FULL/LONG TITLE OF THE STUDY	**Using machine learning to model early-onset neonatal sepsis risk in late preterm and term neonates in Uganda**
SHORT STUDY TITLE / ACRONYM The short title should be: • Sufficiently detailed to make clear to participants what the research is about in simple English • If acronyms are used the full title should explain them. The proposed acronym should not drive the long title	**Using machine learning to model early-onset neonatal sepsis risk in Uganda (NeoRisk)**
PROTOCOL VERSION NUMBER AND DATE Version control: • All draft versions should be numbered 0.1, 0.2 etc. • The final version for submission should be numbered 1.0 • The changes made relative to the previous protocol version should be listed after submission	**1.0 – 30/11/2023** **2.0 - 09/01/2024** **2.1 – 09/02/2024**
IRAS Number:	
JRES Reference Number	
Funder Reference Number:	
This protocol has regard for the HRA guidance and order of content	

SIGNATURE PAGE

The undersigned confirm that the following protocol has been agreed and accepted and that the Chief Investigator agrees to conduct the study in compliance with the approved protocol and will adhere to the principles outlined in the Declaration of Helsinki, the Sponsor's SOPs, and other regulatory requirement.

I agree to ensure that the confidential information contained in this document will not be used for any other purpose other than the evaluation or conduct of the investigation without the prior written consent of the Sponsor.

I also confirm that I will make the findings of the study publicly available through publication or other dissemination tools without any unnecessary delay and that an honest accurate and transparent account of the study will be given; and that any discrepancies from the study as planned in this protocol will be explained.

For and on behalf of the Study Sponsor:

Signature:

.....

Date:
...../...../.....

Name (please print):

.....

Position:

.....

Chief Investigator:

Signature:

.....[redacted].....

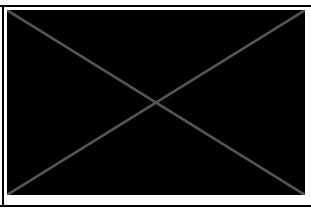
Name: (please print):

.[redacted] .....

[redacted]

ABBREVIATIONS	
CI	Chief Investigator
CRF	Case Report Form
ICF	Informed Consent Form
ISF	Investigator Site File
NHS	National Health Service
NIHR	National Institute for Health Research
PI	Principal Investigator
REC	Research Ethics Committee
SGUL	St Georges, University of London
JRES	(St Georges) Joint Research and Enterprise Services
WHO	World Health Organisation
ROM	Rupture of membranes
EOS	Early onset sepsis
LOS	Late onset sepsis
PCR	Polymerase chain reaction
LRS	Low resource setting

KEY STUDY CONTACTS	
Chief Investigator	[redacted]
Principal Investigator (In-Country)	[redacted]
Study Co-ordinator	[redacted]
Sponsor	St George's, University of London
Funder(s)	CREATE (Africa Health Research Training Programme) Wellcome Trust
Key Protocol Contributors	[redacted]

STUDY SUMMARY	
Study Title	**Using machine learning to model early-onset neonatal sepsis risk in Uganda**
Internal ref. no. (or short title)	NeoRisk
Study Design	Observational
Study Participants	Neonates born at ≥34 weeks' gestational age at Kawempe Hospital in Kampala, Uganda (Phase 1) or Sally Mugabe Central Hospital in Harare, Zimbabwe (Phase 3)
Planned Size of Sample (if applicable)	1600 (Phases 1+2), 900 (Phase 3)
Follow up duration (if applicable)	1 month post-delivery
Planned Study Period	September 2023-May 2024
Research Question/Aim(s)	1. Determine antenatal and perinatal factors associated with early-onset neonatal sepsis in low-resource settings, as diagnosed by a blood culture or a senior clinician 2. Measure the strength of association between the above factors and early-onset neonatal sepsis 3. Combine the above risk factors into a risk stratification model for early-onset neonatal sepsis using machine learning techniques

FUNDING AND SUPPORT	
FUNDER(S)	**FINANCIAL AND NON FINANCIALSUPPORT GIVEN**
CREATE PhD Programme	

ROLES AND RESPONSIBILITIES OF STUDY MANAGEMENT COMMITEES/GROUPS & INDIVIDUALS

The study has been discussed with the Community Engagement team at Makerere University-Johns Hopkins University Research Collaboration in Kampala, Uganda, which works with clinical studies at Kawempe Hospital.

The team have provided feedback on the acceptability of this study, including on the taking of a blood culture sample, recruitment strategy, and on development of patient information resources.

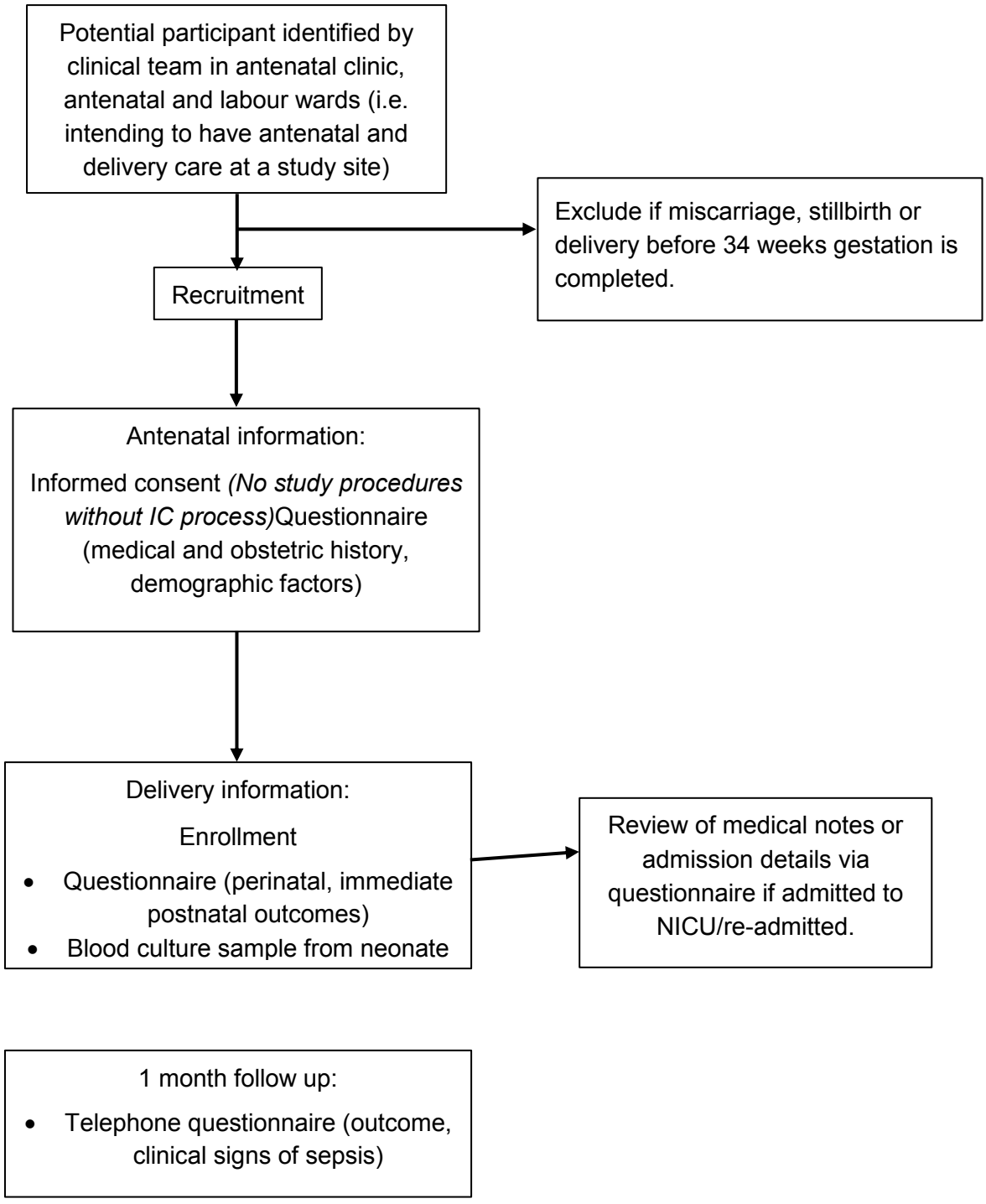
The study team will continue to work closely with the Community Engagement team throughout the study to ensure its acceptability and to disseminate research findings to the community and relevant stakeholders.

PROTOCOL CONTRIBUTORS

No funders have been involved in study design and no funder will be involved in conduct, data analysis and interpretation, manuscript writing, or dissemination of results. The funder does not control the final decision regarding these aspects.

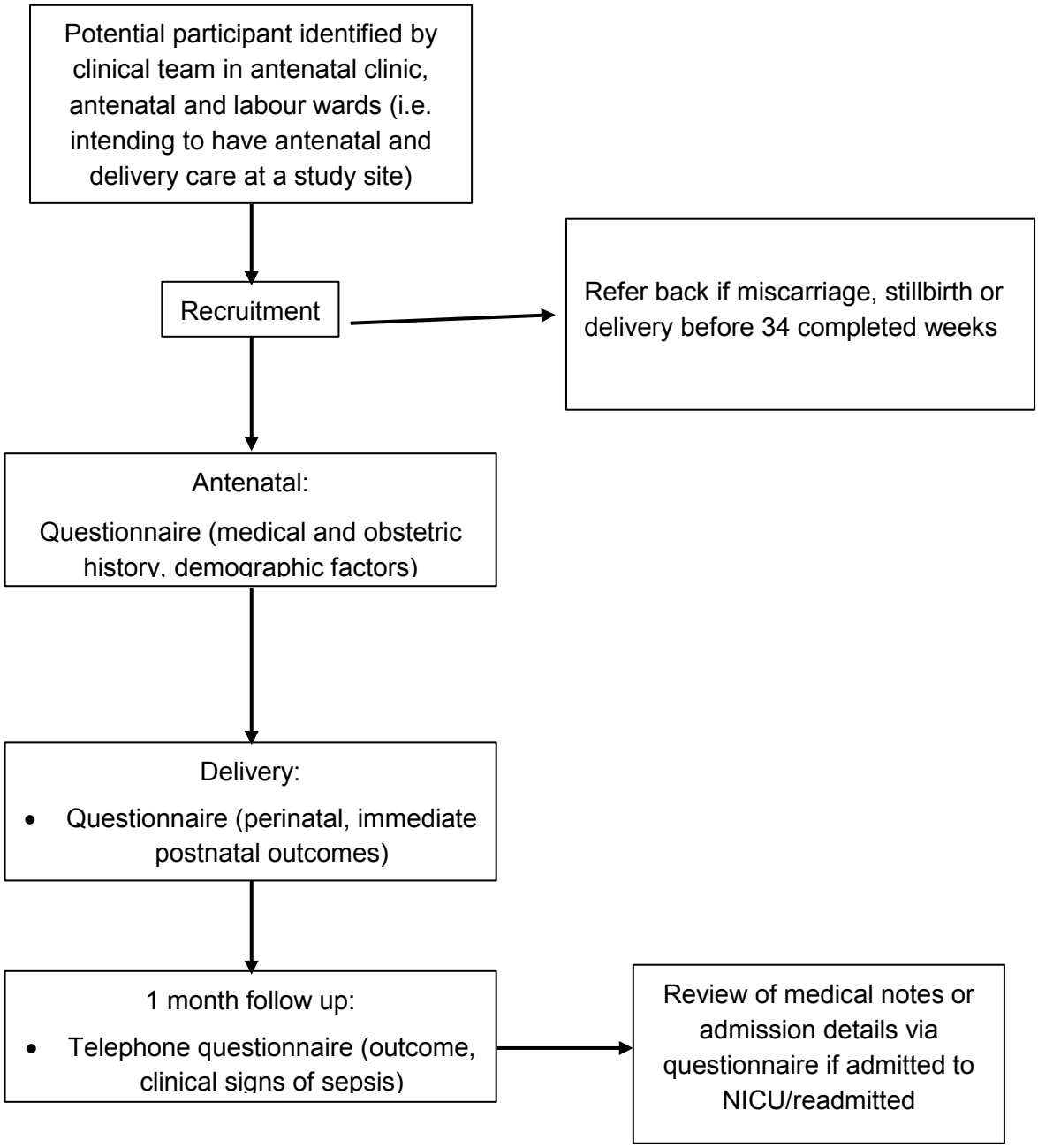
As above, the protocol has been designed in collaboration with the Community Engagement team who represent members of the community, patients, and carers.

