data support this.[29] The methodology applied here will aid the design of future nationally representative studies.

Ethics and information governance

Patients will be informed about the study in all HCCs in England (via patient information leaflets) and will have the opportunity not to be included in the data linkage component. Informed consent will be obtained from all participants before they are enrolled in the survey component of the study. Participants can withdraw from the survey at any time without explanation.

The study has been approved by the Health Research Authority - Research Ethics Committee (Reference 24/YH/0186) and the Confidentiality Advisory Group (24/CAG/0138). The study will adhere to the General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018, ensuring data are de-identified as soon as practicable. Each participant will be assigned a unique patient identification number, and only pseudonymised data will be used for analysis to safeguard personal information. This minimises the risk of data security breaches.

A risk register and monitoring plan will be developed prior to the study's commencement and will be updated as needed to reflect protocol modifications. Regular monitoring will be conducted in accordance with the Risk Assessment Monitoring Plan, and data will be reviewed for adherence to the protocol and accuracy against source documents as outlined in the study-specific monitoring plan.

Investigators are required to report any potential data breaches to the research team. The University of Leeds' formal processes will then be followed. A potential breach refers to a protocol violation or a failure to uphold GCP principles that significantly impacts the safety, physical or mental integrity of study participants or the scientific validity of the research.

On completion of the study, pseudonymised data will be securely archived for 10 years by the University of Leeds in accordance with the sponsor's procedures. Afterwards, arrangements will be made for the confidential destruction of the data. No records will be destroyed without prior written approval from the sponsor (University of Leeds).

Dissemination

Effective dissemination of the results will be crucial to ensure timely and appropriate learning, ultimately leading to the achievement of the programme's objectives and an improved quality of life and care for the target populations. Thus, while the research findings will be reported through traditional academic formats, such as peer-reviewed open-access journal

articles and conference presentations, efforts will also be made to disseminate the results directly to key stakeholders, including patients, families, clinicians, local and national health and social care systems, and relevant third-sector organisations.

To maximise the reach and accessibility of the findings, the results will be:

a. Specifically aimed at the relevant stakeholders and audiences.
b. Available in a variety of formats.
c. Presented through a wide range of media channels.
d. Accessible in open and unrestricted formats, including open-access publications and digital media platforms (while ensuring participants' confidentiality). All results, presented in various formats, will be available on the dedicated HALO study website (https://halo-haemstudy.leeds.ac.uk/).

Author affiliations
[1]School of Medicine and Population Health, University of Sheffield, Sheffield, UK
[2]Leeds Institute for Data Analytics, School of Medicine, University of Leeds, Leeds, UK
[3]Academic Unit of Heath Economics, Leeds Institute of Health Sciences, School of Medicine, University of Leeds, Leeds, UK
[4]Department of Haematology, Leeds Teaching Hospitals NHS Trust, Leeds, UK
[5]Department of Haematology, Sheffield Children's Hospital NHS Foundation Trust, Sheffield, UK
[6]Department of Haematology, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[7]Department of Haematology, Royal Victoria Infirmary, Newcastle upon Tyne, UK
[8]School of Allied Health Professions, Nursing, and Midwifery, University of Sheffield, Sheffield, UK
[9]Leeds Teaching Hospitals NHS Trust, Leeds, UK
[10]Department of Endocrinology, Leeds Teaching Hospitals NHS Trust, Leeds, UK
[11]Department of Haematology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
[12]Department of Haematology, Royal Liverpool University Hospital, Liverpool, UK
[13]Department of Haematology, Manchester University NHS Foundation Trust, Manchester, UK

Acknowledgements The authors gratefully acknowledge the support of the Child Health Outcomes Research at Leeds (CHORAL) programme in conducting this study. The authors also thank the University of Sheffield for funding the article processing charges associated with this publication.

Contributors AG and RF served as the principal investigators of the research project. KAh, IH, KAb, SJM, RM-M, GG, AM, TF, MR, EA, SA, BG, DMG, QH, BJ, MK-W, RDM, CS, DS, JAS, JS and NS contributed to the study design and the development and writing of the study protocol, and approved the final version of the manuscript. KAh acts as the guarantor of the study. Grammarly was used to assist with checking spelling and verb tense consistency.

Funding This study is funded by Leeds Hospitals Charity (Award reference A2001526). For the purpose of open access, the authors has applied a Creative Commons Attributions (CC BY) license to any Author Accepted Manuscript version arising.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Kate Absolom https://orcid.org/0000-0002-5477-6643
Samantha J Mason https://orcid.org/0000-0003-0306-8353
Richard Feltbower https://orcid.org/0000-0002-1728-9408

