DEEPNEO

DEEPNEO: A machine-learning approach to describe variations in clinical practice and patient outcomes in neonatal units in England and Wales

Version 190919

MAIN SPONSOR: Imperial College London
FUNDERS: NIHR Imperial Biomedical Research Centre ITMAT Funding Award
NRES reference:
IRAS Project ID: 273001

Protocol authorised by:

Name & Role	Date	Signature
Professor Modi CI	200919	

Study Management Group

Chief Investigator: Professor Neena Modi, Professor of Neonatal Medicine, n.modi@imperial.ac.uk

Co-investigators:

- Dr Christopher Gale MRC Clinician Scientist, Clinical Reader and Consultant Neonatologist, christopher.gale@imperial.ac.uk
- Kayleigh Ougham Data Analyst, NDAU k.ougham@imperial.ac.uk
- Dr Matthew Hyde Research Associate matthew.hyde02@imperial.ac.uk
- Dr Cheryl Battersby Clinical Senior Lecturer c.battersby@imperial.ac.uk
- Dr Elsa Angelini, Senior Data Scientist, e.angelini@imperial.ac.uk
- Professor Robert Glen, Professor of Computational Medicine r.glen@imperial.ac.uk
- Dr Sam Greenbury Research Associate s.greenbury@imperial.ac.uk
- Jinyi Wu Research Assistant j.wu17@imperial.ac.uk

Statistician: Professor Robert Glen

Study Management: Professor Neena Modi

Study Coordination Centre *Not applicable*

Clinical Queries

Clinical queries should be directed to Professor Modi

Sponsor

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

Joint Research Compliance Office
Imperial College London and Imperial College Healthcare NHS Trust
Room 215, Level 2, Medical School Building
Norfolk Place
London, W2 1PG
Tel: 0207 594 9459

Funder: NIHR Imperial Biomedical Research Centre ITMAT Funding Award

This protocol describes the DEEPNEO study and provides information about procedures. Every care was taken in its drafting, but corrections or amendments may be necessary. These will be circulated to investigators in the study. Problems relating to this study should be referred, in the first instance, to the Chief Investigator.

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

Table of Contents

GLOSSARY OF ABBREVIATIONS

NDAU	Neonatal Data Analysis Unit
NNRD	National Neonatal Research Database

KEYWORDS
Database, Machine learning, Neonatal, Informatics, Patient Outcomes, Retrospective cohort study

STUDY SUMMARY

TITLE DEEPNEO: A machine-learning approach to describe variations in clinical practice and patient outcomes in neonatal units in England and Wales

DESIGN This is a retrospective cohort study, using existing de-identified data held in the National Neonatal Research Database (NNRD). We will not use any patient identifiable information.

We will initially map different components of clinical care. Using feeding as an example, we can categorise this as enteral (milk) and parenteral (intravenous). In the *enteral* component this can be modelled by taking into account the type of milk (mothers own; donor, formula), time of initiation (age), and rate of advancement (ml/kg/day). In the *parenteral* component, we can model this to account for type of formulation, time of initiation, rate of advancement and route of administration.

AIMS We aim to address the research question “Can we identify patterns of combinations of treatments using ML approaches that align with clinical explicability and predict clinical outcomes?”

OUTCOME MEASURES Primary: Any and exclusive breast-feeding at discharge

Secondary outcomes will include (not confined to):

Length of stay (number of days between first neonatal unit admission and final neonatal unit discharge)

Discharge weight (weight for post-menstrual age standard deviation z-score at final neonatal unit discharge)

Weight gain (difference in z-scores between birth and final discharge)

Brain injury (defined as any of grade 3-4 periventricular haemorrhage; periventricular leucomalacia; porencephalic cyst; cerebral abscess; hydrocephalus requiring treatment; stroke); Bronchopulmonary dysplasia defined according to the National Neonatal Audit Programme (NNAP) definition

Necrotising enterocolitis (requiring surgery and/or leading to death)

Survival (defined as alive at final neonatal unit discharge)

Time to full enteral feeds (age in days when parenteral nutrition was stopped)

Retinopathy of prematurity (ROP) (defined as any ROP and ROP receiving treatment)