STUDY Schematic: Phase 1 (Uganda)



Phase 2 comprises analysis and modelling of data collected in Phase 1.

STUDY Schematic: Phase 3 (Zimbabwe)



Using machine learning to model early-onset neonatal sepsis risk in Uganda (NeoRisk)

STUDY PROTOCOL

Using machine learning to model early-onset neonatal sepsis risk in Uganda

1 BACKGROUND

Sepsis is a leading cause of neonatal mortality worldwide(1), affecting 39.3 per 1000 live births in low- and middle-income countries(2). Mortality from early-onset sepsis can be reduced with prompt, correct treatment with antibiotics(3); therefore, identifying those at risk early is crucial. The sepsis rate in Uganda is high, with an incidence of 68 per 1000 live births, and is highest in hospital facilities(4).

Antibiotics are recommended for neonates with specific risk factors for sepsis, and those with signs of illness, including poor feeding and drowsiness(5). However, many low-resource settings diverge from the World Health Organization (WHO)'s recommended regimens, often due to antimicrobial resistance (5–7) or drug availability(8). Ideally, antibiotics should be given within an hour of birth in cases of suspected early-onset sepsis, which accounts for almost ¾ of sepsis cases in Uganda(4,9).

Additionally, risks of antimicrobial resistance are increasing. Neonates with sepsis in low- and middle-income countries show significant rates of resistance to the extent that the majority of proven infections are resistant to first and second line WHO-recommended antibiotics(10,11). Mortality from sepsis is double with resistant versus sensitive organisms(12). Previous studies in Kampala have shown high rates of resistance to first-line antibiotics(13). Prolonged antibiotic therapy without proven infection is associated with morbidity and mortality in preterm infants, and may affect normal flora and immunity(14,15). Therefore, antimicrobial stewardship is vital for low-resource settings. Accurate, early identification of at-risk neonates allows the benefits of early treatment whilst minimising the harms of overtreatment, and unnecessary admission.

Risk stratification systems are established tools in medicine, including in sub-Saharan Africa(16,17), to predict mortality in paediatric inpatients(18) and neonates(19), but previously tested systems for prediction of neonatal sepsis in low-resource settings show modest performance(20). This may be due to limited laboratory testing, so systems are built at least partially on clinician defined sepsis rather than the gold-standard of blood culture diagnosis(20). In high resource settings, risk stratification systems have focused on reducing overtreatment and antibiotic resistance(14,15,21). Neonatal sepsis risk calculators (a risk stratification system with a digital interface), such as the Kaiser Permanente (KP) Early-Onset Sepsis (EOS) calculator, have been adopted in the USA and UK(22). This calculator uses local incidence of EOS, gestational age, time from rupture of membranes (ROM) to delivery, highest maternal temperature during labour, maternal group B Streptococcus status, and type and timing of antepartum antibiotics(23).

In low-resource settings, the consequences of missing neonatal sepsis are more severe due to limited higher acuity care(24). Therefore, sepsis risk assessment in a low-resource setting should prioritise sensitivity over specificity, although prioritisation of limited resources for true cases is also helpful(25). Existing risk calculators may be inappropriate to use in a

low-resource setting; background incidence of EOS is much higher in these settings(1,2). Other clinical and non-clinical factors may mediate risk, such as HIV status or birth location. Few of the usual risk factor inputs are available in a low-resource setting. For example, without adequate antenatal care and imaging, gestational age may be uncertain, and without skilled birth attendants, duration of ROM may be unknown. Finally, diagnostic facilities for sepsis such as biochemical tests for inflammatory markers may be limited or unavailable. This makes identification of those at-risk using clinical factors, as opposed to laboratory testing, even more important. Additionally, lack of access to blood cultures has limited the ability of previous models to include this 'gold standard' of diagnosis in their modelling(20).

2 RATIONALE

Current risk stratification systems are not able to predict neonatal sepsis risk in a low-resource setting with sufficient accuracy, as outlined in section 1. Reliable prediction of neonatal sepsis risk would enable clinicians to prioritise early broad-spectrum antibiotics and close monitoring for the highest-risk neonates, reducing associated morbidity and mortality, and allow earlier discharge of low-risk neonates without antibiotic treatment, reducing the economic and psychological burden to families of remaining in hospital and antimicrobial overuse. In a wider context, the ability to predict sepsis risk more accurately would also aid the development of clinical pathways and service planning within maternity and neonatal services. Finally, the resulting model and code from this study could be used by other researchers in the future to model risk of other infective conditions, or neonatal sepsis in other settings.

To build a model for risk stratification, data are required that encompass not only a broad range of potential risk factors for sepsis, but also accurate outcome data confirming or refuting the diagnosis of sepsis. Previous attempts to create risk stratification systems have lacked blood cultures as the gold standard diagnostic method for sepsis(20), so this study will include a blood culture sample from every participant. This study will also include a range of patient-reportable potential risk factors to maximise the clinical utility of the model in the future.

3 THEORETICAL FRAMEWORK

Aetiology of neonatal sepsis

Neonatal sepsis is commonly defined as a severe systemic infection occurring in the neonatal period. Early-onset neonatal sepsis (EOS) refers to those infections presenting within 72 hours of life and is usually the result of vertical acquisition of the pathogen. This is in contrast to late-onset neonatal sepsis which presents after 72 hours of life usually results from horizontal acquisition of the pathogen (i.e. from the neonate's environment)(26).

Many risk factors are common to both early- and late-onset sepsis, but their importance differs for each of the conditions. EOS relates more to maternal risk factors(26–29), and LOS to neonatal factors(26,27,29)(. Environmental factors will also influence the risk of LOS, as they will affect the risk of horizontal transmission of pathogens(26,27). As this study aims to create a risk stratification system relevant for low-resource settings, it is important to consider patient-

reportable or easily measurable factors that may modify the risk of neonatal sepsis. These have been considered systematically by examining possible related maternal, neonatal, and environmental factors and their associations.

Patient-reportable risk factors

To assess many of the established risk factors for neonatal sepsis, mothers must receive skilled antenatal and delivery care; for example, a diagnosis of chorioamnionitis would require assessment by a midwife or doctor and may also require measuring of vital signs and laboratory tests to confirm the diagnosis. As the purpose of this project is to statistically model risk factors for early-onset neonatal sepsis with the aim to help inform clinical decision making in a low-resource setting, it is necessary to consider the association between neonatal sepsis and factors which are either patient-reportable or easily measurable without the need for laboratory or radiological services that may be unavailable in these settings.

Therefore, in designing the data collection instrument for this study, we have considered which patient-reportable factors might be strongly associated with the established sepsis risk factors above, and determined the following fields for data collection, listed in the table below:

Category	Risk factor	Data collection field	Established	Investigational
Maternal factors prior to pregnancy	Age	(years)	X	
	Tribe			X
	Medical comorbidities increasing infection risk		X	
		HIV	X	
		Chronic infection e.g. hepatitis, tuberculosis	X	
		Regular immunosuppressive medications e.g. steroids	X	
		Previous abdominal surgery		X
	Obstetric history		X	
		Number of previous pregnancies	X	
		Number of previous live deliveries		X
		Number of previous neonatal deaths	X	
	Low socioeconomic status		X	
		Maternal and partner's occupation		X
		Location of dwelling		X
		Construction of dwelling		X
		Occupancy of dwelling, including number of adults, children, and animals		X

		Source of lighting		X
		Maternal and partner's educational level		X
		Cooking facilities at home e.g. open fire indoors, gas stove		X
		Access to refrigeration at home		X
		Food insecurity		X
		Presence or absence of debt		X
	Water, sanitation and hygiene		X	
		Type of sanitation available at home (type of toilet/latrine)		X
		Source of drinking water		X
		Access to handwashing facility at home		X
	Nutritional status		X	
		Mid upper arm circumference	X	
		BMI	X	
		Dietary quality		X
	Substance use	E.g. recreational drugs, smoking, alcohol	X	
Maternal factors during pregnancy	Vaginal colonisation		X	
		Abnormal discharge during pregnancy		X
		Vulval or vaginal discomfort/pruritus		X
		Dyspareunia or postcoital bleeding		X
		Regularity of menstrual cycle		X
		Urinary tract symptoms during pregnancy	X	
		Diarrhoea or vomiting during pregnancy	X	
		Access to sanitation at home		X
		Antibiotic use during pregnancy		X
		Invasive procedures during or prior pregnancy e.g. cervical cerclage, surgical termination of pregnancy	X	
	Insufficient or late antenatal care		X	
		Number of antenatal appointments attended		X
		Number of antenatal ultrasound scans		X
		Screening blood tests taken antenatally		X