REFERENCES

1 Maddams J, Utley M, Møller H. Projections of cancer prevalence in the United Kingdom, 2010-2040. *Br J Cancer* 2012;107:1195–202.
2 Telfer P, Coen P, Chakravorty S, *et al*. Clinical outcomes in children with sickle cell disease living in England: a neonatal cohort in East London. *Haematologica* 2007;92:905–12.
3 van Laar M, Kinsey SE, Feltbower RG. Survival of childhood acute lymphoid leukaemia in Yorkshire by clinical trial era, 1990-2011. *Br J Haematol* 2015;171:116–9.
4 Kane E, Kinsey S, Bonaventure A, *et al*. Excess morbidity and mortality among survivors of childhood acute lymphoblastic leukaemia: 25 years of follow-up from the United Kingdom Childhood Cancer Study (UKCCS) population-based matched cohort. *BMJ Open* 2022;12:e056216.
5 NHS sickle cell and thalassaemia screening programme. Data report 2016 to 2017: trends and performance analysis: phe publications gateway number: 2018025. 2018. Available: https://assets.publishing.service.gov.uk/media/5b112835ed915d2cdb024c3f/sct_data_report_2016_to_2017.pdf
6 Marmot M. Fair Society, Healthy Lives. 2013;1–74.
7 Piel FB, Hay SI, Gupta S, *et al*. Global burden of sickle cell anaemia in children under five, 2010-2050: modelling based on demographics, excess mortality, and interventions. *PLoS Med* 2013;10:e1001484.
8 England PH. NHS sickle cell and thalassaemia screening programme, data report 2016 to 2017: trends and performance analysis. *Public Health England London* 2018.
9 Modell B, Khan M, Darlison M, *et al*. Improved survival of thalassaemia major in the UK and relation to T2* cardiovascular magnetic resonance. *J Cardiovasc Magn Reson* 2008;10:42.
10 Lanzkron S, Carroll CP, Haywood C Jr. Mortality Rates and Age at Death from Sickle Cell Disease: U.S., 1979–2005. *Public Health Rep* 2013;128:110–6.
11 Lubeck D, Agodoa I, Bhakta N, *et al*. Estimated Life Expectancy and Income of Patients With Sickle Cell Disease Compared With Those Without Sickle Cell Disease. *JAMA Netw Open* 2019;2:e1915374.
12 Gardner K, Douiri A, Drasar E, *et al*. Survival in adults with sickle cell disease in a high-income setting. *Blood* 2016;128:1436–8.
13 Ryan K, Pancham S, Telfer P, *et al*. PRO146 HEALTH-RELATED QUALITY OF LIFE AND PRODUCTIVITY/ACTIVITY IMPAIRMENT IN PATIENTS WITH TRANSFUSION DEPENDENT B-THALASSAEMIA IN THE UK NHS: DATA FROM A UK MULTICENTRE OBSERVATIONAL STUDY. *Value Health* 2019;22:S868–9.
14 NHR. n.d. Available: https://nhr.mdsas.com
15 National opt out. n.d. Available: https://www.sbs.nhs.uk/national-data-opt-out-policy/#:~:text=What%20is%20the%20national%20data,out%20choice%20at%20any%20time
16 IQVIA. n.d. Available: https://www.iqvia.com
17 Jones K. Unit costs of health and social care 2022. 2022. Available: https://pure.york.ac.uk/portal/en/publications/unit-costs-of-health-and-social-care-2022
18 National cost collection 2019/2020 report. n.d. Available: https://www.england.nhs.uk/wp-content/uploads/2021/06/National-Cost-Collection-2019-20-Report-FINAL.pdf
19 Quality of ethnicity data. n.d. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/datasets/qualityofethnicitydatainhealthrelatedadministrativedatas

ourceswherepopulationwasrestrictedtothosewithdataforallsociodemographiccharacteristics
20 Smith L, Glaser AW, Greenwood DC, *et al*. Cumulative burden of subsequent neoplasms, cardiovascular and respiratory morbidity in young people surviving cancer. *Cancer Epidemiol* 2020;66:101711.
21 Smith L, Glaser AW, Peckham D, *et al*. Respiratory morbidity in young people surviving cancer: Population-based study of hospital admissions, treatment-related risk factors and subsequent mortality. *Int J Cancer* 2019;145:20–8.
22 van Laar M, Feltbower RG, Gale CP, *et al*. Cardiovascular sequelae in long-term survivors of young peoples' cancer: a linked cohort study. *Br J Cancer* 2014;110:1338–41.
23 Elm E von, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.
24 Ethnic group classification: census. 2021. Available: https://www.ons.gov.uk/census/census2021dictionary/variablesbytopic/ethnicgroupnationalidentitylanguageandreligionvariablescensus2021/ethnicgroup/classifications
25 Hoy R. Health and Deprivation P Townsend P Phillimore A Beattie Health and Deprivation Published by Croom Helm212pp£19.950-7099-4351-2. *Nurs Stand* 1988;2:34.
26 Tennant PWG, Murray EJ, Arnold KF, *et al*. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol* 2021;50:620–32.
27 Mihaylova B, Briggs A, O'Hagan A, *et al*. Review of statistical methods for analysing healthcare resources and costs. *Health Econ* 2011;20:897–916.
28 Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. *Med Decis Making* 2012;32:56–69.
29 Aaberge R. Characterizations of Lorenz curves and income distributions. *Soc Choice Welfare* 2000;17:639–53.
30 Yorkshire specialist registry for cancer in children and young people. n.d. Available: https://ysrccyp.org.uk/research/data-and-infographics
31 NCARDRS. n.d. Available: https://digital.nhs.uk/ndrs/about/ncardrs
32 ONS. n.d. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2020to2022#life-expectancy-at-birth-in-uk-countries
33 NPD. n.d. Available: https://www.data.gov.uk/dataset/9e0a13ef-64c4-4541-a97a-f87cc4032210/national-pupil-database
34 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
35 Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;48:1740–1740g.
36 Group TE. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199–208.
37 Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–70.
38 Cooper O, McBain H, Tangayi S, *et al*. Psychometric analysis of the adult sickle cell quality of life measurement information system (ACSQ-Me) in a UK population. *Health Qual Life Outcomes* 2019;17:74.
39 Keller SD, Yang M, Treadwell MJ, *et al*. Patient reports of health outcome for adults living with sickle cell disease: development and testing of the ASCQ-Me item banks. *Health Qual Life Outcomes* 2014;12:125.
40 Klaassen RJ, Barrowman N, Merelles-Pulcini M, *et al*. Validation and reliability of a disease-specific quality of life measure (the TranQol) in adults and children with thalassaemia major. *Br J Haematol* 2014;164:431–7.