Sepsis (defined according to NNAP definition)

POPULATION Infants within the NNRD admitted to neonatal units in England and Wales between January 1st 2007 and December 31st 2018

ELIGIBILITY Infants within the NNRD admitted to neonatal units in England and Wales between January 1st 2007 and December 31st 2018

DURATION One year

1. INTRODUCTION

1.1 BACKGROUND

Neonatal care is among the top three high-cost, specialised services, needed by 1 in 8 newborn babies in the UK. Neonatal conditions are responsible for 40% of child deaths and are a major contributor to life-long health need. Improving neonatal outcomes is a national policy priority as reflected e.g. in the “national ambition” to reduce mortality and halve neonatal brain injuries in England by 2025, announced by the Secretary of State for Health in 2017.

The advent of machine learning (ML) big-data approaches has opened the possibility of utilising large databases of “Real-World Data” (data derived from real-world settings) to bring about transformational change in clinical care by generating “Real-World Evidence” faster, and at lower cost than has previously been possible. There is great willingness within clinical communities to reduce unwarranted variation in care. The public sector requires rigorous evaluation of processes and outcomes. A systematic ML approach to identify robust relationships between processes and outcomes would have wide applicability and relevance.

We aim to test the hypothesis “Data-driven ML analytic approaches can identify interventions and processes that are linked with defined clinical outcomes”.

1.2 RATIONALE FOR CURRENT STUDY

We aim to address the research question “Can we identify patterns of combinations of treatments using ML approaches that align with clinical explicability and predict clinical outcomes?”

2. STUDY OBJECTIVES

Primary: Any and exclusive breast-feeding at discharge

Secondary outcomes will include but are not confined to:

- Length of stay (number of days between first neonatal unit admission and final neonatal unit discharge)
- Discharge weight (weight for post-menstrual age standard deviation z-score at final neonatal unit discharge)
- Weight gain (difference in z-scores between birth and final discharge)
- Brain injury (defined as any of grade 3-4 periventricular haemorrhage; periventricular leucomalacia; porencephalic cyst; cerebral abscess; hydrocephalus requiring treatment; stroke); Bronchopulmonary dysplasia defined according to the National Neonatal Audit Programme (NNAP) definition
- Necrotising enterocolitis (requiring surgery and/or leading to death)
- Survival (defined as alive at final neonatal unit discharge)
- Time to full enteral feeds (age in days when parenteral nutrition was stopped)
- Retinopathy of prematurity (ROP) (defined as any ROP and ROP receiving treatment)
- Sepsis (defined according to NNAP definition)

3. STUDY DESIGN

This is a retrospective cohort study, using existing de-identified data held in the National Neonatal Research Database (NNRD). We will not use any patient identifiable information.

We will initially map different components of clinical care. Using feeding as an example, we can categorise this as enteral (milk) and parenteral (intravenous). In the enteral component this can be modelled by taking into account the type of milk (mothers own; donor, formula), time of initiation (age), and rate of advancement (ml/kg/day). In the parenteral component, we can model this to account for type of formulation, time of initiation, rate of advancement and route of administration.

4. PARTICIPANT ENTRY

4.1 PRE-REGISTRATION EVALUATIONS

This is a retrospective database study

4.2 INCLUSION CRITERIA

Infants within the NNRD admitted to neonatal units in England and Wales between January 1st 2007 and December 31st 2018

4.3 EXCLUSION CRITERIA

N/A

4.4 WITHDRAWAL CRITERIA

N/A

5. ADVERSE EVENTS

Not applicable

6. ASSESSMENT AND FOLLOW-UP

There will be no follow up, this is a retrospective data base study using a dataset.

7. METHODS AND DATA ANALYSIS

The NNRD is the data source for this study. The NNRD holds data from all infants admitted to NHS neonatal units in England, Scotland and Wales (approximately 100,000 infants annually). The NNRD is formed from data extracted from the neonatal electronic health record system used by health professionals during routine clinical care.