		Provision of antenatal vitamins		X
		Vaccinations given during pregnancy	X	
	Chorioamnionitis		X	
		Clinical symptoms e.g. fever, rigors, malaise, vomiting, abnormal discharge	X	
		Abnormal vital signs (maternal)	X	
		Foetal tachycardia or abnormal CTG	X	
		Reduced foetal movements	X	
		Duration of rupture of membranes	X	
		Duration of labour		X
		Foul-smelling, purulent, or meconium-stained liquor	X	
	Iatrogenic introduction of pathogen		X	
		Invasive procedures during labour e.g. vaginal examination, invasive foetal monitoring, catheterisation	X	
		Location of delivery (e.g. home or hospital) and hand hygiene policies		X
		Cadre of staff present at delivery		X
		Occupancy of ward (antenatal, labour, and neonatal); e.g. number of admissions per day, number of beds		X
		Number of staff on shift (nursing, midwifery, medical)		X
		Duration of hospital admission prior to delivery		X
		Number of attempts at cannulation/intubation/NG tube insertion for neonate		X
Neonatal factors	Low birthweight	Birthweight, head circumference	X	
	Sex		X	
	Gestational age at birth	By antenatal ultrasound scan where available; if not, use of birth weight (see inclusion criteria)	X	
	Congenital anomalies	e.g. Down syndrome	X	
	Mode of delivery	e.g. forceps, spontaneous vaginal	X	

	Low APGAR scores		X	
	Clinical signs of sepsis or meningitis within first 72h of life		X	
		Temperature instability	X	
		Poor feeding	X	
		Lethargy	X	
		Jaundice at <24h of life	X	
		Respiratory distress	X	
		Tachycardia	X	
		Tachypnoea	X	
		Poor perfusion	X	
		Abnormal tone	X	
	Umbilical cord care	e.g. application of herbal remedies		X
	Method of feeding	Breast versus formula	X	
		Water used to make up formula (e.g. filtered, tap)		X

Diagnosis of neonatal sepsis

The 'gold standard' for diagnosis of neonatal sepsis is widely considered to be a microbiological culture of neonatal blood(30), using at least a 1 ml sample(31). However, the proportion of neonatal patients with clinical signs of sepsis whose blood cultures are returned as negative can be significant, with studies reporting between 6 and 16 cases of 'culture-negative sepsis' for every culture-positive case(14). Therefore, although the use of blood culture is the gold standard, other diagnostic methods must be considered in order to capture all cases of neonatal sepsis and subsequently develop an accurate statistical model.

Various proxy biochemical and haematological markers of sepsis are commonly measured and used in clinical practice in high-resource settings, including C-reactive protein, procalcitonin, and full blood counts. These markers are all useful in making a diagnosis of sepsis and monitoring response to treatment, but none have sufficiently high sensitivity or specificity to be used in isolation(30). Additionally, in low-resource settings (including those settings in which this study will take place), access to consumables and laboratory facilities to measure these markers is unreliable. For this study, the use of additional biochemical and haematological markers has therefore been deemed to have limited usefulness in contributing to the statistical model, and so will not be used.

Clinical diagnosis of sepsis is challenging due to the nonspecific signs exhibited by neonates(32) and the lack of a universal definition of clinical sepsis(33). Signs of sepsis are also present in neonates without infection who are transitioning to extra-uterine life, or suffering from other conditions(32,34). However, given the near-universal availability of clinical diagnosis and the limitations of blood cultures to identify all cases of sepsis, we feel it is important to include clinical diagnosis of sepsis as a secondary outcome. Given the subjective nature of an individual clinician's diagnosis, we will also include established criteria for the diagnosis of sepsis that also assess the level of confidence in the clinical diagnosis, to reflect the spectrum of diagnostic certainty. We will therefore use the GAIA neonatal infection case definition criteria to assess cases, which are intended for international use(35,36).

Use of machine learning for risk stratification modelling

Machine learning is a field with increasing prominence, where computer systems are used to create algorithms to represent a data set(37). Evaluation of risk factors for neonatal sepsis could be achieved using traditional statistical techniques, as has been done for risk stratification models such as the Kaiser Permanente Sepsis Calculator(38). However, machine learning techniques allow increased flexibility and are better suited to factors with a non-linear relationship to the outcome(39), in addition to their ability to analyse large and complex data sets(40). Additionally, machine learning techniques may be able to discover novel associations and relationships that, although still requiring human validation as biologically plausible, may not have been discovered by traditional human-led analysis.

Therefore, a machine learning approach could enable the creation of a more accurate risk stratification model, making it the appropriate method to use for statistical analysis in this study. The success of machine learning methods is dependent on having sufficient data to train the model, and the ability to test it on external datasets. This has been accounted for in the design of this study by optimising the quality and accuracy of the training dataset (using a study site where blood culture results are reliably available), before testing the model in a second population to provide an external dataset.

4 RESEARCH QUESTION/AIM(S)

Can machine learning, using a broad range of variables including parent reported risk factors, be used to predict the risk of neonatal early-onset sepsis with greater accuracy than current local guidelines in low-resource settings (LRS)?

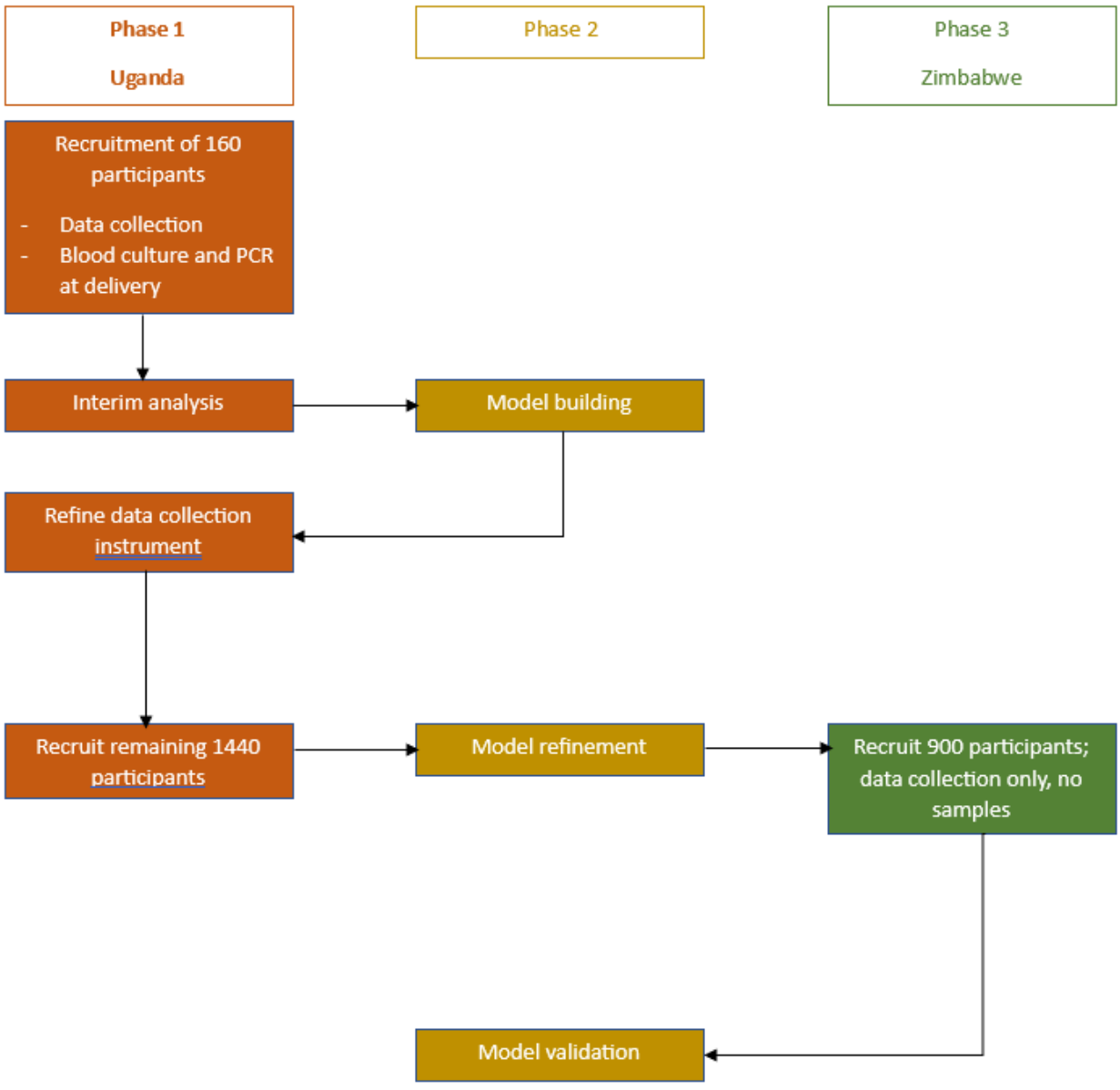
4.1 Objectives

1. Determine risk factors for neonatal early-onset sepsis in Kampala, Uganda
2. Use machine learning techniques to create a risk stratification model for early-onset sepsis in Uganda
3. Explore relationship between culture-positive sepsis and clinical diagnostic features in Uganda using machine learning techniques, including association between level of training of clinician documented diagnosis and blood culture result
4. Validate the model from Uganda externally using an independent dataset from a neonatal population in Harare, Zimbabwe

4.2 Outcome

- Primary outcome for Phase 1: Positive blood culture amongst neonatal participants
- Primary outcome for Phase 3: Clinician (specialist neonatal or paediatric doctor) diagnosis of sepsis amongst neonatal participants as documented in clinical notes or discharge letter
- Clinical features of neonatal sepsis as listed in GAIA neonatal infections case definition
- Admission to neonatal unit
- Readmission to hospital in first 28 days amongst neonatal participants
- Death before 28 days of age
- Sensitivity, specificity, positive and negative predictive value of risk stratification model for participants in Kampala and Harare

5 STUDY DESIGN and METHODS of DATA COLLECTION AND DATA ANALYSIS



There will be three main phases to this study, which is an observational cohort study.

Phase 1 will take place at Kawempe Hospital in Kampala, Uganda. Participants will be recruited from labour wards by study staff, and interim analysis will be completed after recruitment of 10% of the total sample. See section 7.3 for further details.

Data will be collected via an online CRF, including the factors listed in the table in Section 3.

Additionally:

- All neonates within the study will have a **neonatal venous blood sample** taken for culture by designated study staff
    - This must be taken before administration of antibiotics or within 48 hours of birth, whichever is sooner, and regardless of whether the neonate is being investigated and treated for sepsis.
- **Postnatal information** will be collected via telephone at 1 month including maternal and infant outcome, hospital readmission or clinical illness, and vaccines given

All data will be checked for quality control by a member of the study team.

A blood culture will be taken from all neonatal participants at birth, prior to starting antibiotic treatment wherever safe to do so. The volume of blood taken for culture must be 1-2ml and will be verified by weighing blood culture bottles. Training will be provided to paediatric and study teams on blood culture sampling methods, including adequate cleaning of the skin prior to sampling. Feedback on sample volumes and contamination rates (see section 7.3 for definition) will be fed back to paediatric consultants, and additional staff training offered. Where sample is remaining after 3ml is added to the culture bottle, the excess will be retained as a blood spot sample for use in future ethically approved research, where the participant has consented to this.

The results of the microbiological tests will be communicated to the participant's clinical team in real time – see section 7.3.4 for further details. Clinician diagnosis of sepsis will be recorded where a clinician of registrar level or above has recorded a diagnosis of sepsis within 48 hours of birth. To reduce the subjectivity of this measure, GAIA level of certainty for a diagnosis of sepsis will also be assessed (35) and the level of diagnostic certainty recorded (1-5, as per GAIA criteria). Medical records are recorded on paper at Kawempe Hospital, before being coded into an electronic medical record by data clerks. PDF scans of medical records will be available where detailed review is required.

All participant data will be collected by trained study staff, and recorded into an online CRF hosted on REDCap, a secure web application. No patient identifiable data will be stored on the main REDCap database. See section 8.6 for further details of data management plans.

The postnatal information will be completed via a telephone call at 28 days post-delivery, made by study staff. Participants' contact number, a contact number for their partner, and a contact number for a second friend or family member will be taken at recruitment to minimise loss to follow up due to changes to mobile telephone number. Participant identity will be confirmed using three factors (date of birth, full name, address/village). The following information will be collected:

- Health status of mother and baby
- Any visits to medical care for the neonate
    - If attended or admitted to Kawempe hospitals, these electronic medical records will be reviewed by study staff

- If attended elsewhere, the mother will be asked the reason for presenting, whether the neonate was admitted to hospital, the main diagnosis, and any treatments given.

Recruitment of the first 160 participants from labour ward (10% of the overall sample) will be treated as a pilot, and both recruitment and data will be reviewed at this point with particular attention to the following aspects:

- Demographic comparison between recruited women and the population attending Kawempe
- Data completion
- Volume of blood culture samples taken
- Loss to follow up rates
- Association between potential risk factors and clinical sepsis, or culture-positive sepsis

After this pilot analysis, the data collection instrument may be adjusted to be more concise if some risk factors appear to have no relationship with neonatal outcomes. Additionally, the recruitment strategy may be adjusted to include women attending antenatal clinics to improve the representativeness of the sample if the pilot sample is significantly different demographically to the general population of pregnant women attending the same study site antenatally. Finally, feedback and staff training will be given as needed to ensure optimal sample volumes and data completeness.

During Phase 1, two separate focus group discussions with 10 participants each will be held with staff from Kawempe Hospital. One discussion will be with senor clinicians, and one with junior clinicians and midwives. The focus groups will be held on site at Kawempe in English. Focus groups are being held in order to complement the quantitative results of this project by highlighting risk factors currently perceived as significant (which should then be a key part of analysis and results dissemination to confirm their usefulness), inform of current diagnostic difficulties to guide future work with the risk stratification model, and provide context for development and communication of the model to maximise its utility. The focus groups will address the following issues:

- Ascertain knowledge and awareness of neonatal sepsis
- Management of neonatal sepsis
- regarding risks and consequences of neonatal sepsis, and
- Current diagnostic strategies and their perceived strengths and difficulties
- Perceptions of neonatal sepsis and how this might affect diagnosis and treatment
- Recommendations to enhance the diagnosis and treatment of neonatal sepsis

The focus groups will be conducted by members of MUJHU's social sciences team and Dr Sarah Sturrock. A tool has been developed by the social sciences team and Dr Sturrock for these discussions (see attached). The discussions will be recorded, and recordings destroyed once the discussions have been transcribed. The social sciences team will then complete thematic analysis of the discussions.

We anticipate Phase 1 will take place between January and July 2024.

Phase 2 will use the data collected in phase 1 to construct a risk stratification model, in collaboration with the Advanced Research Computing Centre at University College London.

We anticipate Phase 2 will take place between March and September 2024.

Phase 3 will involve external validation of the risk stratification tool. This phase will take place at Sally Mugabe Central Hospital in Harare, Zimbabwe. Participants will be recruited from antenatal clinics and labour wards by study staff.

The same data will be collected from participants in Phase 3 as in Phase 1. Ideally, data collection will take place via the Neotree platform(41) – this will be decided once the factors included in the model have been determined, and whether they are already in routine collection via Neotree. If the Neotree platform will not capture sufficient data to match the inputs of the model, an online CRF will be created as in Phase 1 via REDCap. This decision will be made in collaboration with the local study team prior to the start of recruitment.

The primary outcome for Phase 3 will be senior clinician diagnosis of sepsis recorded within 48 hours of delivery, and blood cultures (when available) will only be taken in neonates judged to be at risk of, or with clinical signs of sepsis as per standard clinical guidelines. Using senior clinician diagnosis of sepsis as the outcome is because supply of blood culture bottles and laboratory analysis of cultures is unreliable and is unlikely to be available to all study participants as in the Phase 1 setting. The risk stratification model will be applied to these data to determine its accuracy in a second population and explore its generalisability.

We anticipate Phase 3 will take place between September 2024 and March 2025.

The TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement have been followed in the design of this study and will be followed in its reporting(42). Adjustments will be made as necessary with the release of upcoming artificial intelligence-specific guidelines.

Data analysis

Phase 2 will constitute the main data analysis for data from Uganda.

There will be two main streams to data analysis, which will be completed using R and in collaboration with the Advanced Research Computing Centre at UCL.

1. Machine learning analysis:
    a. Data from Phase 1 will form a ‘training data set’, with 20% reserved for testing and validation, and be fed into a pre-programmed machine
    b. Previous methods used for sepsis prediction that will be investigated include gradient tree boosting and neural networks
    c. Bootstrapping will be considered
    d. Determination of the exact machine learning model will depend on the data obtained, but the following will be considered:
        i. Weighted inputs: Allows scaling of measurements
        ii. Conformal prediction: Allows the machine to determine how uncertain its prediction is, based on distribution of data it has, and missing data
        iii. Recurrent neural networks: Allows the machine to recognise time-dependent patterns
        iv. Gradient descent: A method of ‘training’ a machine to minimise the error between prediction and reality by gradually adjusting its weights
2. Traditional logistic regression will be used to identify risk factors associated with clinical and microbiological diagnoses of neonatal sepsis. Although the principal model will be created using machine learning techniques, logistic regression will be

used as a second check of results and to provide additional context with which to interpret the machine learning model

For **Phase 3**, data will be inputted to the machine learning model from Phase 2, and the sensitivity, specificity, positive predictive and negative predictive values calculated, as well as the AUC value.

6 STUDY SETTING

Phase 1

Recruitment and data collection will take place at Kawempe Hospital in Kampala, Uganda. These sites are sufficiently busy to recruit the proposed sample size, have the necessary infrastructure to process blood cultures and have significant experience of running research studies within the labour and post-natal ward. Kawempe Hospital is a national referral hospital for high-risk pregnancies and local deliveries. Approximately 25,000 births occur at Kawempe annually, with approximately 11,000 neonatal admissions annually. Mulago Hospital is a large hospital with children's services(43). Electronic health records are currently used in both including antenatal and perinatal information, which are generated by digital data entry from paper hospital records. Blood cultures are currently processed at Makerere Microbiology Laboratories (College of American Pathologists accredited) using BACTEC machines within 12 hours of collection(43).

Phase 2

Data analysis and modelling will take place in collaboration with the Advanced Research Computing Centre at University College London.

Phase 3

Recruitment and data collection will take place at Sally Mugabe Central Hospital in Harare, Zimbabwe. Again, this site is large enough to recruit the proposed sample size, with 22,000 births annually, and has existing research infrastructure to enable quality data collection(41). The Neotree platform is in routine use at Sally Mugabe Central Hospital.

7 SAMPLE AND RECRUITMENT

7.1 Eligibility Criteria

7.1.1 Inclusion criteria

All neonates born at one of the study sites are eligible for recruitment. Where mothers have had an ultrasound scan during pregnancy during routine care, neonates will be eligible for inclusion if they are born at ≥34 weeks gestational age. For mothers who have not had an ultrasound scan during pregnancy or have uncertain dates, neonates will be eligible for

inclusion if their birth weight is ≥1400g which will include the majority of births ≥34 weeks according to the results of the INTERGROWTH-21st Project(44).

7.1.2 Exclusion criteria

Neonates born at <34 weeks gestational age will be excluded due to their higher background risk of sepsis, and automatic admission to the neonatal unit with antibiotic administration. Similarly, neonates with congenital abnormalities or syndromes associated with increased susceptibility to infection will be excluded (e.g. gastroschisis). Neonates whose mothers are unable to provide informed consent will also be excluded. There are no exclusion criteria on the basis of maternal medical or obstetric comorbidities.

7.2 Sampling

7.2.1 Size of sample

Phase 1: There is currently no accepted way to power a machine learning study, the consensus being that more data will always create a better model. However, to ensure the success of this study, we have investigated approximate power to create a diagnostic test to estimate the sample size needed for this study. We have also reviewed this sample size calculation with our collaborators at the UCL Advanced Research Computing centre.

A study of a diagnostic test with 90% sensitivity and a background sepsis rate of 5% (the estimated rate of culture-positive sepsis in Kampala based on previous studies is 12.8%(13) amongst those with clinical signs of sepsis, and up to 5% amongst all neonates according to unpublished data from previous studies at Kawempe) would require 1460 participants (95% confidence interval), as described by Buderer and Jones *et al(45,46).*

This sample size has also been calculated according to three suggested methods for calculating the sample size for a clinical prediction model:

- Calculation of size needing to provide precise estimate of overall outcome risk
- Calculation of size needed to produce predicted values with a small mean error across all individuals
- Size needed considering shrinkage of predictor effects

All of the above methods produced sample sizes from 140-1532, hence the size of 1600 should be sufficient even with some loss to follow up. If, at interim analysis, the rate of culture-positive sepsis is significantly lower, we will explore methods to increase our sample size such as using existing data to create simulated data to train our model.

Phase 3: The sample size will be 900 mothers; this is powered to give 95% confidence for a diagnostic test which has 90% sensitivity with a background rate of clinical sepsis of 16%, using the same calculation method as for the sample size in Phase I. Although there are limited previous studies of neonatal sepsis prevalence in Zimbabwe using blood cultures, particularly amongst all neonates, one study found a culture-positive sepsis rate of 21% amongst NICU admissions(47), and a recent study reported a clinical diagnosis rate of 40.8%(48) . Therefore, this sample size should give an estimated 367 episodes of clinician-diagnosed sepsis, deemed by Collins *et al*. to be sufficient for external validation of a prognostic model(49).

7.2.2 Sampling technique

Convenience sampling will be used for this study. Potential participants will be identified by local clinical staff, and then approached by the study team – see section 7.3 for further details of recruitment procedures.

7.3 Recruitment

7.3.1 Participant identification

Study sensitisation will occur in the antenatal clinic at Kawempe, Mulago and Sally Mugabe Central Hospitals. Posters will be available to assist with this, and local staff will be sensitised to the study and able to direct interested potential participants to study staff to answer any questions.