Briefly, daily clinical information on neonatal unit admissions is recorded in a point-of-care, clinician-entered Electronic Patient Record. A defined data extract, the Neonatal Dataset (NHS Information Standard SCCI595) is transmitted quarterly to the Neonatal Data Analysis Unit at Imperial College London and Chelsea and Westminster NHS Foundation Trust where patient episodes across different hospitals are linked, data are cleaned, and entered into the NNRD. Contributing neonatal units are known as the UK Neonatal Collaborative (UKNC). A GDPR compliant Parent Leaflet is available for neonatal units to give to parents and a poster to display. This explains the purpose of the NNRD and how parents can opt-out of having their baby’s data included.

The NNRD holds the Neonatal Data Set, approximately 450 data items that form a NHS data standard (7). Data items include demographic and admission items (e.g. maternal conditions, birthweight), daily items (entered every day for all infants, e.g. respiratory support, feeding information), discharge items (e.g. feeding and weight at discharge) and ad hoc items (entered if and when they occur e.g. suspected infection, ultrasound scan findings, abdominal x-ray findings).

Data extracted from the neonatal Electronic Health Record are cleaned; records with implausible data configurations are queried and corrected by the treating clinicians. Cleaning is carried out by the Neonatal Data Analysis Unit before data are incorporated into the NNRD. The robustness of core NNRD data (birth weight, sex, length of stay and death) has been previously demonstrated for research purposes. Data held in the NNRD are used for multiple purposes including national audit (the National Neonatal Audit Programme) and analyses for the Department of Health.

We will use classic unsupervised clustering methods such as K-means, community detection methods (e.g. Infomap (1)) and generative models such as Gaussian Mixture Models (2). These can be combined with feature reduction and manifold learning (3) if needed.

We will employ supervised Deep Learning, using denoising autoencoders architectures such as “Deep Patient” (4) to generate “deep” mixed features, which are then exploited in supervised classic classifiers (5) such as SVM and Random Forests for prediction of a specific outcome.

We will apply imputation to deal with missing values (on non-principal background or outcome variables). We will test different approaches, including modal imputation, random imputation, and multiple imputation by chained equations (6).

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

8. REGULATORY ISSUES

8.1 ETHICS APPROVAL

The Chief Investigator has obtained approval from the HRA The HRA has reviewed and approved this study. The study must also receive confirmation of capacity and capability from each participating NHS Trust before accepting participants into the study. The only NHS Trust involved is Chelsea and Westminster NHS Foundation Trust. The study will be conducted in accordance with the recommendations for physicians involved in research on human subjects adopted by the 18th World Medical Assembly, Helsinki 1964 and later revisions.”

8.2 CONSENT

Consent is not applicable. The NNRD has REC approval 16/LO/1093. Parents are given the option to opt out of their or their babies data’s’ inclusion into the NNRD instead.

8.3 CONFIDENTIALITY

The Chief Investigator will preserve the confidentiality of participants taking part in the study and is registered under the Data Protection Act.

8.4 INDEMNITY

Imperial College London holds negligent harm and non-negligent harm insurance policies which apply to this study/ Imperial College Healthcare NHS Trust holds standard NHS Hospital Indemnity and insurance cover with NHS Resolution for NHS Trusts in England, which apply to this study (delete as applicable)

8.5 SPONSOR

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

8.6 FUNDING

Imperial Biomedical Research Centre ITMAT Funding Award.

8.7 AUDITS

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

9. STUDY MANAGEMENT

The day-to-day management of the study will be co-ordinated through Professor Modi.

10. PUBLICATION POLICY

We will report the results of this study in peer-reviewed scientific journals. All members of the study group will be named authors "and the UK Neonatal Collaborative". We will additionally acknowledge all UK Neonatal Collaborative Leads and neonatal units in accordance with standard processes followed at the Neonatal Data Analysis Unit.

10. REFERENCES

1) Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 2008;105:1118-23.
2) Rasmussen C, Williams C. Gaussian Processesfor Machine Learning. Press M, editor2005.
3) Law MHC, Jain AK. Incremental nonlinear dimensionality reduction by manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;28(3):377-91.
4) Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports. 2016;17(6):1-10
5) Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. ACM International Conference on Machine Learning2006. p. 161-8.
6) Buuren SV, KG-O. MICE: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.
7) http://www.datadictionary.nhs.uk/web_site_content/navigation/national_neonatal_data_sets2019