Study staff will approach potential participants (identified by local clinical staff) attending labour wards and antenatal clinics at participating sites, a recruitment technique which has been used successfully in previous studies at Kawempe Hospital. Patient information sheets, translated into local languages, will be provided, and posters and leaflets with study details will also be available on site. Participants in the second stage of labour (fully dilated) will not be approached for recruitment.

Following pilot analysis, if the sample obtained does not seem to be representative of the overall population of pregnant women delivering at the site (for example, if the recruited sample includes a higher percentage of women with uncomplicated deliveries), recruitment will be adjusted to include women attending antenatal clinics at study sites. Participants will be approached if they are attending an antenatal clinic at >34 weeks gestation and planning to attend antenatal and delivery care at a study site.

Normal antenatal and perinatal care will be provided at the same centre as recruitment, so no additional study visits should be required.

7.3.2 Consent

Informed consent will be obtained at the point of recruitment in Uganda, by trained study staff, for data collection and the taking of a sample. Consent must be given by the mother, who will have parental responsibility for the neonate. Wherever possible, consent should be taken from the mother and the father; where it is not possible to take consent from the father, an explanation for this will be entered into the CRF. Neonates born to mothers who die during or after labor must have informed consent taken from their father or caretaker.

Potential participants, including mothers/fathers/caretakers for the neonates, will be provided with an information leaflet and consent documents, in English or Luganda, and will have the opportunity to ask any questions about the study. They will also be allowed time to discuss the study with family members if they wish and will be given contact details for the study team. The study team are experienced with studies involving pregnant people and new parents, including those that involve consenting and data collection during labour.

If happy to proceed, the participant will sign an informed consent form, either in Luganda or English. A photocopy of the signed consent form will be retained by the local study team.

In Zimbabwe, consent procedures will be finalized once the risk stratification model has been constructed and the risk inputs needed to inform the model are known. If all risk inputs are part of the routine Neotree data set, then specific consent will not be sought. If the

model structure is such that additional information needs to be collected from participants, then specific consent will be sought to collect these data.

Consent provisions for collection and use of participant data and biological specimens

No patient identifiable data (name, date of birth, address, national ID, hospital number) will be stored on the main REDCap database – only pseudonymised data with study ID. Telephone numbers and addresses, to enable follow up, will be stored with patient ID and date of birth only, on a separate REDCap database accessible via an encrypted password protected laptop stored in a locked cabinet at the hospital site in Kawempe. These patient identifiable data will be destroyed once follow up is complete. Microbiological samples will be stored and managed according to SGUL and MUJHU's regulations. Consent will be sought from participants for storage of samples for future ethically approved research.

Data from Zimbabwe will be stored securely in-country on a local server, in accordance with local regulations. Pseudonymised data only will be extracted for statistical analysis.

All data will be archived by SGUL after study completion. Pseudonymised data may be used for future research projects by the team at SGUL or collaborators, subject to ethical approval. Data will be stored after the completion of the study in SGUL-managed storage. File format conversion will be done to make the data suitable for long-term retention.

Participants will be able to withdraw from the research at any point – their identifiable data (such as contact details) would then be removed from the database on REDCap, but existing pseudonymised data retained for analysis.

Participants will have access to the results of the microbiological specimens, which will be communicated via their clinical team.

7.3.3 **Data collection tool**

Case Report Forms will be designed by the CI. All CRFs will be electronic. The CI will provide logins to relevant and trained site level members of the research team.

It is the Investigator's responsibility to ensure the accuracy of all data entered and recorded in the CRFs. The Staff Delegation of Responsibilities Log should identify all trial personnel responsible for data collection, entry, handling and managing the database.

Prospective patient-level data will be collected from electronic health records (and participant interviews where needed) into an online data entry system (REDCap). This is a secure web application for building and managing online surveys and databases which supports online and offline data capture. This data entry system has been chosen as it will improve the consistency of the data, and allows them to be immediately available in an

analysable format. The data are also stored securely without being stored on an individual hard drive, ensuring their safety and long-term access without need for physical backup.

These data will be anonymised, and will include the parameters outlined in section 5.

Microbiological samples will be taken from the baby, which will be processed locally for blood culture. Gram stain, organism and resistance profile will be inputted to REDCap.

Each data entry form will include the username of the member of study staff who filled out the form, and the date and time it was filled out. Each form will also include a field to input the date, time, and name of the person marking the form as 'complete'. The data entry forms will be tested by members of study staff before being put into production. Data will be reviewed monthly to ensure that data quality is sufficiently high.

Quality control mechanism

Once the study has begun, trained members of study staff will be given their own logins to REDCap to input data. Study staff will also be able to contact the research team to resolve any queries about how or where to input data. Data will either be collected directly from research participants or from their electronic hospital records, but will be inputted directly into REDCap. Data will be reviewed by two study staff members to check for data completeness with reference to scanned medical records.

All data described above will be inputted into a password protected database designed using REDCap. The database will be stored on a secure server at MUJHU. Access will be strictly controlled and for the study investigators only. The data will be kept on the database server and a file backup server. There will be encrypted back up tapes. These are stored in a secure storage facility with 24-hour security. All computers are connected to a local area network that is protected by firewalls operating to accepted standards and are protected by antivirus software. No patient identifiable information is stored on desktop or portable media.

7.3.4 **Biological Sample Handling**

Blood samples will be taken from each neonate for culture (1-2ml) into BACTEC blood culture bottles and stored at room temperature until arrival at the laboratory. Sample volume will be verified by weight, and samples will be labelled with the neonate's ID.

Samples will be processed at Makerere Microbiology Laboratories (College of American Pathologists accredited) within 12 hours of collection(43). Samples will be incubated for 36 hours. Any presumptive positive culture will be plated onto selective agar and identified by biochemical tests. Couriers will collect samples from study sites and transport them to the Medical and Molecular Laboratories (MML) at Makerere University; coverage is available at weekends and some evenings by request from study staff, so a courier will be arranged if a sample is taken more than 12 hours before the next scheduled courier collection.

Sample results will be classified and handled as follows:

- Positive result: known potential pathogenic bacteria

  - Pathogens included as per GAIA neonatal infection outcome list of 'recognised pathogens'(35)
  - Study staff will notify clinical team and participant's parents by telephone; written notification also to be provided from laboratory (either via electronic system or hard copy) as per routine care
- Contaminant: growth of species considered 'non pathogenic' by GAIA list
  - Review clinical notes to determine presence of risk factors for invasive disease as a result of a normally commensal/contaminant organism (e.g. indwelling venous catheters causing risk of coagulase negative staphylococcal sepsis)
  - If no risk factors, written notification provided to clinical team from laboratory as above
  - If presence of risk factors or uncertainty, study team to additionally contact clinical team via telephone to notify
- Negative result: no growth
  - Written notification provided to clinical team from laboratory as above

Specific written consent will be taken from participants to store any leftover samples for use in future ethically approved research.

8 ETHICAL AND REGULATORY CONSIDERATIONS

8.1 Assessment and management of risk

Risks to participants

As this is an observational cohort study, the only risk to participants is the discomfort of blood sampling for culture. Venous blood sampling is a safe procedure but can cause temporary discomfort and requires the infant to be held still. Sampling will be done by trained members of study staff to minimise discomfort. The risk prediction model will not be shared with clinicians at study sites until study completion to avoid influencing treatment decisions.

There is also a risk of confidentiality loss in the case of any data breach. To minimise this risk, study data will be kept on a separate database to participant identifiable information and all data will be kept on secure, encrypted databases accessible only to study staff.

COVID-19

At present, there is not expected to be any increased risk of COVID-19 for participants or staff as a result of their involvement in the study. However, should this change, a COVID-19 risk assessment and management strategy is given below.

<u>COVID-19 Risk Assessment and Management Strategy</u>

All staff employed by SGUL and/or SGH NHS Foundation Trust are required to complete an ongoing COVID-19 risk assessment prior to undertaking any work on site, which includes research activity. This process is continuously monitored by the responsible line manager.

Participants (unaffected or affected) will not be recruited if they are deemed high risk or are in close contact with someone at risk. The Research Team will contact research participants ahead of scheduled study visits on-site to check for COVID-19 symptoms and the symptom check will be repeated when patients attend the hospital site for the study visit.

Participants will receive information regarding the extra precautions that will be taken in light of the COVID-19 pandemic in the Patient Information Sheet. This will detail steps that patients should take if they have concerns about exposure to COVID-19 through participating in the research, or believe that they are symptomatic or have been in close contact with another person believed to be symptomatic. The Patient Information Sheet will also have contact details for the Research Team for patients to get in touch if they have any concerns or queries about this.

All research personnel are expected to comply with the NHS Trust and University policies on COVID-19.

All patients attending the hospital site for research visits and/or routine clinical follow-up will be expected to abide by the NHS Trust and University policies on COVID-19 which include wearing suitable PPE (provided by the NHS Trust on arrival), adhering to the visitor policy on social distancing and following the one-way routing systems whilst on site.

The schedule of study assessments has been designed so that they align with the current routine clinical pathway for this patient population, and follow up can be done remotely via telephone call.

Therefore, research participants and site staff are not perceived to be at any additional risk of exposure to COVID-19 through participation in this research study.

8.2 Research Ethics Committee (REC) and other Regulatory review & reports

Before the start of the study, a favourable opinion will be sought from an appropriate REC for the study protocol, informed consent forms and other relevant documents e.g. advertisements.

Regulatory Review & Compliance

Before any site can enrol patients into the study, the Chief Investigator/Principal Investigator or designee will ensure that appropriate approvals from participating organisations are in place. Specific arrangements on how to gain approval from participating organisations are in place and comply with the relevant guidance.

Amendments

For any amendment to the study, the Chief Investigator or designee, in agreement with the sponsor will submit information to the appropriate body in order for them to issue approval for the amendment. The Chief Investigator or designee will work with sites (R&D departments at NHS sites as well as the study delivery team) so they can put the necessary arrangements in place to implement the amendment to confirm their support for the study as amended.

8.3 Peer review

This protocol has been peer reviewed by two independent reviewers with expertise in infectious disease and epidemiology.

8.4 Patient & Public Involvement

The study has been reviewed with the Community Advisory Board at MUJHU Care Ltd at an early stage, as well as with local researchers with experience conducting research in Kampala and Harare. As a result, the protocol involves no study visits in-person outside of routine clinical care, as many patients travel long distances in order to attend the study sites for antenatal and labour care.

Further PPI activities will take place in-country, including:

- Paper-based surveys for patients to ascertain knowledge and understanding of neonatal sepsis
- Focus groups with local clinicians to assess opinions on the use of risk stratification systems for clinical decision making
- Regular meetings with the Community Advisory Board

8.5 Protocol compliance

Protocol deviations, non-compliances, or breaches are departures from the approved protocol.

All protocol deviations must be adequately documented on the relevant forms and reported to the REC, Chief Investigator and Sponsor immediately.

Deviations from the protocol which are found to frequently recur are not acceptable, will require immediate action and could potentially be classified as a serious breach.

8.6 Data protection and patient confidentiality

All data should be handled in accordance with the Data Protection Act 2018 (UK implementation of the EU General Data Protection Regulation (GDPR)) and local regulations in Uganda and Zimbabwe.

Any Case Report Forms (CRFs) will not bear the participant's name or other directly identifiable data. The participant's trial Identification Number (ID) only, will be used for identification. The Subject ID log can be used to cross reference participant's identifiable information.

8.7 Indemnity

St George's University of London sponsored research:
As this is a non-interventional, observational study and in accordance with local and SGUL guidelines, insurance is not required to cover injury as a result of the study.

Where the Trial is conducted in a hospital, the hospital has a duty of care to participants. St George's University of London will not accept liability for any breach in the hospital's duty of care, or any negligence on the part of hospital employees.

8.8 Access to the final study dataset

Aggregated and coded anonymised data only will be suitable for sharing.

All data requests will be handled by the SGUL Research Data Management Service, alongside project and university representatives, in accordance with university policies. Data access will be granted for research use, dependent on relevant ethical and institutional approval and only pseudonymised data. Data will only be available to the research team until the data are published, as soon as possible after the end of the study and within 6 months of close of the study.

To enable the anonymised data to be discovered and responsibly accessed by academic parties, it will be publicised through networks in the UK, Uganda and Zimbabwe, presentations, and peer reviewed publications. Metadata will be openly available with instructions on how these data can be accessed via the SGUL Research Data Repository.

9 DISSEMINATION POLICY

9.1 Dissemination policy

Publication: "Any activity that discloses, outside of the circle of trial investigators, any final or interim data or results of the Trial, or any details of the Trial methodology that have not been made public by the Sponsor including, for example, presentations at symposia, national or regional professional meetings, publications in journals, theses or dissertations."

All scientific contributors to the Trial have a responsibility to ensure that results of scientific interest arising from Trial are appropriately published and disseminated. The Sponsor has a firm commitment to publish the results of the Trial in a transparent and unbiased manner without consideration for commercial objectives.

To maximise the impact and scientific validity of the Trial, data shall be consolidated over the duration of the trial, reviewed internally among all investigators and not be submitted for publication prematurely. Lead in any publications arising from the Trial shall lie with the chief investigator in the first instance.

Before the official completion of the Trial,

All publications during this period are subject to permission by the Sponsor. If an investigator wishes to publish a sub-set of data without permission by the Sponsor during this period, the **Steering Committee/the Funder** shall have the final say.

Exempt from this requirement are student theses that can be submitted for confidential evaluation but are subject to embargo for a period not shorter than the anticipated remaining duration of the trial.

Up to 180 days after the official completion of the Trial

During this period the Chief Investigator shall liaise with all investigators and strive to consolidate data and results and submit a manuscript for peer-review with a view to publication in a reputable academic journal or similar outlet as the Main Publication.

- The Chief Investigator shall be senior and corresponding author of the Main Publication.
- Insofar as compatible with the policies of the publication outlet and good academic practice, the other Investigators shall be listed in alphabetic order.
- Providers of analytical or technical services shall be acknowledged, but will only be listed as co-authors if their services were provided in a non-routine manner as part of a scientific collaboration.
- Members of the Steering Group shall only be acknowledged as co-authors if they contributed in other capacities as well.
- If there are disagreements about the substance, content, style, conclusions, or author list of the Main Publication, the Chief Investigator shall ask the Steering Group to arbitrate.

Beyond 180 days after the official completion of the Trial

After the Main Publication or after 180 days from Trial end date any Investigator or group of investigators may prepare further publications. In order to ensure that the Sponsor will be able to make comments and suggestions where pertinent, material for public dissemination will be submitted to the Sponsor for review at least sixty (60) days prior to submission for publication, public dissemination, or review by a publication committee. Sponsor's reasonable comments shall be reflected. All publications related to the Trial shall credit the Chief and Co-Investigators as co-authors where this would be in accordance with normal academic practice and shall acknowledge the Sponsor and the Funders.

9.2 Archiving Arrangements

Each site will be responsible for their onsite level study archiving. The trial essential TMF along with any central trial database will be archived in accordance with the sponsor SOP.

10 REFERENCES

1. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet. 2016 Dec;388(10063):3027–35.

2. Velaphi SC, Westercamp M, Moleleki M, Pondo T, Dangor Z, Wolter N, et al. Surveillance for incidence and etiology of early-onset neonatal sepsis in Soweto, South Africa. Larson SD, editor. PLoS ONE. 2019 Apr 10;14(4):e0214077.

3. Schmatz M, Srinivasan L, Grundmeier RW, Elci OU, Weiss SL, Masino AJ, et al. Surviving Sepsis in a Referral Neonatal Intensive Care Unit: Association between Time to Antibiotic Administration and In-Hospital Outcomes. The Journal of Pediatrics. 2020 Feb;217:59-65.e1.

4. Migamba SM, Kisaakye E, Komakech A, Nakanwagi M, Nakamya P, Mutumba R, et al. Trends and spatial distribution of neonatal sepsis, Uganda, 2016–2020. BMC Pregnancy Childbirth. 2023 Nov 4;23(1):770.

5. World Health Organization. Pocket Book of Hospital Care for Children, 2nd ed. 2013.

6. Li G, Bielicki JA, Ahmed ASMNU, Islam MS, Berezin EN, Gallacci CB, et al. Towards understanding global patterns of antimicrobial use and resistance in neonatal sepsis: insights from the NeoAMR network. Arch Dis Child. 2020 Jan;105(1):26–31.

7. Knowles R, Sharland M, Hsia Y, Magrini N, Moja L, Siyam A, et al. Measuring antibiotic availability and use in 20 low- and middle-income countries. Bull World Health Organ. 2020 Mar 1;98(3):177-187C.

8. Mendelson M, Røttingen JA, Gopinathan U, Hamer DH, Wertheim H, Basnyat B, et al. Maximising access to achieve appropriate human antimicrobial use in low-income and middle-income countries. The Lancet. 2016 Jan;387(10014):188–98.

9. National Institute for Health and Care Excellence. Neonatal infection: antibiotics for prevention and treatment. 2021.

10. Sands K, Carvalho MJ, Portal E, Thomson K, Dyer C, Akpulu C, et al. Characterization of antimicrobial-resistant Gram-negative bacteria that cause neonatal sepsis in seven low- and middle-income countries. Nat Microbiol. 2021 Mar 29;6(4):512–23.

11. Thomson KM, Dyer C, Liu F, Sands K, Portal E, Carvalho MJ, et al. Effects of antibiotic resistance, drug target attainment, bacterial pathogenicity and virulence, and antibiotic access and affordability on outcomes in neonatal sepsis: an international microbiology and drug evaluation prospective substudy (BARNARDS). The Lancet Infectious Diseases. 2021 Dec;21(12):1677–88.

12. Kayange N, Kamugisha E, Mwizamholya DL, Jeremiah S, Mshana SE. Predictors of positive blood culture and deaths among neonates with suspected neonatal sepsis in a tertiary hospital, Mwanza- Tanzania. BMC Pediatr. 2010 Dec;10(1):39.

13. Tumuhamye J, Sommerfelt H, Bwanga F, Ndeezi G, Mukunya D, Napyo A, et al. Neonatal sepsis at Mulago national referral hospital in Uganda: Etiology, antimicrobial resistance, associated factors and case fatality risk. Butaye P, editor. PLoS ONE. 2020 Aug 10;15(8):e0237085.

14. Klingenberg C, Kornelisse RF, Buonocore G, Maier RF, Stocker M. Culture-Negative Early-Onset Neonatal Sepsis — At the Crossroad Between Efficient Sepsis Care and Antimicrobial Stewardship. Front Pediatr. 2018 Oct 9;6:285.

15. Popescu CR, Cavanagh MMM, Tembo B, Chiume M, Lufesi N, Goldfarb DM, et al. Neonatal sepsis in low-income countries: epidemiology, diagnosis and prevention. Expert Review of Anti-infective Therapy. 2020 May 3;18(5):443–52.

16. Kluyts HL, Le Manach Y, Munlemvo DM, Madzimbamuto F, Basenero A, Coulibaly Y, et al. The ASOS Surgical Risk Calculator: development and validation of a tool for identifying African surgical patients at risk of severe postoperative complications. British Journal of Anaesthesia. 2018 Dec;121(6):1357–63.

17. Biccard BM, Du Toit L, Lesosky M, Stephens T, Myer L, Prempeh AB, et al. Enhanced postoperative surveillance versus standard of care to reduce mortality among adult surgical patients in Africa (ASOS-2): a cluster-randomised controlled trial. The Lancet Global Health. 2021 Oct;9(10):e1391–401.

18. Mpimbaza A, Sears D, Sserwanga A, Kigozi R, Rubahika D, Nadler A, et al. Admission Risk Score to Predict Inpatient Pediatric Mortality at Four Public Hospitals in Uganda. Brock G, editor. PLoS ONE. 2015 Jul 28;10(7):e0133950.

19. Medvedev MM, Brotherton H, Gai A, Tann C, Gale C, Waiswa P, et al. Development and validation of a simplified score to predict neonatal mortality risk among neonates weighing 2000 g or less (NMR-2000): an analysis using data from the UK and The Gambia. Lancet Child Adolesc Health. 2020 Apr;4(4):299–311.

20. Neal SR, Fitzgerald F, Chimhuya S, Heys M, Cortina-Borja M, Chimhini G. Diagnosing early-onset neonatal sepsis in low-resource settings: development of a multivariable prediction model. Arch Dis Child. 2023 Aug;108(8):608–15.

21. Folgori L, Ellis SJ, Bielicki JA, Heath PT, Sharland M, Balasegaram M. Tackling antimicrobial resistance in neonatal sepsis. The Lancet Global Health. 2017 Nov;5(11):e1066–8.

22. Goel N, Shrestha S, Smith R, Mehta A, Ketty M, Muxworthy H, et al. Screening for early onset neonatal sepsis: NICE guidance-based practice versus projected application of the Kaiser Permanente sepsis risk calculator in the UK population. Arch Dis Child Fetal Neonatal Ed. 2020 Mar;105(2):118–22.

23. Kaiser Permanente. Probability of Neonatal Early-Onset Sepsis Based on Maternal Risk Factors and the Infant's Clinical Presentation. [Internet]. Available from: https://neonatalsepsiscalculator.kaiserpermanente.org

24. World Health Organization. WHO launches new roadmap on human resource strategies to ensure that all newborns survive and thrive [Internet]. 2020. Available from: https://www.who.int/news/item/17-11-2020-who-launches-new-roadmap-on-human-resource-strategies-to-ensure-that-all-newborns-survive-and-thrive

25. Bonniface M, Nambatya W, Rajab K. An Evaluation of Antibiotic Prescribing Practices in a Rural Refugee Settlement District in Uganda. Antibiotics. 2021 Feb 9;10(2):172.

26. Glaser MA, Hughes LM, Jnah A, Newberry D. Neonatal Sepsis: A Review of Pathophysiology and Current Management Strategies. Advances in Neonatal Care. 2021 Feb;21(1):49–60.

27. Simonsen KA, Anderson-Berry AL, Delair SF, Davies HD. Early-Onset Neonatal Sepsis. Clin Microbiol Rev. 2014 Jan;27(1):21–47.

28. Patras KA, Nizet V. Group B Streptococcal Maternal Colonization and Neonatal Disease: Molecular Mechanisms and Preventative Approaches. Front Pediatr. 2018 Feb 22;6:27.

29. Ocviyanti D, Wahono WT. Risk Factors for Neonatal Sepsis in Pregnant Women with Premature Rupture of the Membrane. J Pregnancy. 2018;2018:4823404.

30. Celik IH, Hanna M, Canpolat FE, Mohan Pammi. Diagnosis of neonatal sepsis: the past, present and future. Pediatr Res. 2022 Jan;91(2):337–50.

31. Huber S, Hetzer B, Crazzolara R, Orth-Höller D. The correct blood volume for paediatric blood cultures: a conundrum? Clinical Microbiology and Infection. 2020 Feb;26(2):168–73.

32. Wynn JL. Defining neonatal sepsis. Current Opinion in Pediatrics. 2016 Apr;28(2):135–40.

33. on behalf of the Infection, Inflammation, Immunology and Immunisation (I4) section of the ESPR, McGovern M, Giannoni E, Kuester H, Turner MA, Van Den Hoogen A, et al. Challenges in developing a consensus definition of neonatal sepsis. Pediatr Res. 2020 Jul;88(1):14–26.

34. Sands K, Spiller OB, Thomson K, Portal EA, Iregbu KC, Walsh TR. Early-Onset Neonatal Sepsis in Low- and Middle-Income Countries: Current Challenges and Future Opportunities. IDR. 2022 Mar;Volume 15:933–46.

35. Vergnano S, Buttery J, Cailes B, Chandrasekaran R, Chiappini E, Clark E, et al. Neonatal infections: Case definition and guidelines for data collection, analysis, and presentation of immunisation safety data. Vaccine. 2016 Dec;34(49):6038–46.

36. Munoz FM, Eckert LO, Katz MA, Lambach P, Ortiz JR, Bauwens J, et al. Key terms for the assessment of the safety of vaccines in pregnancy: Results of a global consultative process to initiate harmonization of adverse event definitions. Vaccine. 2015 Nov;33(47):6441–52.

37. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14.

38. Kuzniewicz MW, Walsh EM, Li S, Fischer A, Escobar GJ. Development and Implementation of an Early-Onset Sepsis Calculator to Guide Antibiotic Management in Late Preterm and Term Neonates. The Joint Commission Journal on Quality and Patient Safety. 2016 May;42(5):232–9.

39. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards: Critical Care Medicine. 2016 Feb;44(2):368–74.

40. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clinical Infectious Diseases. 2018 Jan 6;66(1):149–53.

41. Gannon H, Chimhuya S, Chimhini G, Neal SR, Shaw LP, Crehan C, et al. Electronic application to improve management of infections in low-income neonatal units: pilot implementation of the NeoTree beta app in a public sector hospital in Zimbabwe. BMJ Open Qual. 2021 Jan;10(1):e001043.

42. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350(jan07 4):g7594–g7594.

43. Kyohere M, Davies HG, Musoke P, Nakimuli A, Tusubira V, Tasimwa HB, et al. Seroepidemiology of maternally-derived antibody against Group B Streptococcus (GBS) in Mulago/Kawempe Hospitals Uganda - PROGRESS GBS. Gates Open Res. 2020 Nov 13;4:155.

44. Villar J, Giuliani F, Bhutta ZA, Bertino E, Ohuma EO, Ismail LC, et al. Postnatal growth standards for preterm infants: the Preterm Postnatal Follow-up Study of the INTERGROWTH-21 st Project. The Lancet Global Health. 2015 Nov;3(11):e681–91.

45. Buderer NMF. Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. Academic Emergency Medicine. 1996 Sep;3(9):895–900.

46. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003 Sep;20(5):453–8.

47. Nathoo KJ, Mason PR, Chimbira TH. Neonatal septicaemia in Harare Hospital: aetiology and risk factors. The Puerperal Sepsis Study Group. Cent Afr J Med. 1990 Jun;36(6):150–6.

48. Mathela SK, Mapfumo C, Mzingwane ML. Risk factors and practices contributing to sepsis in neonates admitted at a Children's Hospital in Zimbabwe [Internet]. In Review; 2022 May [cited 2023 Dec 6]. Available from: https://www.researchsquare.com/article/rs-1641840/v1

49. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine. 2016 Jan 30;35(2):214–26.

11. APPENDICIES

11.1 Appendix 1

Schedule of Procedures				
Procedures	**Visits (insert visit numbers as appropriate)**			
	Screening	**Antenatal**	**Delivery**	**1 Month**
Informed consent	x			
Completion of relevant CRF		x	x	x
Blood culture (neonatal)			x	
Blood spot for PCR (neonatal)			x	
Follow up (telephone)				x

11.2 Appendix 2

Amendment Log				
Amendment No.	**Protocol version no.**	**Date issued**	**Author(s) of changes**	**Details of changes made**
1	2.1	09.02.2024	[redacted]	Clarification on focus group rationale; see section 5

11.3 Appendix 3

Complete the form below. It will require review and sign-off by the Institute Director (SGUL) or the Care Group Lead (SGHFT).

Research Data Protection Impact Assessment (DPIA)

Using machine learning to model early-onset neonatal sepsis risk in Uganda (NeoRisk)

Data Protection Impact Assessments (DPIAs) are a tool which can help organisations identify the most effective way to comply with their data protection obligations under the Data Protection Act 2018 (DPA 18) and meet individuals' expectations of privacy.

A DPIA helps identify data privacy risks when planning new, or revising existing, projects and to identify actions to mitigate these risks. In the rare cases where risks cannot be mitigated at all it may be necessary to consult with the Information Commissioner's Office (ICO). Under data protection legislation it is a legal requirement to complete a DPIA in the following circumstances:

- • where data processing is likely to result in a high risk of harm to individuals, e.g. new, invasive technology is proposed
- • when large volumes of personal data are processed, e.g. use of behavioural profiles based on website usage
- • when processing special category personal data on a large scale, e.g. healthcare data, genetic tests to assess and predict the disease/health risks

• where publicly accessible areas are monitored, e.g. CCTV or when filming public areas

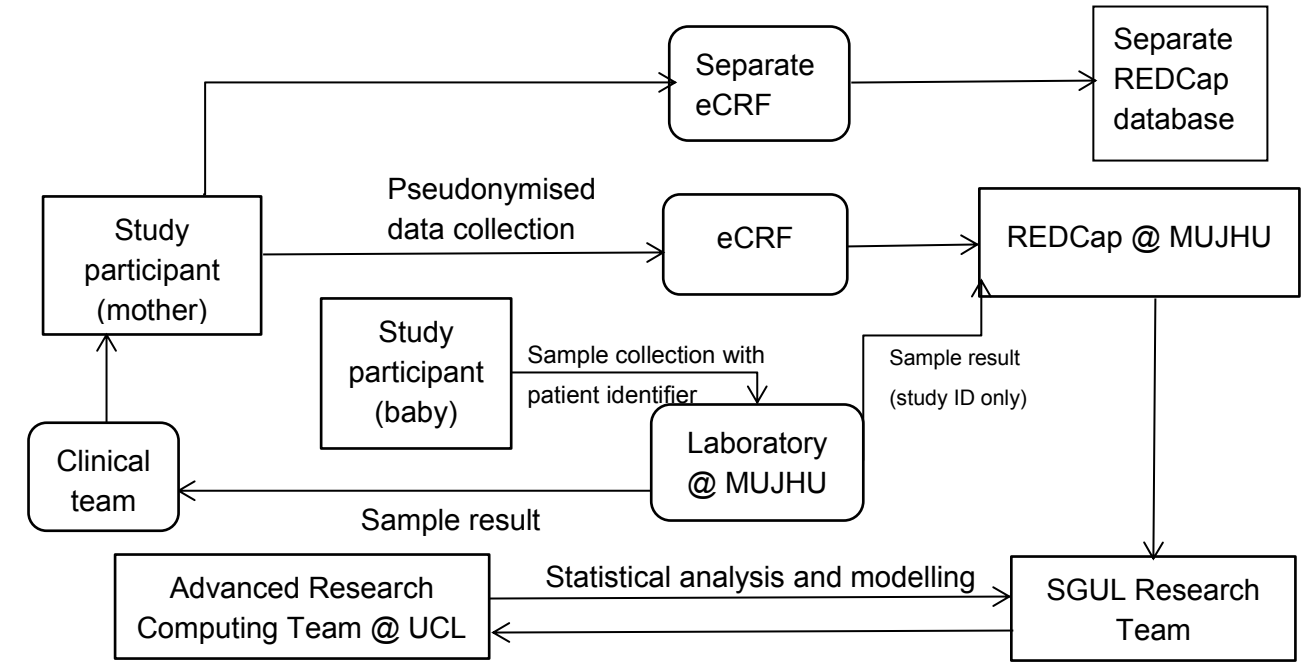
Therefore a DPIA will be carried out for both internal and partnership projects which require the collection/processing of personal data in any format for the purpose of research.

The DPIA should be carried out towards the start of the project, in order to identify any associated information risks and mitigate in the early stages, before you start processing.

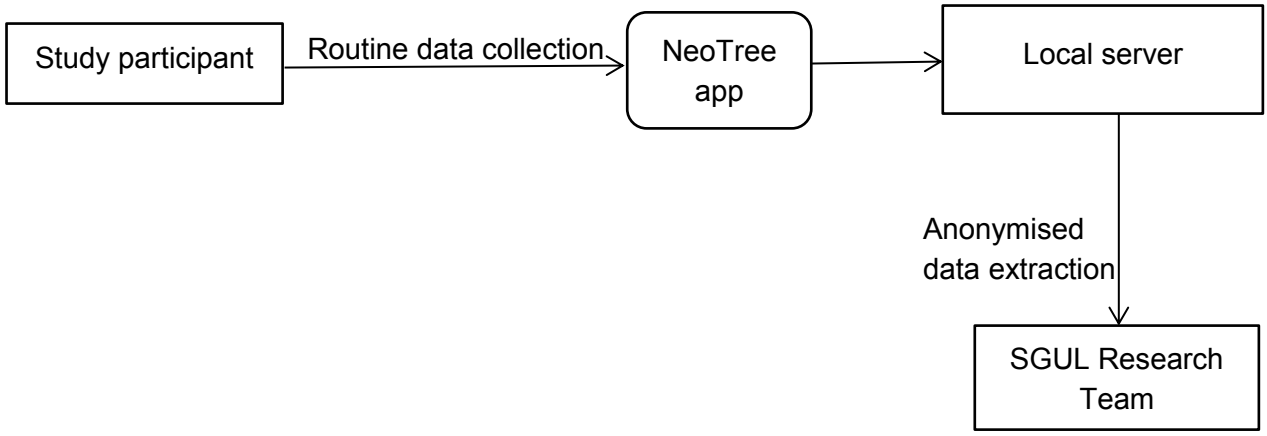
Study Title/Acronym:	**Using machine learning to model early-onset neonatal sepsis risk in Uganda**
JRES Reference Number:	TBD
Chief Investigator Name:	[redacted]
Chief Investigator Email Address:	[redacted]

PROJECT DETAILS
Project / process description: - include / attach processing operations (include a flow diagram or another way of explaining data flows), the purpose and, where applicable, what St George's lawful basis is for the processing of the information.

Phase 1 & 2



Phase 3



All research data will be anonymised at source, upon entrance to REDCap where a study ID will be used and no patient-identifiable information. This anonymised data will be transferred from MUJHU and BRTI respectively to SGUL for statistical analysis. Anonymised data only will be transferred between SGUL and UCL (Advanced Research Computing team) for the purposes of statistical modelling.

Data will be held locally on secure servers at MUJHU and BRTI, and centrally at SGUL.

What personal data do you intend to use, and why? (List all categories)

The data listed in the table in Section 3 will be collected via an online CRF.

These data will be used because all are potentially related to an infant's risk of early-onset neonatal sepsis, and therefore are relevant and required to answer the study's research question.

Will the personal data be identifiable, pseudonymised or anonymised (if a mix tick accordingly)

Identifiable		
*Pseudonymised	Yes	
Anonymised		

**Confirm that the key to this data is kept securely away from the used data with strict controlled access*

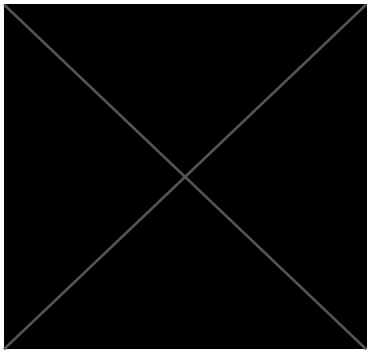
List all organisations / agencies which will have access to the personal data collection used for this project / process

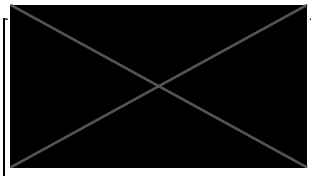
St George's, University of London
Makerere University Johns Hopkins University collaboration
Biomedical Research and Training Institute, Zimbabwe
University College London
Primary institutional review boards and monitors

Length of the study – include an assessment of the necessity and proportionality of the processing in relation to the purpose. Also include who, internally & externally, has been consulted in the preparation of this DPIA.

Recruitment will take place over an 18 month period from September 2023-March 2025. The overall project is scheduled to be completed by March 2026.

The followin eo le have been consulted in preparation of this DPIA:





If external organisations / agencies are involved, is there a contract or information sharing agreement in place with suitable clauses for data protection and data incident reporting,? If not why not?

A data sharing agreement will be in place between SGUL and UCL, SGUL and MUJHU, and SGUL and BRTI prior to the transfer or sharing of any data in this study, with suitable clauses for data protection and data incident reporting. Only pseudonymised data will be shared between SGUL and UCL, for purposes of statistical analysis and modelling only.

RISK

Can you achieve your objectives using anonymised data? – see ICO Code of Practice on Anonymisation

Yes	X		
No		Why not?	

What are the benefits to the individual of their personal data being used for this purpose?

There are no benefits to specific individual, but this study will hopefully help to develop a tool that can be applied in future to aid in timely diagnosis and management of neonatal sepsis in their community.

What are the organisational benefits of the individual's personal data being used for this purpose?

Use of personal data for the purposes of this study will be used to generate a risk stratification model for neonatal sepsis. Any resulting manuscripts or conference presentations will recognise the contribution of the organisation in creating this research, and will assist in building the organisation's profile within neonatal and paediatric infectious diseases research. Additionally, although the code used to create the model will be available via Github for use by others, this model may be used to develop an app or web interface to aid in neonatal sepsis risk stratification, which would also raise the organisation's profile within health technology.

What are potential negative impacts to the individual of their personal data being used for this purpose in the event of a Data Breach occurring?

All sensitive information will be fully anonymised at the point of data entry into REDCap, so in the event of a Data Breach an individual will not be identifiable. Similarly, a data breach during data sharing between MUJHU or BRTI and SGUL should not result in the release of any patient identifiable information or data.

How will you avoid causing unwarranted or substantial damage/distress to the individual when using their personal data for this purpose?

All sensitive information will be anonymised at the point of entry into REDCap using a study ID.
All data will be collected and stored in accordance with SGUL and MUJHU's information governance and data management policies, which are compliant with the NHS Data Security and Protection Toolkit and are Cyber Essentials Plus certified. SGUL and MUJHU are REDCap consortium partners, and each have a dedicated server which uses Secure Sockets Layer (SSL) to protect data in transit. Data

sharing agreements will be in place between SGUL and UCL, and SGUL and BRTI to enable secure data sharing for statistical analysis.

Is the data already held by St George's?			
Yes			
No	X	Prospective study	
Is it held by one of the partner organisations / agencies involved in this process/project?			
Yes			
No	X	Which agency will be collecting the data	MUJHU, BRTI
Have you told the individuals whose personal data you want to use for this purpose, how and why you intend to use their data?			
Yes			
No	X	Prospective collection	
If not, are you intending to tell them?			
Yes	X	Will be included in Informed Consent form and Patient Information Sheet	
No		Why not?	
Do you already have the individual's consent to use their data for this purpose?			
Yes			
No	X	Why not?	Prospective study – participants will provide consent at recruitment
If not, are you going to ask for their permission?			
Yes	X		
No		Why not?	
Have individuals been given the opportunity to refuse us permission to use their data for this purpose?			
Yes	X	Declining to consent to participation in the study will have no impact on their clinical care	
No			

How will you make sure that the personal data you are using is kept accurate and up to date?

Information will be verified at each study visit, with particular attention paid to contact information (e.g. telephone number) to enable follow up. This patient identifiable information will be stored with patient ID and date of birth only, on a separate REDCap database accessible via an encrypted password protected laptop stored in a locked cabinet at the hospital site in Kawempe. These data will be destroyed once follow up is complete.

What steps or controls are you taking to minimise risks to privacy?
Please tick those which apply and provide details of how each is ensured

• Risks to individual privacy are minimal • Personal data is pseudonymised • Encryption of data at rest, i.e. when stored • Encryption used in transfers	• Limited identifiable data collected • Pseudonymisation at point of online CRF

• Information compliance training for staff has been completed - data protection, information security, FOI • Special category personal data is not used • Participant opt out at any stage of the research • Research is not used to make decisions directly affecting individuals • Short retention limits • Restricted access controls	• Encryption used in storage and transfer by using REDCap and data sharing agreements • Information compliance training will be completed by all study staff • No special category personal data will be stored in REDCap • Participants can withdraw at any point • Personal data will be kept in country (Uganda or Zimbabwe) • Data will only be stored for as long as is necessary • Access restricted to trained study staff

How long will you need to hold the personal data for after the study has completed?

Personal data to enable contact, and identifiable personal data, such as telephone numbers and address to enable follow up will be destroyed by the local site once follow up is complete.
Study data will be archived by SGUL after study completion in accordance with standard SGUL procedures.

How will you make sure that you are holding data for the appropriate length of time and no longer?

The study protocol and data management plan will be peer reviewed, and reviewed by the relevant research ethics authorities.

How will the data be held /stored?

No patient identifiable data (name, date of birth, address, national ID or hospital number) will be stored on the main REDCap database. Telephone numbers and addresses, to enable follow up, will be stored with patient ID and date of birth only, on a separate REDCap database accessible via an encrypted password protected laptop stored in a locked cabinet at the hospital site in Kawempe. These data will be destroyed once follow up is complete. Microbiological samples will be stored and managed according to SGUL and MUJHU's regulations. Consent will be sought from participants for storage of samples for future ethically approved research.

Data will be archived after the completion of the study in SGUL-managed storage. File format conversion will be done to make the data suitable for long-term retention.

Will you be using any electronic and/or paper Case Report Forms (CRFs) to collect data? If so what are these and how will they be held securely and managed at the end of the project?

Prospective patient-level data will be collected from electronic health records (and participant interviews where needed) into an online data entry system (REDCap). This is a secure web application for building and managing online surveys and databases which supports online and offline data capture. This data entry system has been chosen as it will improve the consistency of the data, and allows them to be immediately available in an analysable format. The data are also stored securely without being stored on an individual hard drive, ensuring their safety and long-term access without need for physical backup.

Informed Consent Forms will be stored securely by the local research team according to MUJHU policies, and electronic medical records data will be held on a secure local areas network at Kawempe National Referral Hospital and extracted periodically to a secure server at SGUL.

Will personal data be transferred/shared between the organisations involved in this project? If so how?		
No identifiable data will be shared between organisations. Data sharing agreements will be in place for sharing of anonymised data		
Will you be transferring personal data to a country or territory outside of the UK? *If yes, name countries and receiving parties.*		
Yes – within EEA		
Yes – outside of EEA		
No	X	Anonymised data will be transferred from countries outside the UK to the UK for analysis

How will you ensure that third parties will comply with data protection obligations?

Study staff members will be required to complete study-specific training before being given access to REDCap. Study staff members will also need to show evidence of locally relevant research training, such as Good Clinical Practice certification, GDPR training and international equivalent.

What measures are in place to ensure only appropriate and authorised access to and use of, personal data?

All study staff will have an individual log in to REDCap to input data with specific permissions granted for data entry and/or editing as required. All will have an end date added to their access to REDCap, and this end date will be amended as necessary in the case of staff leaving the study team.

The access list on REDCap will be reviewed by the study management team on a monthly basis to ensure it is current and accurate.

How will technical and organisational security be monitored/audited?

The PI will have primary responsibility for data management and storage. The project's dedicated data manager, project manager, and study coordinator will be responsible for daily administration of the data, data quality and data curation. Data security responsibility will be shared between project members and SGUL IT staff.

Declaration

I confirm that the information recorded on this form is, to the best of my knowledge, an accurate and complete assessment of the potential privacy impacts of this study.

Name: [redacted]

Signature: [redacted]

Date [redacted]

Institute Director (SGUL) or Care Group Lead (SGHFT)

Name:

Signature:

Date:

JRES Reviewer

Name:

Signature:

Date: