ASMHI 7PAVREPR PROJECT DETAILS

2022/23

Last Updated 03/10/22

MSC APPLIED STATISTICAL MODELLING AND HEALTH INFORMATICS (ASMHI)

BIOSTATISTICS AND HEALTH INFORMATICS DEPARTMENT IOPPN, KING'S COLLEGE LONDON

TABLE OF CONTENTS

Project ID 1 3

Project ID 2 6

Project ID 3 9

Project ID 4 12

Project ID 5 15

Project ID 6 18

Project ID 7 21

Project ID 8 24

Project ID 9 27

Project ID 10 30

Project ID 11 32

Project ID 12 35

Project ID 13 38

Project ID 15 41

Project ID 16 45

Project ID 17 48

Project ID 18 51

Project ID 19 54

Project ID 20 57

Project ID 21 59

Project ID 22 62

Project ID 23 64

Project ID 24 67

Project ID 25 70

Project ID 26 75

Project ID 27 78

Project ID 28 81

Project ID 29 84

Project ID 30 87

Project ID 31 90

Project ID 32 93

Project ID 33 96

PROJECT ID 1

PRIMARY SUPERVISOR

Mohammad Mahdi Karimi

EMAIL ADDRESS

mohammad.karimi@kcl.ac.uk

DEPARTMENT

Comprehensive Cancer Centre

SECONDARY SUPERVISOR

Sheeba Irshad

EMAIL ADDRESS

sheeba.irshad@kcl.ac.uk

PROJECT TITLE

A deep learning approach for prediction of spatial transcriptomics from histological images in breast cancer

RESEARCH AREA

Bioinformatics, Spatial transcriptomics, Genomics

PROJECT AIMS

- Establishing a TBNC longitudinal spatial transcriptomic atlas
- Recapitulating spatial transcriptomic information through machine learning

PROJECT OUTLINE

Triple-negative breast cancer (TNBC) is defined as a type of breast cancer lacking ER, PgR and HER2 expression. In comparison to other types of breast cancer, TNBC is characterised by its clinically aggressive nature, younger age at presentation, distinct metastatic patterns and lack of effective targeted therapies, thus representing an important clinical challenge. The high heterogeneity of TNBC within individual tumours and across patients has been widely suggested as the reason for diverse response rates to traditional treatments or new targeted therapies, leading to discrepant patient outcomes.

Multigene expression signatures, such as the Prosigna 50 gene-based test, provide a molecular subdivision of early breast cancer associated with patient outcome. While in ER+/HER2− disease, gene signatures appear to contribute additional prognostic value even at a relatively short follow-up time, significantly less apparent prognostic value has been observed in TNBC. This underlines the critical importance of combining genetic information with other markers to understand disease progression and response. The GeoMx Digital Spatial Profiler is a unique instrument in enabling multiplexed spatial profiling of both RNA and protein, even in challenging FFPE samples. Since its introduction, GeoMx enabled spatial transcriptomic (ST) atlases for several human diseases which are unlocking new therapeutic options and molecular understanding. Generating a

GeoMx atlas of TNBC therapeutic response will provide genetic and molecular background as well as the cellular interplay within the tumour microenvironment (TME) which determines/underlies disease progression and response to treatment. Our banked tissue provides an extensive, digitised resource to interrogate ST across TNBC samples of variable grade, stage and genetic background, alongside full clinical data. Using our recently funded Wellcome-LEAP grant as matched funding for this proposal, we plan to generate a fully annotated human TNBC ST atlas.

The ST TNBC atlas data generated from this combined dataset will be the ideal substrate on which we can apply deep learning methods as predictive tools to improve clinical decision-making and reduce diagnostic costs. The cost of ST and NGS is too expensive for routine clinical operation (£7000/slide) especially within the context of treatment in the NHS; and its throughput is insufficiently low (5/slide). To circumvent these limitations, we propose to apply deep-learning approaches such as convolutional neural networks (CNN) to predict ST information. CNN combined with other state-of-the-art deep learning methods will provide the tool to reconstruct the ST information from hematoxylin and eosin (H&E)-stained section images. This approach provides spatial transcriptomic-equivalent (STe) information that will contribute to the generation of the patient avatar for stratification, treatment selection and evaluation. Moreover, the deep learning model will accelerate 60-fold the time of sample collection to data generation, from 5 days every 6 slides for ST to 1 day per 50 slides for STe, and reduce the cost of the data generation for GeoMx+NGS to £20 for H&E histological slides including imaging and computational time. We will provide open access to this deep learning model to generate STe TNBC for new human samples.

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

The student will learn about the applications of deep learning in the genomics field. All deep learning modelling will be performed on our recently purchased GPU server integrated into CREATE, the KCL HPC cluster. The student will be involved in running scripts on the KCL HPC and writing/troubleshooting of Python scripts using the scikit-learn library.

STARTER REFERENCES

1. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, Zou J. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng. 2020 Aug;4(8):827-834. doi: 10.1038/s41551-020-0578-x. Epub 2020 Jun 22. PMID: 32572199.

2. Li Y, Stanojevic S, Garmire LX. Emerging artificial intelligence applications in Spatial Transcriptomics analysis. Comput Struct Biotechnol J. 2022 Jun 2;20:2895-2908. doi: 10.1016/j.csbj.2022.05.056. PMID: 35765645; PMCID: PMC9201012.

EXPECTED OUTPUTS AND TASKS

We will seek to have a data scientist MSc student to apply deep learning methods on ST data. This student will be the primary person responsible for deploying single cell analysis pipelines on Linux server, and test the performance of deep learning tools.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

Yes

STUDENT REQUIREMENTS

NA

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7BBG2014 Bioinformatics, interpretation and data quality assurance in genome analysis;7BBG2016 Advanced Bioinformatics: practical bioinformatics data skills;

PROJECT ID 2

PRIMARY SUPERVISOR

Ben Carter

EMAIL ADDRESS

ben.carter@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Roxanna Short

EMAIL ADDRESS

roxanna.short@kcl.ac.uk

PROJECT TITLE

The non-linear relationship between age and mortality in research of older people

RESEARCH AREA

Statistical modelling of older people

PROJECT AIMS

To determine the optimal parameterisation of the prognostic covariate age in modelling mortality in research for older people.

PROJECT OUTLINE

Researchers often report a non-linear relationship between age and mortality in geriatrics research. Little understanding of the consistency of this relationship, or over the clinical meaning of the interpretation.

Using four established older people dataset, students will use four established cohorts and fit patient age with an optimal generalised additive model and compare model fit statistics to the published models.

This project will be hosted within the CLARITY Research group

https://www.nbt.nhs.uk/research-innovation/our-research/current-research/ageing-research-hub/clarity-team

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Statistical modelling, Statistical Computing (Stata), communication

STARTER REFERENCES

Perera, M. and Tsokos, C. (2018) A Statistical Model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis. Advances in Breast Cancer Research, 7, 65-89. https://doi.org/10.4236/abcr.2018.71005

Frøslie, K.F., Røislien, J., Laake, P. et al. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study. BMC Med Res Methodol 10, 103 (2010). https://doi.org/10.1186/1471-2288-10-103

Carter B, Short R, Bouamra O, Parry F, Shipman D, Thompson J, Baxter M, Lecky F and Braude P. A national study of twenty-three major trauma centres to investigate the effect of frailty on clinical outcomes of older people admitted with major trauma (FiTR): a multicentre observational study. The Lancet Healthy Longevity, 2022

Hewitt J, Carter B, Vilches-Moraga A, Quinn TJ, Braude P, Verduri A, Pearce L, Stechman M, Short R, Price A, Collins JT, Bruce E, Einarsson A, Rickard F, Mitchell E, Holloway M, Hesford J, Barlow-Pay F, Clini E, The effect of frailty on survival in patients with COVID-19 (COPE): a multicentre, European, observational cohort study. Lancet Public Health. 2020 Jun 30:S2468-2667(20)30146-8

Palmer, K. L., Law, J., Carter, B., Hewitt, J., Boyle, J., Maitra, C., Farrell, I. & Moug, on behalf of the ELF Study Group. S. Frailty in Older Patients undergoing Emergency Laparotomy: results from the observational ELF Study (Emergency Laparotomy and Frailty). Annals of Surgery; 2019; Jun 7.

J Hewitt, B Carter, K McCarthy, L Pearce, J Law, F V Wilson, H S Tay, C McCormack, M J Stechman, S J Moug, P K Myint, Frailty predicts mortality in all emergency surgical admissions regardless of age. An observational study, Age and Ageing, 2019, May 1;48(3):388-394. doi: 10.1093/ageing/afy217

EXPECTED OUTPUTS AND TASKS

The expectation is that this will be part of a project to establish how age is modeled within journals for older people

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes, four

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Access to Stata

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPRMD Prediction Modelling;7PAVCTMH Clinical Trials: A practical Approach;

PROJECT ID 3

PRIMARY SUPERVISOR

Daniel Stahl

EMAIL ADDRESS

daniel.r.stahl@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Paolo Fusar-Poli

EMAIL ADDRESS

paolo.fusar-poli@kcl.ac.uk

PROJECT TITLE

To develop a prediction model using machine learning methods to identify patients at risk of developing psychoses within two years, to compare its performance with a current theory-driven model and to assess the influence of sample size on prediction accuracy

RESEARCH AREA

Clinical prediction modelling, psychosis

PROJECT AIMS

AIM(S):

To develop a machine learning model to improve the detection of individuals at risk of developing psychosis among patients accessing secondary mental health care when using all available data and adequately handling missing data

To assess the influence of the number of variables relative to sample size on prediction accuracy using simulations by random sampling from the patient data sets.

PROJECT OUTLINE

Psychosis is a mental health problem that causes people to perceive or interpret things differently from those around them. This might involve hallucinations or delusions. Psychosis can be a symptom of serious mental illnesses such as bipolar disorder and schizophrenia. Approximately 1% of the UK population suffers from psychotic disorders with onsets often in children and young people. The onset of psychotic illness represents a potential personal disaster in the life course of a young individual and current treatments offer minimal help.

This project will focus on the development and validation of individualized predictive models and risk stratification in patients at high risk for psychosis using modern machine learning methods, such as xgboost or neural networks. Model building will follow the guidelines of Steyerberg and Vergouwe (2014) with special attention to accurate model performance assessment using internal and external validation methods and adequate handling of missing data to get accurate estimates of prediction accuracy.

The dataset will be a clinical register-based cohort with more than 100,000 patients drawn from electronic clinical records relating to secondary health care in South London and the Maudsley National Health Service Foundation trust.

The project builds on the study by Paolo-Fusar Poli et al (2017), which used only relatively simple statistical learning methods (regularized regression) and is based on a complete-case analysis. The aims of the project are to assess

i) if a machine learning model can improve the detection of individuals at risk of developing psychosis among patients accessing secondary mental health care when using all available data and adequately handling missing data

ii) the influence of the number of variables relative to sample size on prediction accuracy using simulations by random sampling from the patient data sets.

Further emphasis will be on explainability of the “black box” model to understand why the machine learning model makes the decisions they make.

The outputs of this project will be a revised prognostic tool and guidance on sample size considerations.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Developing robust prediction models using real-life data

2. Develop problem-solving and decision-making skills in clinical decision modelling

3. Learn clinical concepts about psychoses

STARTER REFERENCES

Fusar-Poli P, Rutigliano G, Stahl D, Davies C, Bonoldi I, Reilly T, McGuire P. (2017) Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. JAMA Psychiatry. 2017 May 1;74(5):493-500. doi: 10.1001/jamapsychiatry.2017.0284.

Fusar-Poli P, Stringer, D. , … Stahl D,   (2019) Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. Transl Psychiatry.; 9: 259. doi: 10.1038/s41398-019-0600-9

Introduction to Prediction modelling

Steyerberg EW, Vergouwe Y. (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. ;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

Introduction to missing data

SWJ Nijmanet al (2021) Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review,Journal of Clinical Epidemiology,142,218-229,

https://doi.org/10.1016/j.jclinepi.2021.11.023.

Introduction to Psychoses:

Basseer, M et al (2017) Early psychosis for the non-specialist doctor. BMJ 2017;357: j4578 doi: 10.1136/bmj.j4578

EXPECTED OUTPUTS AND TASKS

The outputs of this project will be a revised clinical prognostic tool using modern machine learning methods with adequately handling of missing data and guidance on sample size considerations for clinical prediction modelling.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

A research passport to access the CRIS data sets is needed.

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 4

PRIMARY SUPERVISOR

Dr Delia Fuhrmann

EMAIL ADDRESS

delia.fuhrmann@kcl.ac.uk

DEPARTMENT

Psychology

SECONDARY SUPERVISOR

Dr Kathryn Bates

EMAIL ADDRESS

kathryn.2.bates@kcl.ac.uk

PROJECT TITLE

Modeling sensitive periods for social isolation and the development of loneliness in youth

RESEARCH AREA

Developmental Psychology, Psychometrics

PROJECT AIMS

- Understanding at what age young people are sensitive to social isolation and loneliness

- Leveraging Latent Change Score Models and SEM Trees to study sensitive periods of development

PROJECT OUTLINE

There are dynamic biological, psychological, and social transitions during adolescence, which could make young people more vulnerable to stress, especially social stressors such as isolation and deprivation. However, it remains unclear at what age young people are most likely to experience loneliness in response to social isolation threats and when loneliness is most likely to emerge. Identifying age groups of heightened susceptibility to loneliness will inform when in youth, interventions are most needed and likely to be successful. Because prior research has been limited by difficulties manipulating social isolation experimentally in humans, we will use experiences of COVID-19 lockdowns to address this research question. As part of an ESRC-funded project, we will study how COVID-linked social isolation threat is linked to loneliness across different age groups within youth. We will leverage existing, large-scale data from population representative samples like the UK Understanding Society COVID-19 sample (N = 5,700, aged 16-24 years) and the emerging Co-SPACE study (N ~ 1,000, aged 11-16 years) to quantify changes in loneliness during the COVID-19 pandemic. Across this period, young people experienced nationwide lockdowns designed to halt the spread of COVID-19. We will compare how the threat of social isolation (during lockdown phases) was differentially related to changes in loneliness across age groups. We will use Latent Change Score Models to quantify changes in subjective experiences of loneliness during the pandemic. We will extract change scores for loneliness from these models and feed them into a Structural Equation Modelling Tree – an exploratory technique that can assess which age groups are most susceptible to non-adaptive loneliness without imposing arbitrary groupings a priori. This study can address long-standing questions about when loneliness is most likely to emerge and become non-adaptive, helping to identify target age groups for intervention.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

latent variable modeling, longitudinal modelling, data wrangling and analysis in R, Open Science approaches

STARTER REFERENCES

Fuhrmann, D., Knoll, L. J., & Blakemore, S. J. (2015). Adolescence as a Sensitive Period of Brain Development. Trends in cognitive sciences, 19(10), 558–566. https://doi.org/10.1016/j.tics.2015.07.008

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A. L., de Mooij, S., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., NSPN Consortium, Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. Developmental cognitive neuroscience, 33, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

Brandmaier AM, von Oertzen T, McArdle JJ, Lindenberger U. Structural equation model trees. Psychol Methods. 2013 Mar;18(1):71-86. doi: 10.1037/a0030001. Epub 2012 Sep 17. PMID: 22984789; PMCID: PMC4386908.

EXPECTED OUTPUTS AND TASKS

- Development of data wrangling scripts

- Development of analysis code

- Writing up of analysis

- Attendance of and contribution to lab meetings, methods, and journal clubs in the lab

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

NA

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPSYC Contemporary Psychometrics;7PAVCIAE Causal Modelling and Evaluation;

PROJECT ID 5

PRIMARY SUPERVISOR

Kim Goldsmith

EMAIL ADDRESS

kimberley.goldsmith@kcl.ac.uk

DEPARTMENT

Biostatistics & Health Informatics

SECONDARY SUPERVISOR

Trudie Chalder

EMAIL ADDRESS

trudie.chalder@kcl.ac.uk

PROJECT TITLE

Heterogeneity in Irritable Bowel Syndrome: latent class analysis of the ACTIB trial of cognitive behavioural therapy for irritable bowel syndrome

RESEARCH AREA

Irritable Bowel Syndrome and Latent Class Analysis

PROJECT AIMS

To investigate to what extent there are identifiable sub-groups (heterogeneity) at baseline of people with irritable bowel syndrome (IBS) in the ACTIB trial of cognitive behavioural therapy for IBS.

PROJECT OUTLINE

The ACTIB trial of CBT for IBS tested the effectiveness of Telephone-delivered and Web-delivered CBT for refractory IBS (IBS that is resistant to treatment). Both were effective in reducing IBS symptoms and improving work and social adjustment relative to treatment as usual (Everitt et al, 2015 and 2019). Further understanding of potential heterogeneity in IBS patients and how they might form groups with a certain profile of characteristics is the first step in understanding for whom these treatments might provide more benefit.

Latent class analysis (LCA) (see Weller et al for an overview and other useful references) will allow us to study heterogeneity in individuals with IBS, and possibly predictors of such heterogeneity. Latent class analysis takes an individual- rather than variable-focused approach, in that these analyses focus more on classification of individuals rather than on associations between variables. This type of analysis could provide a different way to study predictors of outcome, as opposed to looking at individual variables in isolation.

Heterogeneity will be assessed in the ACTIB population by fitting LCA models to baseline data. Models with different numbers of classes will be fitted. The most suitable model will be chosen by balancing minimisation of information criteria, such as the Bayesian Information Criterion, and theoretical considerations such as plausibility of the classes and parsimony. The latent class analysis will include baseline variables such as: age, gender, IBS type, duration, symptom severity, distress, behavioural responses, cognitive responses questionnaire, etc.

Dependent on the student's interest and skill level, the project could go on to look at whether the discovered latent classes differentially predict outcome.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Gain technical expertise in fitting latent class structural equation models.

2. Learn how to calculate and estimate effects from these models and how to display their results effectively.

3. Experience working with clinical colleagues and explaining how your findings could impact on their clinical practice.

STARTER REFERENCES

Everitt H, Landau S, Little P et al. Assessing Cognitive behavioural Therapy in Irritable Bowel (ACTIB): protocol for a randomised controlled trial of clinical-effectiveness and cost-effectiveness of therapist delivered cognitive behavioural therapy and web-based self-management in irritable bowel syndrome in adults

BMJ Open 2015;5:e008622. doi: 10.1136/bmjopen-2015-008622

Everitt HA, Landau S, O'Reilly G, Sibelli A, Stephanie H et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. Gut. 2019 Sep 1;68(9):1613-1623. https://doi.org/10.1136/gutjnl-2018-317805

Weller BE, Bowen NK, and Faubert SJ. Latent Class Analysis: A Guide to Best Practice. Journal of Black Psychology 2020, Vol. 46(4) 287–311. Doi: 10.1177/0095798420930932

EXPECTED OUTPUTS AND TASKS

Besides the dissertation, in some cases the work could form a publication, or the basis or part of a future publication.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

None

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVPSYC Contemporary Psychometrics;7PAVMALM Multilevel and Longitudinal Moddelling;

PROJECT ID 6

PRIMARY SUPERVISOR

Kimberley Goldsmith

EMAIL ADDRESS

kimberley.goldsmith@kcl.ac.uk

DEPARTMENT

Biostatistics & Health Informatics

SECONDARY SUPERVISOR

Trudie Chalder

EMAIL ADDRESS

trudie.chalder@kcl.ac.uk

PROJECT TITLE

For whom does cognitive behavioural therapy (CBT) work and how: moderation in mediated models of CBT for chronic fatigue syndrome

RESEARCH AREA

Chronic Fatigue Syndrome, mediation and moderation regression modelling, causal modelling

PROJECT AIMS

To more realistically model cognitive behavioural treatment processes in individuals with chronic fatigue syndrome by incorporating moderating variables into mediation models.

PROJECT OUTLINE

Understanding how treatments work (mechanisms) and for whom (in which subgroups) is key. Mediation and moderation analysis can answer such questions. Mediation is by its nature a longitudinal process, and can be modelled using different frameworks (MacKinnon et al, 2007). There is strong clinical interest in modelling overarching mediation and moderation processes, however these aspects are often studied separately. More comprehensive models have been of particular interest in understanding the treatment mechanisms of cognitive behavioural therapy (CBT) for people with chronic fatigue syndrome (CFS).

This treatment and its mechanisms were studied in the large robust PACE trial of treatments for CFS (Chalder et al 2015, White et al 2011). This was a trial of three rehabilitative therapies for CFS: CBT, graded exercise therapy and adaptive pacing therapy, in comparison to a control (specialist physician medical care). The trial recruited 641 participants, and measured a number of cognitive behavioural mediators (for example, fear avoidance and avoidance behaviour) and outcomes (physical functioning and fatigue) at baseline and at multiple time points post-randomisation (mid-treatment, post-treatment and one year follow-up).

The student will extend mediation models that have already been fitted to the PACE data to include interactions with potential moderators. The student will look at relationships between mediator, moderator and outcome variables. They may then go on to replicate previous mediation models (Chalder et al 2015) using the causal mediation programme “paramed”. The student will apply and compare these approaches to estimate moderated effects.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Gain technical expertise in fitting regression models that realistically represent processes.

2. Learn how to calculate and estimate mediated and moderated effects from statistical models and how to display these effectively.

3. Experience working with clinical colleagues and explaining how your findings could impact on their clinical practice.

STARTER REFERENCES

MacKinnon, Fairchild, Fritz. Mediation Analysis. Annu. Rev. Psychol, 2007; 58:593–614. doi: 10.1146/annurev.psych.58.110405.085542

Chalder, Goldsmith, et al. Rehabilitative therapies for chronic fatigue syndrome: a secondary mediation analysis of the PACE trial. Lancet Psychiatry, 2015; 2(2):141-152. doi: 10.1016/S2215-0366(14)00069-8.

White, Goldsmith et al. Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial. Lancet, 2011; 377(9738):823-836. doi: 10.1016/S0140-6736(11)60096-2.

Stata paramed: https://ideas.repec.org/c/boc/bocode/s457581.html

EXPECTED OUTPUTS AND TASKS

Besides the dissertation, in some cases the work could form a publication, or the basis or part of a future publication.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVCIAE Causal Modelling and Evaluation;7PAVCTMH Clinical Trials: A practical Approach;

PROJECT ID 7

PRIMARY SUPERVISOR

Kimberley Goldsmith

EMAIL ADDRESS

kimberley.goldsmith@kcl.ac.uk

DEPARTMENT

Biostatistics & Health Informatics

SECONDARY SUPERVISOR

Edmund Sonuga-Barke

EMAIL ADDRESS

edmund.sonuga-barke@kcl.ac.uk

PROJECT TITLE

Elucidating meaningful app usage pattern clustering: the SPARKLE trial of the Parent Positive app within the Co-SPACE cohort

RESEARCH AREA

Parenting support, app usage metrics, latent class analysis

PROJECT AIMS

To further analyse and characterise patterns of different types of usage of the Parent Positive app in the SPARKLE trial, investigate whether there are identifiable app usage sub-groups (heterogeneity), and look at predictors of these usage groups.

PROJECT OUTLINE

The SPARKLE trial of the Parent Positive parenting app (https://www.kcl.ac.uk/research/supporting-parents-and-kids-through-lockdown-experiences-sparkle-trial) was conducted as a trial within a cohort (TwiC) within the Co-SPACE cohort (http://cospaceoxford.org/). The Co-SPACE study was designed to assess how parents and families coped during COVID and what was and wasn't working for them in terms of supporting their children's mental health. Co-SPACE found an increase in child behaviour problems and family stress during lockdown, so the Parent Positive app was designed to support parents in addressing these and other problems via short videos on pertinent topics delivered by celebrity parents along with other resources to support these topics (boosters), high-quality evidence-based parenting web resources and a parenting exchange zone moderated by parent facilitators where parents could raise issues and discuss with other parents, as well as obtain advice from trained experts.

We gathered Parent Positive usage data, including: number of sessions, time spent accessing the boosters, the number of posts and comments on the parenting exchange, and the number of expert videos watched; these were gathered over the baseline to first outcome time point, and between the first and second outcome time points. We have looked at a small subset of these data for the main trial paper, but this project would delve deeper into the different types of data and possibly explore different patterns of usage in a longitudinal fashion.

The student could usefully start with some descriptive and correlational analysis of all the available data, and then apply latent class analysis (LCA, see Weller et al for an overview and other useful references) to study heterogeneity in app usage data. Models with different numbers of classes will be fitted. The most suitable model will be chosen by balancing minimisation of information criteria, such as the Bayesian Information Criterion, and theoretical considerations such as plausibility of the classes and parsimony. LCA takes an individual- rather than variable-focused approach, in that these analyses focus more on classification of individuals rather than on associations between variables. This type of analysis could provide a different way to study a group of variables, as opposed to looking at individual variables in isolation. The project could then go on to look at baseline predictors of being classified in the different app usage subgroups. We will likely start with app usage data from one of the follow-up periods, but dependent on the student's interest and skill level, they could look at modelling the data longitudinally, and possibly whether the latent classes differentially predict outcome.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Gain technical expertise in fitting longitudinal and latent class structural equation models.

2. Learn how to calculate and estimate effects from these models and how to display their results effectively.

3. Experience working with clinical colleagues and explaining how your findings could impact on their clinical practice.

STARTER REFERENCES

Kostyrka-Allchorne K, Creswell C, Byford S, Day C, Goldsmith K…Palmer M… & Sonuga-Barke E. Supporting Parents & Kids Through Lockdown Experiences (SPARKLE): A digital parenting support app implemented in an ongoing general population cohort study during the COVID-19 pandemic: A structured summary of a study protocol for a randomised controlled trial. Trials, 2021; 22: 267, doi:10.1186/s13063-021-05226-4.

Milne-Ives M, van Velthoven MH, Meinert E. Mobile apps for real-world evidence in health care. J Am Med Inform Assoc. 2020 Jun 1;27(6):976-980. doi: 10.1093/jamia/ocaa036.

Weller BE, Bowen NK, and Faubert SJ. Latent Class Analysis: A Guide to Best Practice. Journal of Black Psychology 2020, Vol. 46(4) 287–311. Doi: 10.1177/0095798420930932

EXPECTED OUTPUTS AND TASKS

Besides the dissertation, in some cases the work could form a publication, or the basis or part of a future publication.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

None needed

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPSYC Contemporary Psychometrics;

PROJECT ID 8

PRIMARY SUPERVISOR

Raquel Iniesta

EMAIL ADDRESS

raquel.iniesta@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Phil Chowienczyk

EMAIL ADDRESS

phil.chowienczyk@kcl.ac.uk

PROJECT TITLE

Exploring the utility of optically recorded pulse waveforms using data from UK Biobank

RESEARCH AREA

Machine Learning for precision medicine

PROJECT AIMS

The aim of this project is to use machine learning models to determine if features of the pulse waveform predict major adverse clinical events (death, heart attack, stroke) by using cross sectional and time to event data.

PROJECT OUTLINE

A pulse waveform can be recorded by optical means from exposed parts of the body (finger, wrist, face) by “wearable” sensors or from a display screen (smart phone, laptop, desktop etc). The waveform is characterised by an initial fast upstroke followed by a slowly decaying component exhibiting an inflection point. There has been much interest in health information that could be derived from such waveforms but, as yet little systematic evaluation of this potential monitoring technology. In this project general and dynamic machine learning models will be developed to determine if features of the pulse waveform predict major adverse clinical events (death, heart attack, stroke) by using cross sectional and time to event data. The main data source (subject to necessary permissions) will be UK Biobank where such waveforms are available together with information on clinical outcomes in approx. 500,000 individuals, but data is also available in other cohorts. Secondly, having identified such features, to relate these to cardiovascular properties (e.g. blood pressure and properties of the heart and arteries derived from imaging) in a sub-sample of approx. 40,000 individuals.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Knowing how to work with real data scenarios

2. Knowing how to use dynamic machine learning algorithms to build predictive models using survival data

3. Knowing the structure of biobank data

STARTER REFERENCES

[1] Jin W, Chowienczyk P, Alastruey J (2021). Estimating pulse wave velocity from the radial pressure wave using machine learning algorithms. PLoS ONE 16(6): e0245026.

[2] Article by Jason Brownlee on August 31, 2016 in XGBoost https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmachinelearningmastery.com%2Ffeature-importance-and-feature-selection-with-xgboost-in-python%2F&amp;data=05%7C01%7Craquel.iniesta%40kcl.ac.uk%7Cf7b6c10e91aa4258c78908da963e6cd0%7C8370cf1416f34c16b83c724071654356%7C0%7C0%7C637987492093793981%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&amp;sdata=m%2FEmHLX3XFc7et58lUa2fs0rSFWVkT3%2FK1VXn5rc6LA%3D&amp;reserved=0

[3] Article by Kurtis Pykes published in Towards Data Science https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Ftowardsdatascience.com%2Foversampling-and-undersampling-5e2bbaf56dcf&amp;data=05%7C01%7Craquel.iniesta%40kcl.ac.uk%7Cf7b6c10e91aa4258c78908da963e6cd0%7C8370cf1416f34c16b83c724071654356%7C0%7C0%7C637987492093950202%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&amp;sdata=Cy%2Bq1Hfltmwb%2FzKY5lUTzK%2Fey2qaHovLET3dJahGyEY%3D&amp;reserved=0

EXPECTED OUTPUTS AND TASKS

The student is expected to understand data aspects, as variables definition and codification, develop a machine learning algorithm (including training and testing) using a language of his choice (usually R and/or Python), produce messures to assess the performance of the algorithm and be able to interpret the results. On a broader perspective, the student is expected to perform a literature review in the field of prediction of cardiovascular events, that will help them properly discuss the results of the present project with regards to previous published works.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Apply to biobank access (that's usually quick to get)

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 9

PRIMARY SUPERVISOR

Raquel Iniesta

EMAIL ADDRESS

raquel.iniesta@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Nadine Seward

EMAIL ADDRESS

nadine.seward@kcl.ac.uk

PROJECT TITLE

Predictors and moderators of a psychological treatment for antenatal depression in improving outcomes of depression

RESEARCH AREA

Machine Learning for personalised medicine

PROJECT AIMS

Main research questions:

-What are the predictors for antidepressant treatment response?

-What are the variables that interacted with treatment allocation that influenced recovery from perinatal depression?

-What are the predictors of potential mediators including: number of sessions attended, perceived social support, and other clinical factors?

PROJECT OUTLINE

A double-blind individually randomised trial was conducted in two antenatal clinics in peri-urban settlement of Khayelitsha, Cape Town. Treatment response to antidepressant drugs was measured using two different outcome measures including the Hamilton Depression Rating Scale and the Edinburgh Postnatal Depression Score (EPDS). The primary outcome was response on the Hamilton Depression Rating Scale at three months postpartum (minimum 40% score reduction from baseline) among participants who did not experience pregnancy or infant loss.

The project seeks to identify predictors of treatment response to add insight into reasons behind why there was no treatment response for some patients. We also wish to explore potential moderators that may have influenced treatment response.

Several state-of-the-art machine learning algorithms including support vector machines and Random Forests will be trained to investigate the best performing model that can estimate response to treatment for a meaningful clinical application.

Findings from this analysis can provide insight into areas that the intervention can target to improve outcomes in future trials.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Knowing how to work with real data scenarios

2. Knowing how to use machine learning algorithms to build predictive models for precision medicine

3. Knowing the structure of data related to clinical trials

STARTER REFERENCES

-Combining clinical variables to optimize prediction of antidepressant treatment outcomes

Raquel Iniesta et al. J Psychiatr Res. 2016 Jul.

-Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables

Raquel Iniesta et al. Sci Rep. 2018.

EXPECTED OUTPUTS AND TASKS

The student is expected to understand data aspects, as variables definition and codification, develop a machine learning algorithm (including training and testing) using a language of his choice (usually R and/or Python), assess the performance of the algorithm and be able to interpret the results. On a broader perspective, the student is expected to perform a literature review in the field of prediction of antidepressant treatment response, that will help them properly frame and discuss the results of the present project with regards to previous published works.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 10

PRIMARY SUPERVISOR

Angus Roberts

EMAIL ADDRESS

angus.roberts@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Tao Wang

EMAIL ADDRESS

tao.wang@kcl.ac.uk

PROJECT TITLE

Restoring clinical timeline using relation extraction from clinical text

RESEARCH AREA

Natural Language Processing, Clinical Knowledge Graph

PROJECT AIMS

This project aims to develop a new natural language processing (NLP) method to extract temporal relations from clinical text, so that clinical timeline can be restored from multiple clinical notes for clinical decision support.

PROJECT OUTLINE

Understanding the clinical timeline is vital in determining a patient's diagnosis and planning treatment [1]. Although electronic health records (EHRs) contain rich, temporal information for patient care, much of the critical information is only recorded in unstructured text documents such as narrative notes [2], e.g., "she had a serious stroke before admission". Natural Language Processing (NLP), a technique enabling automated processing of human language, has been successfully applied to extract fine-grained patient timelines from clinical text to support health research and well-informed care decision making [3]. Temporal relation extraction (TRE), which aims to extract temporal relations among entities or events mentioned in clinical text data (e.g. a medication was given prior to a surgery), has been a prime target for developing automated NLP techniques for clinical text, which a wide range of downstream clinical applications such as detecting adverse drug events, assessing response to treatments and understanding patient trajectories.

This project aims to develop a new NLP method to extract temporal relations from clinical text. Specifically, this project will use open data such as i2b2-2012 [1] and THYME [3] datasets, where temporal relations between medical entities/events are annotated by human experts, for model development and evaluation.

Unlike existing methods which either focus on extracting the relation between events and document creation time (i.e., absolute timelines in terms of calendar time) or extracting relationships between events (i.e., relative timelines in terms of sequence), this project aims to develop a joint model to extract both types of temporal relations by deep learning and multiple-task learning, so that both absolute and relative timelines can be harmonized and a clinically meaningful patient timeline can be restored.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Programming (Python preferable), Machine learning, and Text data analysis.

STARTER REFERENCES

[1] Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. "Evaluating temporal relations in clinical text: 2012 i2b2 challenge." Journal of the American Medical Informatics Association 20.5 (2013): 806-813.

[2] Dalianis, H. (2018). Clinical text mining: Secondary use of electronic patient records. Springer Nature.

[3] Zhao, Shiyi, and Lishuang Li. "Temporal information extraction with the scalable cross-sentence context for electronic health records." Journal of Biomedical Informatics 128 (2022): 104052.

EXPECTED OUTPUTS AND TASKS

Expected output includes dissertation and scientific publications in conferences or journals.

Tasks include: data analysis, programming, result presentation and interpretation, writing.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Online application for data access

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVNLPS Natural Language Processing;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 11

PRIMARY SUPERVISOR

Raquel Iniesta

EMAIL ADDRESS

raquel.iniesta@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Polychronis Pavlidis

EMAIL ADDRESS

polychronis.pavlidis@kcl.ac.uk

PROJECT TITLE

Deploying a machine learning algorithm to identify differential prognostic trajectories in IBD patients

RESEARCH AREA

Machine Learning for personalised medicine

PROJECT AIMS

We aim to test the hypothesis that features collected as part of routine clinical care may be used to segregate patients to groups with different prognosis allowing patients to identify early those patients who may have a less than favourable clinical trajectory.

PROJECT OUTLINE

Personalised, precision medicine approaches are needed to revolutionise management and change outcomes for patients with inflammatory bowel disease (IBD). Comprising of ulcerative colitis (UC) and Crohn's disease (CD), IBD is a chronic immune mediated inflammatory disease of the gastrointestinal tract without a medical cure. Currently, management algorithms are becoming increasingly complicated because there are now multiple different classes of licenced biologics and small molecules. Despite the multiple treatment options available, few patients with inflammatory bowel disease (IBD) achieve sustained remission with any of these agents. The development of robust biomarkers to predict response to these therapies promises to revolutionize the drug selection process in IBD, allowing individual patients to be fast-tracked to the right biological agent on the basis of their probability of responding to it, such that improved outcomes for individual patients can be achieved at the earliest time point possible. The concept of personalized or individualized medicine is a highly sought after goal by clinicians, and is regarded by patients, clinicians and researchers as the number one research priority in IBD, according to the James Lind Alliance IBD research priority-setting partnership.

In this project we aim to test the hypothesis that features collected as part of routine clinical care may be used to segregate patients to groups with different prognosis allowing patients to identify early those patients who may have a less than favourable clinical trajectory. The project will seek to apply an unsupervised and supervised machine learning approach to segregate patients with inflammatory bowel disease in groups of different prognostic trajectory. The clustering will be based on features collected as part of routine clinical care and associate these clusters with long term outcomes of clinical interest.

The project will involve data from:

(a) the GSTT IBD cohort, a longitudinal dataset for 5,000 patients with IBD (~40,000 patient years of follow up) collected in routine clinical care and including clinical indices of disease activity, biochemical, endoscopic and radiological markers of disease activity. Genomic data available in 2,000 patients.

(b) the KCH IBD cohort, a longitudinal data for 2,000 patients with IBD (~20,000 patient years of follow up) collected in routine clinical care and including clinical indices of disease activity, biochemical, endoscopic and radiological markers of disease activity. Genomic data available in 1,000 patients.

(c) the IBD Bioresource. The IBD BioResource is an integral part of the NIHR BioResource for Translational Research which provides the major nationally-accessible resource of over 200,000 volunteers from the general population and patients with common and rare diseases. The NIHR BioResource and IBD BioResource are working collaboratively in supporting studies looking at how genes and other factors may influence diseases. https://www.ibdbioresource.nihr.ac.uk

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Knowing how to work with real data scenarios

2. Knowing how to develop supervised and unsupervised machine learning algorithms

3. Knowing the structure of a large dataset

STARTER REFERENCES

-A retrospective cohort study: pre-operative oral enteral nutritional optimisation for Crohn’s disease in a UK tertiary IBD Centre. Meade, S., Patel, K. V., Luber, R. P., O'Hanlon, D., Caracostea, A., Pavlidis, P., Honap, S., Anandarajah, C., Gryffin, N., Zeki, S., Ray, S., Mawdsley, J., Samaan, M. A., Anderson, S. H., Darakhshan, A., Adams, K., Williams, A., Sanderson, J. D., Lomer, M. & Irving, P. M., Aug 2022, In: Alimentary Pharmacology and Therapeutics. 56, 4, p. 646-663 18 p. Research output: Contribution to journal › Article › peer-review. DOIs: https://doi.org/10.1111/apt.17055

-Cyclin-dependent kinase 9 as a potential target for anti-TNF resistant inflammatory bowel disease. Omer, O. S., Hertweck, A., Roberts, L. B., Lo, J. W., Clough, J. N., Jackson, I., Pantazi, E. D., Irving, P. M., MacDonald, T. T., Pavlidis, P., Jenner, R. G. & Lord, G. M., 1 Jun 2022, (E-pub ahead of print) In: Cellular and molecular gastroenterology and hepatology. Research output: Contribution to journal › Article › peer-review. DOIs: https://doi.org/10.1016/j.jcmgh.2022.05.011

EXPECTED OUTPUTS AND TASKS

The student is expected to preprocess data, understand data aspects, as variables definition and codification, develop a machine learning algorithm (including training and testing) using a language of his choice (usually R and/or Python), assess the performance of the algorithm and be able to interpret the results. On a broader perspective, the student is expected to perform a literature review in the field of prediction of response to IBD treatment, that will help them properly discuss the results of the present project with regards to previous published works.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 12

PRIMARY SUPERVISOR

Angus Roberts

EMAIL ADDRESS

angus.roberts@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Leo Xinyue Zhang

EMAIL ADDRESS

leo.xinyue.zhang@kcl.ac.uk

PROJECT TITLE

Does parameter sharing between named entity recognition and relations extraction models improve model performance for i2b2 medical concept and relation extraction?

RESEARCH AREA

Natural language processing of electronic health records

PROJECT AIMS

Named entity recognition and relation extraction are two crucial tasks for extracting information from electronic health records. These two tasks are traditionally done independently. However, research shows that they could benefit from each other by sharing information during training. One way to do this is to share parameters between the two task models. The aim of this project is to identify if parameter sharing between named entity recognition and relation extraction models help improve model performance for i2b2 medical concept and relation extraction. The objectives include building independent named entity recognition and relation extraction models, build parameter sharing models and compare the reported performance of the two based on i2b2 2018 Track two dataset.

PROJECT OUTLINE

Phase I: literature exploration and self-learning (Dec/2022-Mar/2023)

Phase II: Data exploration and data processing (Apr/2023)

Phase III: Independent NER and RE model building (May/2023-June/2023)

Phase IV: parameter sharing model building (June/2023- July/2023)

Phase V: performance comparison and writing up (July/2023-August/2023)

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Development of neural network based natural language processing

2. Applying NLP to real world electronic health records.

3. Understanding of the messiness of real world data and gain experience on clearing up these data for further use.

4. Researching related literature and online tutorials to solve problems.

STARTER REFERENCES

About the dataset

Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association : JAMIA, 27(1), 3–12. https://doi.org/10.1093/jamia/ocz166

About NER and RE tasks

Dan Jurafsky and James H. Martin, Speech and Language Processing (Chapter 11 and 17), https://web.stanford.edu/~jurafsky/slp3/

An example paper of parameter sharing for NER and RE

Miwa, Makoto and Bansal, Mohit. End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770, 2016. https://arxiv.org/pdf/1601.00770.pdf

Paper and tutorials on one of the widely used NLP models

Paper: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018). https://arxiv.org/abs/1810.04805

Tutorial https://huggingface.co/course/chapter7/2?fw=pt

EXPECTED OUTPUTS AND TASKS

Task 1: Read literature and self-learning related materials.

Output 1: A short literature report

Task 2: Explore the dataset, preprocessing the dataset to make it ready for model input

Output 2: Report on data statistics, pre-processed dataset

Task 3: Build independent named entity recognition model and relation extraction model, and report performance

Output 3: performance report for independent models

Task 4: Build parameter sharing models for named entity recognition model and relation extraction, report and compare performance.

Output 4: performance report for parameter sharing models, performance comparison.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes, public available data - i2b2 2018 Track 2. Available upon finishing online training which takes about a couple of hours.

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Strong Python programming skills are essential for this project, i.e. Python experience prior to this MSc programme. The student will also need to have sufficient background in machine learning and natural language processing prior to starting the project

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7PAVNLPS Natural Language Processing;

PROJECT ID 13

PRIMARY SUPERVISOR

Daniel Stahl

EMAIL ADDRESS

daniel.r.stahl@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Olesya Ajnakina,

EMAIL ADDRESS

olesya.ajnakina@kcl.ac.uk

PROJECT TITLE

Understanding the “Black Box”

RESEARCH AREA

PRedciton modelling, machine learning, explainable AI

PROJECT AIMS

The aim of the project is to interpret a black box prediction model for predicting the risk of developing dementia by using model-agnostic explainability approaches.

PROJECT OUTLINE

Understanding the “Black Box”

Clinical research is becoming increasingly focused on applying machine learning techniques to complex problems. We can read a lot about machine learning, AI and automated clinical decision-making tools. The research has produced some impressive results , such as about the prediction of the risk of the onset of psychoses among young people with mental health problems, the risk of developing dementia or the best treatment for breast cancer.

However, these tools have not been realised in clinical practice. One reason for this lack of translation is increasing concerns about the ethical and medico-legal impact of using machine learning algorithms for medical decision-making. Machine learning algorithms, such as xgboost or neural networks, are often “black box” models that make inscrutable predictions resulting in a lack of trust in prediction models among clinicians and patients. One reason is the possible promotion of biases, which can disadvantage certain groups of patients Often, we are not aware of any data-specific biases that unfairly penalize groups of individuals and it is important to detect and correct them. Famous examples of biases are in facial recognition systems which fail to correctly identify faces of colour or algorithms which present lower-paid job ads to women. Other biases are more subtle, such as using prior use of health care as a surrogate for disease severity. This can lead to racial and social disparities because previous use of resources (i.e. going to the GP) itself was an indicator of barriers to care and existing healthcare inequity and, therefore, does not accurately representative of the need for services. Race/ethnicity are often important predictors in decision tools but may lead to stigmata of ethnic groups, esp. in mental health. To build trust in these models, the system must provide the reasons behind a machine learning decision and open up the operation of the processes for the health organizations, clinicians, service users and the public that are affected by them.

In this project, we will develop a model to predict the onset of dementia by applying an xgboost survival model to a large data set using the English Longitudinal Study of Ageing (ELSA) data set. This data set is a representative sample of the English population above 50 years of age. The data set includes approx. 7500 participants and 197 potential predictor variables about people's physical and mental health, wellbeing, finances and attitudes around ageing.

The aim of the project is to interpret the black box model by using model-agnostic explainability approaches ranging from simple feature variable importance measures, global methods that describe the average behaviour of a model (partial dependency plot or global surrogates) and local methods that explain individual predictions (i.e. Individual Conditional Expectation, Local surrogate models or Shapley values). The output of the project would be an explainable model that automatically produces a report explaining the results of the prediction in a way that is understandable and useful to human users.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Developing a robust prediction model

Develop problem-solving and decision-making skills in clinical prediction modelling

Critical evaluation and understanding of models and its ethical implications

Learn about concepts of dementia

STARTER REFERENCES

Black box and biases

Vokinger, K.N., Feuerriegel, S. & Kesselheim, A.S. Mitigating bias in machine learning for medicine. Commun Med 1, 25 (2021). https://doi.org/10.1038/s43856-021-00028-w

Colin G Walsh, et al (2020) Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence, JAMIA Open, 3(1), 9–15, https://doi.org/10.1093/jamiaopen/ooz054

Heyen, N.B., Salloch, S. (2021) The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. BMC Med Ethics 22, 112. https://doi.org/10.1186/s12910-021-00679-3

ELSA study

https://www.elsa-project.ac.uk/

Ajnakina O, Murray R, Steptoe A, Cadar D. (2022) The long-term effects of a polygenetic predisposition to general cognition on healthy cognitive ageing: evidence from the English Longitudinal Study of Ageing. Psychol Med. 2022 Feb 10:1-9.

doi: 10.1017/S0033291721004827.

Stamate, D., Musto, H., Ajnakina, O., Stahl, D. (2022). Predicting Risk of Dementia with Survival Machine Learning and Statistical Methods: Results on the English Longitudinal Study of Ageing Cohort. IFIP Advances in Information and Communication Technology, vol 652. Springer, Cham. https://doi.org/10.1007/978-3-031-08341-9_35

Methodology

Christoph Molnar (2020) Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

EXPECTED OUTPUTS AND TASKS

The aim of the project is to interpret the black box model by using model-agnostic explainability approaches ranging from simple feature variable importance measures, global methods that describe the average behaviour of a model (partial dependency plot or global surrogates) and local methods that explain individual predictions (i.e. Individual Conditional Expectation, Local surrogate models or Shapley values). The output of the project would be an explainable model that automatically produces a report explaining the results of the prediction in a way that is understandable and useful to human users.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

N/A

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 15

PRIMARY SUPERVISOR

Angus Roberts

EMAIL ADDRESS

angus.roberts@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Margaret Heslin

EMAIL ADDRESS

margaret.heslin@kcl.ac.uk

PROJECT TITLE

Extracting LGBT+ status from the natural language text of mental health records

RESEARCH AREA

Natural language processing of electronic health records

PROJECT AIMS

1. Develop and training models and applications to extract LGBT+ status from CRIS records. This is expected to include baseline NLP methods as well as state-of-the-art context embedding and deep learning models.

2. Designing and executing a robust evaluation methodology for these models.

PROJECT OUTLINE

Background

(LGBT+) have significantly increased risk for mental health problems compared with non LGBT+ individuals (Chakraborty et al, 2011; King et al, 2008; Semlyen et al, 2016). Discrimination relating to sexual orientation (both experienced and anticipated), and trauma appear to contribute to this elevated risk (Chakraborty et al, 2011; Woodhead et al, 2016). The increased trauma, stigma and discrimination experienced by LGBT+ individuals may adversely impact on not only their risk for developing psychological problems but also their ability to benefit from treatment. Indeed, research indicates that lesbian and bisexual women have worse outcomes following treatment by Improving Access to Psychological Therapies (IAPT) services (Rimes et al, 2018). However, little research has been conducted on treatment outcomes following treatment by secondary mental health services. Our ability to investigate this using secondary data is limited due to the lack of structured data on sexuality and gender.

A high proportion of clinically relevant information is, however, held in the unstructured, free text portion of the electronic health record (EHR), which documents letters between clinician, and notes of patient encounters. Natural language processing (NLP) is increasingly used to derive variables from this text for health research, and for real time processing to support clinical care.

This project therefore aims to investigate whether natural language processing can be used to determine sexual and gender status from this unstructured, free text portion of mental health records.

Methods

This project will make use of over 30 million free text patient documents and notes from the South London and Maudsley NHS Foundation Trust (SLaM), available in de-identified form from the NIHR Maudsley Biomedical Research Centres's CRIS database (Perera et al, 2016). NLP is routinely used on CRIS text, with around 80 applications used in a production environment to extract information such as symptoms, cognitive function, social context, and medications. A full range of techniques are used in these applications, including regex, complex rules, simple supervised machine learning, and deep learning.

A sample of the dataset will be pre-labelled by domain experts with mentions of LGBT+ status

The project will include:

- Developing and training models and applications to extract LGBT+ status from CRIS records. This is expected to include baseline NLP methods as well as state-of-the-art context embedding and deep learning models.

- Designing and executing a robust evaluation methodology.

- If there is sufficient time, the project may also include working with CRIS analysts to operationalise the application in the CRIS NLP production service, and using outputs in a small cross sectional study.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Knowledge of the full NLP development lifecycle.

2. Experience of working with messy, real world medical record text.

3. Experience of working alongside epidemiologists and health service staff.

STARTER REFERENCES

Dataset:

Perera et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource BMJ Open 2016;6:e008721.

NLP of the EHR:

Chapter 17 in Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky/slp3/

Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. 1st ed. 2018. Springer International Publishing; 2018.

LGBTQ+ status and mental health:

Chakraborty, A., McManus, S., Bhugra, T. S., Bebbington, P. and King, M.

(2011). Mental health of the non-heterosexual population of England.

British Journal of Psychiatry, 198, 143–148. doi:

10.1192/bjp.bp.110.082271

King, M., Semlyen, J., Tai, S. S., Killaspy, H., Osborn, D., Popelyuk,

D. and Nazareth, I. (2008). A systematic review of mental disorder,

suicide, and deliberate self harm in lesbian, gay and bisexual people.

BMC Psychiatry, 8, 70. doi: 10.1186/1471-244X-8-70

Rimes KA, Broadbent M, Holden R, Rahman Q, Hambrook D, Hatch SL,

Wingrove J. Comparison of treatment outcomes between lesbian, gay,

bisexual and heterosexual individuals receiving a primary care

psychological intervention. Behavioural and cognitive psychotherapy.

2018 May;46(3):332-49.

Semlyen, J., King, M., Varney, J. and Hagger-Johnson, G. (2016).

Sexual orientation and symptoms of common mental disorder or low wellbeing:

combined meta-analysis of 12 UK population health surveys. BMC

Psychiatry, 16, 67. doi: 10.1186/s12888-016-0767-z

Woodhead, C., Gazard, B., Hotopf, M., Rahman, Q., Rimes, K. A. and

Hatch, S. L. (2016). Mental health among UK inner city

non-heterosexuals: the role of risk factors, protective factors and

place. Epidemiology and Psychiatric Sciences, 25, 450–461. doi:

10.1017/S2045796015000645

EXPECTED OUTPUTS AND TASKS

1. A working NLP application to extract LGBT+ status from CRIS text.

2. An evaluation of this application.

3. If time allows, use of extracted data in a small study.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes, CRIS

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Letter of access; SLaM NHS Foundation Trust network account

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7PAVNLPS Natural Language Processing;

PROJECT ID 16

PRIMARY SUPERVISOR

Daniel Stahl

EMAIL ADDRESS

daniel.r.stahl@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Diana Shamsutdinova

EMAIL ADDRESS

diana.shamsutdinova@kcl.ac.uk

PROJECT TITLE

Development of the user-friendly approach to test and explain non-linearity in longitudinal health data using ensembles of classical and machine learning methods.

RESEARCH AREA

Prediction modelling, survival analysis.

PROJECT AIMS

To develop a user-friendly function in R language that 1) fits and internally validates several ensemble methods and a baseline Cox model, and 2) makes inferences on the presence of non-linear dependencies in the data structure.

PROJECT OUTLINE

Survival analysis is a collection of methods used for time-to-event outcomes such as death, disease onset, recovery and others. Classical Cox Proportionate Hazard model developed over 50 years ago has proved to be one of the most popular survival analysis models. It has important advantages such as robustness, low computational load and easy interpretation. However, data with more complex relationships between the outcome and risk factors may require more elaborate methods. Machine learning methods such as classification trees, random forests and neural networks have recently been extended to accommodate survival data and can fit complex data, though those models often lack prediction transparency.

Here, we will work on further developing ensemble methods that combine classical and tree-based ensemble methods in a way that improves prediction performance and preserves interpretability. The primary goal of this project would be creating a user-friendly function that health researchers may apply to their data and test if non-linear algorithms outperform classical linear model, quantify marginal prediction performance and try to locate data non-linearities.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Working on this project will advance student's knowledge of the classical and modern survival methods widely used to analyse longitudinal time-to-event data. Student will improve/gain R programming skills, and learn and apply various prediction modelling techniques.

STARTER REFERENCES

[1] George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol 2014;21:686–94. https://doi.org/10.1007/s12350-014-9908-2.

[2] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. Br J Cancer 2003;89:781–6. https://doi.org/10.1038/sj.bjc.6601117.

[3] Ishwaran, H., Lauer, M.S., Blackstone, E.H., Lu, M.: randomForestSRC: Random Survival Forests Vignette (2021)

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA (1984)

[5]Shamsutdinova, D., Stamate, D., Roberts, A., & Stahl, D. (2022). Combining Cox Model and Tree-Based Algorithms to Boost Performance and Preserve Interpretability for Health Outcomes. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 170-181). Springer, Cham.

EXPECTED OUTPUTS AND TASKS

Expected tasks in this project are:

1) familiarize with the existing survival models [e.g. starter references 1-4]

2) study the ensemble methods described in [5] paper

3)[main goal] transform or re-write existing R code into a user-friendly function targeted for health research professionals

4) develop output format such that it

a) displays relative performance of linear (Cox) and non-linear methods and

b) makes automated inference if the baseline model is good enough or non-linear methods are better suited to model the data

c) suggests where non-linearities are

5) test the function on various simulated and real-life data

6)[bonus task if time permits] add other baseline and interpretable ensemble methods with similar aims of disentangling linear and more complex dependencies and improve prediction performance

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

there are simulated data that can be used and openly available survival-type data for testing the function

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

access to some datasets may be needed to test the methods

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

1-2

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVITHI Introduction to Health Informatics;7PAVMALM Multilevel and Longitudinal Moddelling;7PAVAIHA Artificial Intelligence for Health Analytics;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 17

PRIMARY SUPERVISOR

Angus Roberts

EMAIL ADDRESS

angus.roberts@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Grace Crowley

EMAIL ADDRESS

Grace.Crowley@slam.nhs.uk

PROJECT TITLE

Text mining health records to understand the experiences of forced migrants accessing secondary mental health services

RESEARCH AREA

The project has the potential to span natural language processing, epidemiology, mental health and population science

PROJECT AIMS

To develop and evaluate NLP models on an annotated training dataset to identify experiences of forced migration mentioned in CRIS, a mental health case register. Beyond this project, the models will be used in population mental health research.

PROJECT OUTLINE

Mental health clinical records can provide a rich repository of free text information detailing the clinical characteristics and social circumstances of people accessing mental health services. Structured fields related to migration status, however, are poorly completed. The mental health needs of forced migrants (including people seeking asylum and those with refugee status) are potentially much greater compared to those who have not had these experiences. Forced migrants are more likely to have experienced adversity such as conflict, torture, sexual exploitation and dangerous journeys, compared to the host population and economic migrants. Once in the UK, they may face challenges to everyday life such as socioeconomic deprivation, unsuitable accommodation and discrimination, as well as psychological difficulties related to uncertainty, mistrust and threats of detention or deportation. In addition, there are a myriad of barriers to accessing healthcare and psychological support. It is not surprising that the rates of mental illness are high among this population, yet very little is known about those who access secondary mental health services or their experiences of these encounters.

In this project, we will use free text data from clinical mental health records to inform the development of Natural Language Processing (NLP) algorithms, which may be used to identify people who have experienced forced migration. The work will inform a larger project on mental health course and outcomes in refugee and asylum seeker groups. The algorithms will go on to be used to explore clinical characteristics, service use, migration-related factors and social circumstances among this population, with the broader aim of informing improvements in mental health services for people who have experienced forced migration. This is a timely project in light of the on-going hostile environment policy and the recent introduction of the Nationality and Borders Act, which has the potential to impact on the mental health of those seeking asylum in the UK.

Development will use a pre-existing set of labelled mentions of forced migration in the CRIS dataset. This will be used to train and evaluate several machine learning models for extraction of these mentions. Baseline "keyword" string matching based models will be compared to support vector classifiers and embedding based neural net models. It is expected that the complexity of the language may require some work on feature engineering. Evaluation will be on a held-out blind dataset. If time allows, the student may also carry out a small prevalence study on a subset of CRIS.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. A working knowledge of the NLP development process.

2. Knowledge of working with real-world healthcare data, including text.

3. Experience of collaborating directly with clinical and other hospital staff on the dataset.

STARTER REFERENCES

Dataset:

Perera G, Broadbent M, Callard F, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):e008721-. doi:10.1136/bmjopen-2015-008721

NLP of the EHR:

Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. 1st ed. 2018. Springer International Publishing; 2018.

Information Extraction:

Chapter 17 in Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin https://web.stanford.edu/~jurafsky/slp3/

Mental health and forced displacement:

Pollard T, Howard N. Mental healthcare for asylum-seekers and refugees residing in the United Kingdom: a scoping review of policies, barriers, and enablers. Int J Ment Health Syst. 2021 Jun 14;15(1):60. doi: 10.1186/s13033-021-00473-z. PMID: 34127043; PMCID: PMC8201739.

Jannesari S, Molyneaux E, Lawrence V. What affects the mental health of people seeking asylum in the UK? A narrative analysis of migration stories. Qualitative Research in Psychology. 2019 Mar 7; 19:2, 295-315. doi: 10.1080/14780887.2019.1581311

Satinsky E, Fuhr DC, Woodward A, Sondorp E, Roberts B. Mental health care utilisation and access among refugees and asylum seekers in Europe: A systematic review. Health Policy. 2019 Sep;123(9):851-863. doi: 10.1016/j.healthpol.2019.02.007. Epub 2019 Feb 22. PMID: 30850148.

EXPECTED OUTPUTS AND TASKS

1. NLP models for the extraction of mentions of forced migration in EHRs.

2. An evaluation of these models.

3. If time allows, a small cross sectional / period prevalence study on forced migration in a portion of the dataset.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

CRIS dataset

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Letter of access, hospital network account

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7PAVNLPS Natural Language Processing;

PROJECT ID 18

PRIMARY SUPERVISOR

Nicholas Cummins

EMAIL ADDRESS

nickcummins41@gmail.com

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Depend on the student and their aims

EMAIL ADDRESS

N/A

PROJECT TITLE

Remote speech-based health assessment

RESEARCH AREA

Speech Processing, Machine Learning

PROJECT AIMS

1. Identify properties of speech indicative of changes in health status

2. Explore factors affecting the quality of remotely recorded speech

PROJECT OUTLINE

Speech is uniquely placed as a digital phenotype; no other signal contains its singular combination of cognitive, neuromuscular and physiological information. Models developed in speech studies have the real potential to provide unique preventive and predictive information about mental health to provide opportunities for enhanced self-management or screening services. These advantages aside, the potential of speech as a digital phenotype of mental health is yet to be realised.The work in this project will focus on quantify such changes on a new large speech databases collected remotely using smartphones. It is a unique opportunity to help determine if findings in the literature, collected in very controlled laboratory studies, are transferable into larger-scale remote monitoring studies and help realise the potential of speech as core personalised and precision medicine signal.

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Gain skills in organising and structuring real-world data for statistical and machine learning modelling

2. Acquire and expand data analytics knowledge and competencies

3. Gain new knowledge in the study of speech signals and related processing methods

STARTER REFERENCES

- Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology. 2020 Feb;5(1):96-116.

- Cummins N, Baird A, Schuller BW. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. Methods. 2018 Dec 1;151:41-54.

- Noffs G, Perera T, Kolbe SC, Shanahan CJ, Boonstra FM, Evans A, Butzkueven H, van der Walt A, Vogel AP. What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis. Autoimmunity reviews. 2018 Dec 1;17(12):1202-9.

- Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J, Uher T, Vaneckova M, Andelova M, Novotna K, Kadrnozkova L. Characteristics of motor speech phenotypes in multiple sclerosis. Multiple sclerosis and related disorders. 2018 Jan 1;19:62-9.

- Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication. 2015 Jul 1;71:10-49.

EXPECTED OUTPUTS AND TASKS

Task can vary from more conventional statical modelling through to contemporary deep learning model. The exact nature of the analysis is discussed and decided upon in conjunction with students who undertake the project.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

none

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

3

USEFUL OPTIONAL MODULES

7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7PAVITHI Introduction to Health Informatics;

PROJECT ID 19

PRIMARY SUPERVISOR

Silia Vitoratou

EMAIL ADDRESS

silia.vitoratou@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Jane Gregory

EMAIL ADDRESS

jane.gregory@linacre.ox.ac.uk

PROJECT TITLE

Evaluating psychometric assessments for misophonia

RESEARCH AREA

Psychometrics, latent variable modelling, structural equation modelling

PROJECT AIMS

To identify a set of items suitable for the measurement of misophonia

PROJECT OUTLINE

Misophonia, acknowledged as a disorder by consensus in 2021 (1), has attracted increased interest in the literature the past five years. To sufficiently study the disorder, its prevalence, and the effect of treatments, it is essential to have reliable and valid tools for its measurement. At the Psychometrics and Measurement lab we use state-of-the-art methods in contemporary psychometrics to provide tools for the measurement of misophonia. We initiated a large study which recently receive a substantial grant to proceed with our goals. We have established a team of all levels of expertise (including students) who help with the tasks that are listed below and we are happy to welcome more members to our team.

We currently focus our efforts on a large sample representative of the US population, to ensure the generalizability of the results and a large sample of people who identify with misophonia. As a starting point, we will use the S-Five (2), a short and robust measurement tool for the severity of misophonia, which has been shown to be reliable and valid in nine independent samples (in six languages). We also consider the Duke Misophonia Questionnaire (DMQ; (3)), the Revised Amsterdam Misophonia Scale (AMISOS-R; (4)), and the Misophonia Questionnaire (MisoQuest (5)).

For each measurement tool we will evaluate its dimensionality (structural validity) and reliability (internal consistency, stability), and we will present for the first time the norms of the corresponding scale scores for the US population. We will investigate each tool's ability to discriminate between those with and without the disorder and, based on these results, we will proceed with the estimation of the misophonia prevalence for the US population (discriminative validity). We will thoroughly explore the construct validity (convergent and discriminant validity) of each measure. Findings on the nomological network of misophonia will be presented, in relation to the four measures and measures of mental health conditions, auditory disorders, and general psychopathology. At the last phase of the proposed study, we aim to synthesize the utilities of the existing tools for the development of a combined measurement tool, a self-assessment to assess the severity of symptoms, suitable for research and clinical purposes.

1. Swedo S, Baguley DM, Denys D, Dixon LJ, Erfanian M, Fioretti A, et al. A Consensus Definition of Misophonia: Using a Delphi Process to Reach Expert Agreement. medRxiv. 2021:2021.04.05.21254951.

2. Vitoratou S, Uglik-Marucha N, Hayes C, Gregory J. Listening to people with misophonia: exploring the multiple dimensions of sound intolerance using a new psychometric tool, the S-Five, in a large sample of individuals identifying with the condition. Psych. 2021;3:639-62.

3. Rosenthal MZ, Anand D, Cassiello-Robbins C, Williams ZJ, Guetta RE, Trumbull J, et al. Development and Initial Validation of the Duke Misophonia Questionnaire. Frontiers in Psychology. 2021;12(4197).

4. Jager I, de Koning P, Bost T, Denys D, Vulink N. Misophonia: Phenomenology, comorbidity and demographics in a large sample. PLoS One. 2020;15(4):e0231390.

5. Siepsiak M, Sliwerski A, Lukasz Dragan W. Development and Psychometric Properties of MisoQuest-A New Self-Report Questionnaire for Misophonia. International Journal of Environmental Research and Public Health. 2020;17(5).

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

- Write up of ethics applications

- Online survey development and dissemination

- Clean and prepare data

- Conduct advanced psychometric analyses and other SEM analyses

- Write up reports and papers

STARTER REFERENCES

Swedo S, Baguley DM, Denys D, Dixon LJ, Erfanian M, Fioretti A, et al. A Consensus Definition of Misophonia: Using a Delphi Process to Reach Expert Agreement. medRxiv. 2021:2021.04.05.21254951.

Vitoratou S, Uglik-Marucha N, Hayes C, Gregory J. Listening to people with misophonia: exploring the multiple dimensions of sound intolerance using a new psychometric tool, the S-Five, in a large sample of individuals identifying with the condition. Psych. 2021;3:639-62.

EXPECTED OUTPUTS AND TASKS

- Sample data for the 5 questionnaires

-Conduct part of the analyses

-Assist with report and paper writing (co-authorship expected should the tasks be completed)

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Some yes, some to be collected

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

Yes

STUDENT REQUIREMENTS

NA

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVPSYC Contemporary Psychometrics;7PAVMALM Multilevel and Longitudinal Moddelling;7PAVCIAE Causal Modelling and Evaluation;

PROJECT ID 20

PRIMARY SUPERVISOR

Shaoxiong Sun

EMAIL ADDRESS

shaoxiong.sun@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Richard Dobson and Amos Folarin

EMAIL ADDRESS

richard.j.dobson@kcl.ac.uk; amos.folarin@kcl.ac.uk

PROJECT TITLE

A machine/deep learning approach to infer depressive status from data collected through wearable devices and smartphones

RESEARCH AREA

Mobile health; Machine learning; Deep learning

PROJECT AIMS

To develop machine/deep learning approaches to estimate depressive symptom severity from mobile health data

PROJECT OUTLINE

Major Depressive Disorder (aka depression) is a potentially debilitating disease affecting 264 million people worldwide. Its recurring nature makes it important to monitor depressive patients continuously and provide timely interventions when needed. In this project, data were collected from 623 patients suffering from depression over a duration of two years. These patients were equipped with Fitbit devices from which step count, heart rate, and sleep data were gathered. This information, pilot studied by multiple research groups, has been shown to be indicative of depressive status. Precisely, the research question is how to use individual data modality of step count, heart rate, and sleep, or their combination to estimate patients’ depressive status, which is measured through PHQ-8 questionnaire. This study focuses on machine/deep learning approaches and mines the data in fine granularity to extract informative and predictive patterns.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Understanding of roles of mobile devices in mental health

2. Data cleaning (e.g. handling missing data)

3. Hands-on experience of applying machine/deep learning technique in mobile health

STARTER REFERENCES

Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. JMIR Mhealth Uhealth 2018;6(8):e165

Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. JMIR Mhealth Uhealth 2021;9(4):e24604

EXPECTED OUTPUTS AND TASKS

1. To understand literature

2. To clean and preprocess data

3. To extract features and/or develop feature representations

4. To explore existing and develop suitable machine/deep learning models

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 21

PRIMARY SUPERVISOR

Diana Shamsutdinova

EMAIL ADDRESS

diana.shamsutdinova@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Daniel Stahl

EMAIL ADDRESS

daniel.r.stahl@kcl.ac.uk

PROJECT TITLE

Development of a user-friendly approach to test and explain non-linearity in longitudinal health data using ensembles of classical and machine learning methods.

RESEARCH AREA

Prediction modelling, survival analysis, machine learning ensemble methods

PROJECT AIMS

To develop a user-friendly function in R language that 1) fits and internally validates several ensemble methods and a baseline Cox model, and 2) makes inferences on the presence of non-linear dependencies in the data structure.

PROJECT OUTLINE

Survival analysis is a collection of methods used for time-to-event outcomes such as death, disease onset, recovery, and others. Classical Cox Proportionate Hazard model developed over 50 years ago has proved to be one of the most popular survival analysis models. It has important advantages such as robustness, low computational load and easy interpretation. However, data with more complex relationships between the outcome and risk factors may require more elaborate methods. Machine learning methods such as classification trees, random forests and neural networks have recently been extended to accommodate survival data and can fit complex data, though those models often lack prediction transparency.

Here, we will work on further developing ensemble methods that combine classical and tree-based ensemble methods in a way that improves prediction performance and preserves interpretability. The primary goal of this project would be creating a user-friendly function that health researchers may apply to their data and test if non-linear algorithms outperform classical linear model, quantify marginal prediction performance, and try to locate data non-linearities.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Working on this project will advance student's knowledge of the classical and modern survival methods widely used to analyse longitudinal time-to-event data. Student will improve/gain R programming skills, and learn and apply various prediction modelling techniques.

STARTER REFERENCES

References:

[1] George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol 2014;21:686–94. https://doi.org/10.1007/s12350-014-9908-2.

[2] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. Br J Cancer 2003;89:781–6. https://doi.org/10.1038/sj.bjc.6601117.

[3] Ishwaran, H., Lauer, M.S., Blackstone, E.H., Lu, M.: randomForestSRC: Random Survival Forests Vignette (2021)

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA (1984)

[5]Shamsutdinova, D., Stamate, D., Roberts, A., & Stahl, D. (2022). Combining Cox Model and Tree-Based Algorithms to Boost Performance and Preserve Interpretability for Health Outcomes. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 170-181). Springer, Cham.

EXPECTED OUTPUTS AND TASKS

Expected tasks in this project would be:

1) familiarising with the existing survival models

2) study the ensemble methods described in [5] paper

3) transform or re-write existing R code into a user-friendly function targeted for health research professionals

4) develop output format such that it

a) displays relative performance of linear (Cox) and non-linear methods and

b) makes automated inference if the baseline model is good enough or non-linear methods are better suited to model the data

c) suggests where non-linearities are

5) test the function on various simulated and real-life data

6)* add other baseline and interpretable ensemble methods with similar aims of disentangling linear and more complex dependencies and improve prediction performance

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

there are simulated data that can be used and openly available survival-type data for testing the methods

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

may be required if a new data set will required for testing the methods, but not necessary

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Maybe

NUMBER OF MSC STUDENTS PREFERRED

1-2

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 22

PRIMARY SUPERVISOR

Shaoxiong Sun

EMAIL ADDRESS

shaoxiong.sun@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Richard Dobson, Amos Folarin, Jonna Kuntsi

EMAIL ADDRESS

richard.j.dobson@kcl.ac.uk; amos.folarin@kcl.ac.uk; jonna.kuntsi@kcl.ac.uk

PROJECT TITLE

The analysis of movement patterns for people with attention deficit hyperactivity disorder (ADHD) using remote monitoring technology

RESEARCH AREA

Mobile health; remote monitoring technology; ADHD; mental health

PROJECT AIMS

To examine the utility of remote monitoring technologies for measuring characteristics associated with attention deficit hyperactivity disorder (ADHD).

PROJECT OUTLINE

The availability of remote monitoring technology (RMT) such as smartphones and wearable devices provides the opportunity to remotely monitor people with attention deficit hyperactivity disorder (ADHD) in real time for extended periods of time. Preliminary research showed the promise of gait patterns in controlled dual-task paradigms in the laboratory and questionnaire-measured smartphone use as ADHD-sensitive biomarkers. In this project, we will examine differences between individuals with ADHD and controls in gait patterns under typical real-world dual-task conditions such as walking while using phones and step count pattern in daily life. We will work on an available dataset with smartphone and wearable device data remotely collected from 20 individuals with ADHD and 20 control participants for 10 weeks.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Data cleaning

2. Gait analysis from smartphone acceleration signal

3. Movement pattern analysis from Fitbit step count

4. Statistical analysis and machine learning

STARTER REFERENCES

Simons L, Valentine A, Falconer C, Groom M, Daley D, Craven M, Young Z, Hall C, Hollis C

Developing mHealth Remote Monitoring Technology for Attention Deficit Hyperactivity Disorder: A Qualitative Study Eliciting User Priorities and Needs. JMIR Mhealth Uhealth 2016;4(1):e31. URL: https://mhealth.jmir.org/2016/1/e31

EXPECTED OUTPUTS AND TASKS

1. Extract gait features and step count features

2. Conduct statistical analysis on the derived features to identify group differences

3. Conduct machine learning on the derived features to classify different groups

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVITHI Introduction to Health Informatics;

PROJECT ID 23

PRIMARY SUPERVISOR

Shaoxiong Sun

EMAIL ADDRESS

shaoxiong.sun@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Richard Dobson, Amos Folarin, Callum Stewart, Yatharth Ranjan

EMAIL ADDRESS

richard.j.dobson@kcl.ac.uk; amos.folarin@kcl.ac.uk; callum.stewart@kcl.ac.uk; yatharth.ranjan@kcl.ac.uk

PROJECT TITLE

The analysis of massive mobile health data to understand and address key issues in relation to COVID-19

RESEARCH AREA

COVID-19, Mobile health, Remote monitoring technology; Data analytics; Machine learning; Deep learning

PROJECT AIMS

To analyse data collected from smartphones and wearable devices to study their utility in understanding and addressing key issues in relation to COVID-19

PROJECT OUTLINE

There have been 611 million confirmed cases of COVID-19, leading to 6.52 million deaths. The resulting physical and mental suffering and financial costs of the pandemic highlight the need for mobile health and remote monitoring technology, which have a potential to curb the spread of the diseases, understand its aetiology and symptomatology, and provide tailored and personalised treatments. In response to this need, our group initiated Covid Collab (https://covid-collab.org/), an observational mobile health study beginning in June 2020. 17000+ participants enrolled through the study app, Mass Science, were prompted to complete regular surveys on COVID symptoms experienced, vaccination and diagnosis status, mood, and mental well-being. In this master's project, you are encouraged to study how different data modalities sensed by smartphones and wearable devices can help in detecting COVID infection, identifying the trajectory of the Long COVID, understanding physiological response to vaccination, and examining the variations of mental health during the pandemic. These data modalities include GPS location, battery level, activity level, step count, heart rate, and sleep.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Data cleaning

2. Feature engineering and representation

3. Statistical analysis

4. Machine/deep learning

STARTER REFERENCES

Stewart C, Ranjan Y, Conde P, Rashid Z, Sankesara H, Bai X, Dobson R, Folarin A

Investigating the Use of Digital Health Technology to Monitor COVID-19 and Its Effects: Protocol for an Observational Study (Covid Collab Study). JMIR Res Protoc 2021;10(12):e32587 URL: https://www.researchprotocols.org/2021/12/e32587

Sun S, Folarin A, Ranjan Y, Rashid Z, Conde P, Stewart C, Cummins N, Matcham F, Dalla Costa G, Simblett S, Leocani L, Lamers F, Sørensen P, Buron M, Zabalza A, Guerrero Pérez A, Penninx B, Siddi S, Haro J, Myin-Germeys I, Rintala A, Wykes T, Narayan V, Comi G, Hotopf M, Dobson R, RADAR-CNS Consortium

Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19

J Med Internet Res 2020;22(9):e19992

EXPECTED OUTPUTS AND TASKS

You are expected to select one from the above-mentioned research topics (detecting COVID infection, identifying the trajectory of the Long COVID, understanding physiological response to vaccination, and examining the variations of mental health during the pandemic). You will do feature engineering and conduct statistical analysis and machine/deep learning to answer the research question of your interest.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALM Multilevel and Longitudinal Moddelling;7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 24

PRIMARY SUPERVISOR

Petroula Laiou

EMAIL ADDRESS

petroula.laiou@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Callum Stewart

EMAIL ADDRESS

callum.stewart@kcl.ac.uk

PROJECT TITLE

Estimating the brain regions that seizures in start using machine learning

RESEARCH AREA

Machine learning, Neuroscience, Signal processing

PROJECT AIMS

The aim of this project is to investigate whether we can estimate the brain hemisphere that seizures start in epilepsy patients using resting state electroencephalographic (EEG) recordings. Thus, the candidate student will 1) analyse EEG recordings from epilepsy patients whose seizure onset is on the left or right hemisphere 2) apply signal preprocessing techniques 3) extract quantitative features from the EEG signals and 4) develop a machine learning classification algorithm to classify the hemisphere of the seizure onset.

PROJECT OUTLINE

Epilepsy is a common neurological disorder that is characterised by the occurrence of repeated seizures. For an accurate epilepsy diagnosis and treatment, it is crucial that clinicians know the brain regions that seizure start. Hence, often times epilepsy patients undergo a prolonged hospitalization with continuous video-electroencephalography (video-EEG) monitoring. In this process, patients are wearing 24/7 EEG caps that monitor brain activity and produce continuous electroencephalographic recordings. Clinicians assess the seizure activity from the EEG recordings and make conclusions regarding the brain region of the seizure onset. Although this process might provide some insights about seizures, it is not always successful because often patients do not experience seizures during their hospitalization. Hence, a quantitative tool that estimates the brain region that seizure start using non-seizure EEG data (interictal data) is in great demand.

In this project the candidate student will analyse longitudinal, interictal EEG recordings from epilepsy patients whose seizure onset is either on the left or right hemisphere. From those EEG recordings several features will be extracted that quantify the power of the signals as well as the brain functional structure (phase locking value, graph theory measures). Using machine learning classification algorithms (logistic regression, SVM, random forest) the candidate student will investigate whether such algorithms can lateralize the hemisphere of the seizure onset. A positive project outcome would suggest that the seizure onset could be characterised from non-seizure EEG data with quantitative tools and machine learning. Therefore, such methods can serve as a support clinical tool for epilepsy diagnosis and seizure lateralization.

PROJECT TYPE

Wet Lab

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

The student will develop skills on 1) signal preprocessing of electroencephalographic recordings, 2) computation of functional network measures (phase locking value) and graph theory network metrics 3) machine learning classification algorithms. In addition, the student will gain general knowledge about epilepsy.

STARTER REFERENCES

1) Automated diagnosis of temporal lobe epilepsy in the absence of interictal spikes, Verhoeven et al 2018

2) Temporal evolution of multiday, epileptic functional networks prior to seizure occurrence, Laiou et al 2022

EXPECTED OUTPUTS AND TASKS

It is expected that the candidate student will develop a machine learning classification algorithm (logistic regression, SVM). In particular, with the guidance and support of the supervisors the student will 1) apply signal preprocessing techniques on the EEG recordings, 2) compute features from the signals 3) develop and evaluate a Machine Learning classification algorithm. It is expected that the candidate student has strong programming skills in Python and/or MATLAB.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

The dataset has already been collected via the European RADAR-CNS study, and hence no data collection is needed.

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

Yes

STUDENT REQUIREMENTS

N/A

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVCPNS Computational Neuroscience;7PAVMALE Machine Learning for Health and Bioinformatics;7BBG2016 Advanced Bioinformatics: practical bioinformatics data skills;

PROJECT ID 25

PRIMARY SUPERVISOR

Prof Sabine Landau

EMAIL ADDRESS

sabine.landau@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Professor Declan McLoughlin / Ana Jelovac

EMAIL ADDRESS

d.mcloughlin@tcd.ie / ajelovac@tcd.ie

PROJECT TITLE

Mediation and moderation modelling of response and side effects to electroconvulsive therapy in severe depression

RESEARCH AREA

Electroconvulsive Therapy (ECT) is the most acutely effective treatment for treatment resistant, sometimes life-threatening, depression (Kirov et al, 2021). Nevertheless, its use remains controversial, mainly because of cognitive side effects, especially

PROJECT AIMS

Secondary data analyses of the ECT observational database assembled by the clinical team at St Patrick's University Hospital in Dublin in order to assess various mechanistic hypotheses regarding how ECT works and who might experience cognitive side effects.

PROJECT OUTLINE

This current collaboration between the Department of Biostatistics at King's College London and the Department of Psychiatry & Trinity College Institute of Neuroscience at Trinity College Dublin aims to investigate ECT mechanistic hypotheses in three broad areas.

1. Mediation analysis to understand why older severely depressed patients benefit more from ECT than younger patients

It is well established that older patients benefit more from ECT. Why is this? One mechanism that has been proposed is that the illness profile of severe depression varies with age. For example, psychomotor disturbance (characterised by psychomotor retardation or agitation) and psychotic features may be more common in older adults and this illness profile may, in turn, increase ECT response (Heijnen et al., 2019; van Diermen et al., 2020). Thus, we might hypothesise that there is a pathway from age to an aspect of the illness profile to ECT depression response. Such mechanistic hypotheses can be formally assessed by using mediation analysis. In this context, the choice of the putative mediator variable is crucial. Some putative mediator variables might be proposed by existing theories (Heijnen et al., 2019; van Diermen et al., 2020; Waite et al., 2022). It might also be possible to model the whole multidimensional illness profile via a latent mediator variable (as done in the structural equation modelling approach by van Diermen et al, 2020) or by using a two-stage approach whereby the illness profile data are summarized by a categorical or predicted latent variable first, and then this variable is considered as the mediator in a second stage.

If we can understand the mechanisms by which older patients benefit more from ECT, treatment delivery could be targeted to patients most likely to respond and avoid exposing patients unlikely to benefit to the unnecessary risk of cognitive side effects.

2. Baseline demographic and clinical predictors of cognitive side effects

The clinical team have assembled a second database of patients with severe depression but not receiving ECT, providing the opportunity to compare retrograde amnesia between those receiving and not receiving ECT. First, we would expect the severity of retrograde amnesia to be stronger in those receiving ECT as retrograde amnesia is a known side effect of ECT treatment. Second, several baseline variables might be hypothesized to predict the severity of retrograde amnesia for depressed patients, such as age, gender, educational attainment, and psychotic symptoms. Third, the strength of these links might vary between depressed patients who do or do not receive ECT treatment. Such prediction and moderation hypotheses could be addressed by relevant interaction modelling.

3. Do ECT treatment parameters or early ECT session outcomes (time to reorientation) predict post-treatment retrograde amnesia?

In those treated with ECT, treatment parameters or clinical characteristics of the ECT sessions might predict longer-term side effects. For example, we have already established in the EFFECT-Dep Trial that bitemporal ECT causes stronger retrograde amnesia than unilateral ECT. It is also possible that shorter times to re-orientation after ECT sessions indicate increased patient resilience and thus predict less retrograde amnesia (Martin et al., 2015; Sobin et al., 1995). Such prediction hypotheses could be addressed by relevant statistical modelling approaches. Again, the treatment session information is multidimensional (several variables assessed longitudinally) and multivariate techniques could be explored (e.g. as done in latent class analysis by Rhebergen et al., 2015) to reduce dimensionality and empirically identify relevant session predictor variables.

If we can discover which baseline or session characteristics predict retrograde amnesia after ECT, we may be able to modify ECT delivery to minimise such side effects.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

1. Develop your understanding of advanced modelling techniques

2. Learn how to use methods from the field of causal inference to address real-life research questions regarding ECT for depression

3. Learn how to plan a research project

4. Develop skills in typical work processes, in particular the process of translating clinical questions into quantitative research hypotheses and the identification of appropriate analysis approaches to address them.

5. Gain experience of working in a medical research environment

6. Take part in a real-life collaboration providing experience of working in a multidisciplinary team consisting of statisticians/data scientists and clinical researchers

7. Provide the opportunity to contribute to research publication(s)

STARTER REFERENCES

1. Heijnen, W. T. C. J., Kamperman, A. M., Tjokrodipo, L. D., Hoogendijk, W. J. G., van den Broek, W. W., & Birkenhager, T. K. (2019). Influence of age on ECT efficacy in depression and the mediating role of psychomotor retardation and psychotic features. J Psychiatr Res, 109, 41-47.

https://doi.org/10.1016/j.jpsychires.2018.11.014

2. Kirov G, Jauhar S, Sienaert P, Kellner CH, McLoughlin DM (2021) Electroconvulsive therapy for depression: 80 years of progress. Br J Psychiatry Nov;219(5):594-597.

https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/electroconvulsive-therapy-for-depression-80-years-of-progress/EA419A2EDF02EB803D8417B437779060

3. Martin, D. M., Galvez, V., & Loo, C. K. (2015). Predicting Retrograde Autobiographical Memory Changes Following Electroconvulsive Therapy: Relationships between Individual, Treatment, and Early Clinical Factors. Int J Neuropsychopharmacol, 18(12).

https://doi.org/10.1093/ijnp/pyv067

4. Rhebergen, D., Huisman, A., Bouckaert, F., Kho, K., Kok, R., Sienaert, P., Spaans, H. P., & Stek, M. (2015). Older age is associated with rapid remission of depression after electroconvulsive therapy: a latent class growth analysis. Am J Geriatr Psychiatry, 23(3), 274-282.
https://www.sciencedirect.com/science/article/abs/pii/S1064748114001432?via%3Dihub

5. Semkovska, M., Landau, S., Dunne, R., Kolshus, E., Kavanagh, A., Jelovac, A., Noone, M., Carton, M., Lambe, S., McHugh, C., & McLoughlin, D. M. (2016). Bitemporal Versus High-Dose Unilateral Twice-Weekly Electroconvulsive Therapy for Depression (EFFECT-Dep): A Pragmatic, Randomized, Non-Inferiority Trial. Am J Psychiatry, 173(4), 408-417.

https://doi.org/10.1176/appi.ajp.2015.15030372

6. Semkovska, M., & McLoughlin, D. M. (2013). Measuring retrograde autobiographical amnesia following electroconvulsive therapy: historical perspective and current issues. J ECT, 29(2), 127-133.
https://doi.org/10.1097/YCT.0b013e318279c2c9

7. Sobin, C., Sackeim, H. A., Prudic, J., Devanand, D. P., Moody, B. J., & McElhiney, M. C. (1995). Predictors of retrograde amnesia following ECT. Am J Psychiatry, 152(7), 995-1001.

https://doi.org/10.1176/ajp.152.7.995

8. van Diermen, L., Poljac, E., Van der Mast, R., Plasmans, K., Van den Ameele, S., Heijnen, W., Birkenhager, T., Schrijvers, D., & Kamperman, A. (2020). Toward Targeted ECT: The Interdependence of Predictors of Treatment Response in Depression Further Explained. J Clin Psychiatry, 82(1).

https://doi.org/10.4088/JCP.20m13287

9. Waite, S., Tor, P. C., Mohan, T., Davidson, D., Hussain, S., Dong, V., Loo, C. K., & Martin, D. M. (2022). The utility of the Sydney Melancholia Prototype Index (SMPI) for predicting response to electroconvulsive therapy in depression: A CARE Network study. J Psychiatr Res, 155, 180-185.

https://doi.org/10.1016/j.jpsychires.2022.08.011

EXPECTED OUTPUTS AND TASKS

We would expect each student:

- to identify a broad mechanistic hypothesis that they can address in discussion with the clinical team;

- to conduct a literature review (the most relevant papers have already been identified by the clinical supervisors);

- in discussion with the first supervisor identify methods for describing and formally analysing the existing pooled data; where more advanced methods such as mediation modelling are required, and which are only taught in the course later in the year, the first supervisor will provide relevant training.

- write a report detailing their work, including the results of their analyses and a discussion of their findings.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

YES

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

Yes

STUDENT REQUIREMENTS

Letter from principal investigator(s) of ECT studies allowing access to the observational dataset for this research project.

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVCIAE Causal Modelling and Evaluation;7PAVMALM Multilevel and Longitudinal Moddelling;

PROJECT ID 26

PRIMARY SUPERVISOR

Jack Wu

EMAIL ADDRESS

ho_chung.wu@kcl.ac.uk

DEPARTMENT

Biostatistics & Health Informatics

SECONDARY SUPERVISOR

Dan Bean

EMAIL ADDRESS

daniel.bean@kcl.ac.uk

PROJECT TITLE

Automated matching patients to clinical trials with natural language processing

RESEARCH AREA

Natural Language Processing; Information Retrieval;

PROJECT AIMS

The aim of the project is to develop an effective algorithm for matching patient summaries to clinical trial descriptions. Thereby making the patient recruitment process easier for clinical trials.

PROJECT OUTLINE

Patient recruitment is a challenging task in clinical trials. Failure to meet the patient recruitment timeline and/or failure to recruit the minimum number of patients for clinical trials can hinder medical research. This project aims to use the data available in electronic health records to match patient summaries with clinical trial descriptions and build a patient matching system to help identify eligible patients.

In this project, a retrieval system is developed by analysing the text in patient summaries and clinical trial descriptions using natural language processing techniques. Inclusion and exclusion criteria from clinical trial descriptions are extracted and evaluated against the patient summaries.

Techniques in processing text and building retrieval models including tokenisation, named entity recognition, indexing of documents, scoring formula implementation, etc., are required.

This project can benefit researchers in medical research by helping them to automate the patient recruitment process by identifying eligible patients.

The datasets including the documents and queries are publicly available at https://www.trec-cds.org/2021.html.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

The student can develop natural language processing skills (including tokenization, data cleaning and named entity recognition) in analysing free text. The student can also gain experience in building information retrieval models to rank the documents effectively according to queries.

STARTER REFERENCES

Hersh, W. R. (2007). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Clin Pharmacol Ther, 81, 126-128.

Koopman, B., & Zuccon, G. (2016, July). A test collection for matching patients to clinical trials. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 669-672).

Tseo, Y., Salkola, M. I., Mohamed, A., Kumar, A., & Abnousi, F. (2020). Information extraction of clinical trial eligibility criteria. arXiv preprint arXiv:2006.07296.

EXPECTED OUTPUTS AND TASKS

To process text data including tokenization, stop-words removal, stemming, named entity recognition and linking.

To identify and extract inclusion / exclusion criteria in clinical trial descriptions.

To develop and evaluate retrieval models for ranking documents according to their degree of relevance to the queries.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

The dataset is publicly available and free to download.

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

N/A

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVNLPS Natural Language Processing;

PROJECT ID 27

PRIMARY SUPERVISOR

Dr Ioannis Bakolis

EMAIL ADDRESS

ioannis.bakolis@kcl.ac.uk

DEPARTMENT

Biostatistics and health Informatics

SECONDARY SUPERVISOR

Dr Amy Ronaldson

EMAIL ADDRESS

amy.ronaldson@kcl.ac.uk

PROJECT TITLE

AIR POLLUTION AND SEVERITY OF NEUROLOGICAL AND PSYCHIATRIC DISORDERS

RESEARCH AREA

Biostatistics, epidemiology, air pollution

PROJECT AIMS

AIM1: A LITERATURE REVIEW OF THE ASSOCIATIONS BETWEEN AIR POLLUTION AND MENTAL HEALTH

AIM2: CONDUCT EPIDEMIOLOGICAL ANALYSES

AIM3: SYNTHESISE FINDINGS AND DRAFT A LIST OF RECOMMENDATIONS FOR TACKLING AIR POLLUTION LEVELS AND IDENTIFYING VULNERABLE POPULATIONS.

PROJECT OUTLINE

THE WORLD HEALTH ORGANIZATION (WHO) RECENTLY ESTIMATED THAT AMBIENT AIR POLLUTION CAUSES 482,000 PREMATURE DEATHS WITHIN THE WHO EUROPEAN REGION WITH AN ESTIMATED ECONOMIC COST OF 1.575 TRILLION US$ INCLUDING MORBIDITY COSTS. HOWEVER, THE SIGNIFICANT POTENTIAL HEALTH AND SOCIETAL COSTS OF POOR MENTAL HEALTH IN RELATION TO AIR QUALITY ARE NOT REPRESENTED IN THE WHO REPORT DUE TO LIMITED EVIDENCE AND GAPS AND UNCERTAINTIES IN OUR KNOWLEDGE OF THE UNDERLYING PATHOPHYSIOLOGIC MECHANISMS THAT DRIVE THE REPORTED ASSOCIATIONS. BENEFITING FROM COLLABORATIONS WITHIN KING'S COLLEGE LONDON AND IMPERIAL COLLEGE LONDON IN PROVIDING ACCESS TO DATA ON AIR POLLUTION (E.G. CMAQURBAN), AND MENTAL HEALTH (E.G. E-RISK, CLINICAL RECORD INTERACTIVE SEARCH (CRIS)), THIS PROJECT AIMS TO SYSTEMATICALLY EXPLORE ASSOCIATIONS BETWEEN AIR POLLUTION AND NEUROLOGICAL AND PSYCHIATRIC DISORDERS THROUGHOUT THE PROJECT. THE MSC STUDENT WILL USE A RANGE OF STANDARD STATISTICAL TECHNIQUES AND EPIDEMIOLOGICAL DESIGNS (UNDER MAIN SUPERVISION OF IB) TO GAIN A DEEP UNDERSTANDING OF HOW AIR POLLUTION

STRESSORS COULD AFFECT MENTAL HEALTH AND THE POTENTIAL MECHANISMS AND MODERATORS (UNDER MAIN SUPERVISION OF AR).

AIM1: A LITERATURE REVIEW OF THE ASSOCIATIONS BETWEEN AIR POLLUTION AND MENTAL HEALTH

AIM2: CONDUCT EPIDEMIOLOGICAL ANALYSES

AIM3: SYNTHESISE FINDINGS AND DRAFT A LIST OF RECOMMENDATIONS FOR TACKLING AIR POLLUTION LEVELS AND IDENTIFYING VULNERABLE POPULATIONS.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Study design, statistics, epidemiology, policy related work

STARTER REFERENCES

BAKOLIS I, HAMMOUD R, STEWART R, BEEVERS S, DAJNAK D ET AL. MENTAL HEALTH CONSEQUENCES OF URBAN AIR POLLUTION: PROSPECTIVE POPULATION-BASED LONGITUDINAL SURVEY. SOCIAL PSYCHIATRY AND PSYCHIATRIC EPIDEMIOLOGY. 2020 OCT 24. HTTPS://DOI.ORG/10.1007/S00127-020-01966-X

NEWBURY, JB, ARSENEAULT, L, BEEVERS, S, KITWIROON, N, ROBERTS, S, PARIANTE, CM, KELLY, FJ & FISHER, HL 2019, 'ASSOCIATION OF AIR POLLUTION EXPOSURE WITH PSYCHOTIC EXPERIENCES DURING ADOLESCENCE', JAMA PSYCHIATRY, VOL. 76, NO. 6, PP. 614-623. HTTPS://DOI.ORG/10.1001/JAMAPSYCHIATRY.2019.0056

EXPECTED OUTPUTS AND TASKS

Disseration, paper publication subject to project's quality

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

CRIS research passport

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALM Multilevel and Longitudinal Moddelling;7PAVCIAE Causal Modelling and Evaluation;

PROJECT ID 28

PRIMARY SUPERVISOR

Zina Ibrahim

EMAIL ADDRESS

zina.ibrahim@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Inforamtics

SECONDARY SUPERVISOR

None

EMAIL ADDRESS

None

PROJECT TITLE

Neurosymbolic Approaches to Predicting Clinical Deterioration

RESEARCH AREA

Artificial Intelligence, Reasoning, Learning

PROJECT AIMS

To improve the robustness, explainability and domain acceptance of computational clinical deterioration warning systems by using hybrid approaches that combine learning and reasoning about domain knowledge (in the form of clinical guidelines).

PROJECT OUTLINE

Description: Clinical early warning scores (EWS) are used by hospital teams to estimate a patient's risk of deterioration from vital signs. EWS trigger warnings using preset thresholds to initiate more intensive care, but suffer from low sensitivity. Deep Learning-based EWS (DEWS) aims to improve the predictive power of traditional EWS. However, clinical uptake of DEWS is absent; the purely data-driven black-box DEWS are disconnected from well-established clinical guidelines. They are also data-intensive and lack robustness in settings where monitoring is less frequent, e.g. general hospital wards. This project will explore a number of variants of computational EWS that aim to overcome the current limitations of pure DEWS in the aim of achieving explainable and robust models that are aligned with the requirements of clinical practice. Available avenues for research:

1. Improve the predictive power of DEWS through Transformer-based architectures and explore the resulting explainability mechanisms and their alignment with clinical requirements.

2. Explore issues hindering the robust generalization of DEWS, including data quality, imputation and skewed distributions.

3. Explore the possibility of incorporating domain knowledge (through clinical guidelines) into the input of a graph neural network, to steer the training and learning of a DEWS under settings where the patient population is diverse and data is scares.

4. Explore the modes of interaction between a neural network and a symbolic model and their role in improving the alignment between the interpretations of the model outcomes generated and established clinical guidelines and causal pathways.

The models will use real hospital data and the student will choose from a number of outcomes including mortality and admission to intensive care. The students will get the chance to work with the latest developments in neurosymbolic AI, via graph neural networks [1] as well as the most recent developments in deep learning, via the Transformer architecture [2]. The project builds on existing work by the supervisor's team in building clinical outcome prediction models [3] and the utilization of industrial standard executable clinical guidelines through OpenClinical [4]. The project requires students to have background in Computer Science.

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The Graph Neural Network Model," in IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. Kaiser, L., Polosukhin, I. Attention is All you Need. NIPS, 2017.

[3] Z. M. Ibrahim et al., "A Knowledge Distillation Ensemble Framework for Predicting

Short- and Long-Term Hospitalization Outcomes From Electronic Health Records Data," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 1, pp. 423-435, Jan. 2022, doi: 10.1109/JBHI.2021.3089287.

[4] https://www.openclinical.net/

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

The student will come out with excellent knowledge in optimising deep learning models and will develop hand-on in identifying and overcoming the limitation of purely learning-based systems in environments where common sense knowledge drives the decision making process. Please note that excellent programming skills are required in this model, in addition to knowledge of algorithmic methodologies (e.g. case base reasoning, object-oriented programming, functional programming, etc...)

STARTER REFERENCES

Please see project description.

EXPECTED OUTPUTS AND TASKS

1. Output 1: Transformer architecture to predict clinical deterioration - Objective: Improve the predictive power of DEWS through Transformer-based architectures and explore the resulting explainability mechanisms and their alignment with clinical requirements.

2. Output 2: Improved imputation methods for clinical data - Objective: Explore issues hindering the robust generalization of DEWS, including data quality, imputation and skewed distributions.

3. Output 3: A graph neural network incorporating clinical guidelines for a given condition (e.g. sepsis) - Objective: Explore the possibility of incorporating domain knowledge (through clinical guidelines) into the input of a graph neural network, to steer the training and learning of a DEWS under settings where the patient population is diverse and data is scares.

4. Output 4: A reconciliation platform between a DL model and clinical guidelines to establish the robustness of explanations generations by a model - Objective - Explore the modes of interaction between a neural network and a symbolic model and their role in improving the alignment between the interpretations of the model outcomes generated and established clinical guidelines and causal pathways.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

The project will use the MIMIC-III database, which is freely available upon obtaining a security certificate (supervisor will provide guidance).

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

None

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

2

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 29

PRIMARY SUPERVISOR

Mohammad Mahdi Karimi

EMAIL ADDRESS

mohammad.karimi@kcl.ac.uk

DEPARTMENT

Comprehensive Cancer Centre

SECONDARY SUPERVISOR

Giorgio Napolitani

EMAIL ADDRESS

giorgio.napolitani@kcl.ac.uk

PROJECT TITLE

A multi-omics bioinformatics pipeline for identification of non-coding tumour-specific antigens

RESEARCH AREA

Genomics, Immunology

PROJECT AIMS

Our main objective is to optimise and apply the current proteogenomic analysis workflows to detect noncoding regions contributing to the Tumour-specific antigens (TSAs) landscape. Our specific aims are:

(1) Creating personalized protein database derived from tumour RNA sequencing (RNA-seq) data

(2) Searching peptides on the personalized protein database

PROJECT OUTLINE

Tumour-specific antigens (TSAs) are proteins or other molecules found only on cancer but not normal cells. TSAs can assist the body in mounting an immune response against cancer cells and are ideal targets for cancer immunotherapy, but only a few have been discovered so far. Mutated TSAs (mTSAs), also known as neoantigens, have received a lot of attention in the search for vaccines against solid tumours [1]. This is due to the superior immunogenicity of mTSAs attributed to their selective expression on cancers, minimizing the risk of immune tolerance. However, mTSAs are generally patient-specific and are less common than previously thought. Aberrantly expressed TSAs (aeTSAs) are another class of TSAs derived from a variety of genetic and epigenetic changes leading to the transcription and translation of genomic sequences normally not expressed in normal cells, such as non-coding genomic regions.

While mTSAs are patient-specific, aeTSAs can be shared by multiple patients with the same type of tumour and are preferable for vaccine development. Systematic detection of aeTSAs can only be achievable by high-throughput mass spectrometry (MS) analysis of major histocompatibility complex class I (MHC I)–associated peptides. To detect aeTSAs in each patient, we need to use MS analysis software searching MHC I–associated peptides in a personalized protein database derived from tumour RNA sequencing (RNA-seq) data containing the expression of both coding and non-coding regions in each patient [2]. Using public datasets of matched MS and RNA-seq data for tumours, we will optimise and apply the current proteogenomic analysis workflows to create personalised protein databases from RNA-seq libraries and search MHC I–associated peptides on them.

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

- Getting to know the data produced by the state-of-the-art biological experiments such as MS/MS and RNA-seq
- Applying proteogenomic bioinformatics tools such as DE-kupl, PEAKS, etc.
- Learning interdisciplinary skills

STARTER REFERENCES

1. Efremova M, Finotello F, Rieder D, Trajanoski Z. Neoantigens Generated by Individual Mutations and Their Role in Cancer Immunity and Immunotherapy. Front Immunol. 2017 Nov 28;8:1679.

2. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure JP, Gendron P, Courcelles M, Hardy MP, Côté C, Durette C, St-Pierre C, Benhammadi M, Lanoix J, Vobecky S, Haddad E, Lemieux S, Thibault P, Perreault C. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018 Dec 5;10(470):eaau5516.

EXPECTED OUTPUTS AND TASKS

Analysis pipelines to create personalised protein databases from RNA-seq and search MHC I–associated peptides on them.

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

We will use public datasets for initial analysis and optimisation. Then, the analysis methods might be applied to our in-house data.

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

Access to the Rosalind HPC facility at KCL.

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVITHI Introduction to Health Informatics;7PAVMALE Machine Learning for Health and Bioinformatics;7PAVAIHA Artificial Intelligence for Health Analytics;7BBG2014 Bioinformatics, interpretation and data quality assurance in genome analysis;7BBG2016 Advanced Bioinformatics: practical bioinformatics data skills;

PROJECT ID 30

PRIMARY SUPERVISOR

Shaoxiong Sun

EMAIL ADDRESS

shaoxiong.sun@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Richard Dobson and Amos Folarin

EMAIL ADDRESS

richard.j.dobson@kcl.ac.uk; amos.folarin@kcl.ac.uk

PROJECT TITLE

A machine/deep learning approach for blood pressure monitoring technology based wearable PPG

RESEARCH AREA

Machine learning; Deep learning; Signal Processing

PROJECT AIMS

To design features and machine/deep learning models for blood pressure monitoring

PROJECT OUTLINE

Hypertension, or high blood pressure, is one of the most significant risk factors for heart disease and stroke with an estimated prevalence of 26.2% in UK adults over the age of 16. Despite the availability of numerous anti-hypertensive medications that are effective for the majority of patients, rates of uncontrolled hypertension remain high across the globe with a prevalence of 82.9% in an international cohort of hypertensive adults over 50. This is due in large part to challenges in how blood pressure is measured. In practice, blood pressure is predominantly monitored using a brachial cuff in the daily life setting or using an invasive catheter in the acute care setting. While the cuff-based method cannot provide continuous monitoring, the invasive method exposes patients to risks of infection and other complications. There is a critical need for a continuous, cuffless, and unobtrusive way of monitoring and managing blood pressure. In this study, we aim to utilize machine/deep learning techniques to estimate blood pressure or its variations from photoplethysmography signals, which can be easily acquired from wrist.

PROJECT TYPE

Dry Lab (Secondary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Machine learning

Signal (pre)processing

Statistics

Feature engineering

Wearable computing

STARTER REFERENCES

Mieloszyk R, Twede H, Lester J, Wander J, Basu S, Cohn G, Smith G, Morris D, Gupta S, Tan D, Villar N, Wolf M, Malladi S, Mickelson M, Ryan L, Kim L, Kepple J, Kirchner S, Wampler E, Terada R, Robinson J, Paulsen R, Saponas TS. A Comparison of Wearable Tonometry, Photoplethysmography, and Electrocardiography for Cuffless Measurement of Blood Pressure in an Ambulatory Setting. IEEE J Biomed Health Inform. 2022 Jul;26(7):2864-2875. doi: 10.1109/JBHI.2022.3153259. Epub 2022 Jul 1. PMID: 35201992.

Shaoxiong Sun, Erik Bresch, Jens Muehlsteff, Lars Schmitt, Xi Long, Rick Bezemer, Igor Paulussen, Gerrit J. Noordergraaf, Ronald M. Aarts, Systolic blood pressure estimation using ECG and PPG in patients undergoing surgery, Biomedical Signal Processing and Control, Volume 79, Part 1, 2023, 104040

EXPECTED OUTPUTS AND TASKS

Feature extraction of existing features and potentially proposed features

The application of conventional machine learning and deep learning models

Evaluation of results with different metrics and statistics

Opportunity to publish

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

Yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

Yes

STUDENT REQUIREMENTS

No

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

No

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;

PROJECT ID 31

PRIMARY SUPERVISOR

alfredo iacoangeli

EMAIL ADDRESS

alfredo.iacoangeli@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Daniel Bean

EMAIL ADDRESS

daniel.bean@kcl.ac.uk

PROJECT TITLE

Predicting disease risk genes from knowledge graphs

RESEARCH AREA

Genomics, Machine Learning, knowledge Graphs, Genetics, Neurodegeneration, Dementia, precision medicine

PROJECT AIMS

1) To test the potential of our machine learning method for the prediction of genetic causes of a wide range of diseases

2) To further develop our analysis software

3) To develop a webserver to allow users to use our method interactively and in real time

PROJECT OUTLINE

Genetic association studies require very large sample sizes to uncover disease-linked variants, and as yet the variants that have been identified are unable to account for all (or even most) of the genetic risk of disease. In this project, we are attempting to use known genetic risk factors to predict additional (currently unknown) disease-linked genes. These predicted associations can then be validated using targeted studies, potentially leveraging pre-existing data.

Specifically, we are applying a recent predictive algorithm we developed in the SGDP based on knowledge graphs (Bean et al. 2017). Knowledge graphs represent facts as a network, for example a network representing known disease-gene associations would have genes and diseases as nodes, and they would be connected with an edge if there’s a known association. Our algorithm predicts missing edges in such graphs, which in the case of a disease-gene graph means predicting unknown genetic associations. Importantly, the knowledge graph can combine many different types of information, for example we could add genetic interactions between genes. We have shown that this approach algorithm outperformed standard methods (logistic regression, decision trees and support vectors) at predicting unknown adverse reactions to drugs already on the market. These predictions were validated using electronic records from the Maudsley Hospital. Moreover, we recently applied this method to Amyotrophic Lateral Sclerosis successfully and demonstrated its usability in the study of the genetic causes of Neurodegeneration.

In this project, the student will have the opportunity to work on one of the following aspects of our research: 1) to apply our predictive algorithm for the prediction of novel disease risk genes of other neurodegenerative disorders, e.g. Alzheimer's disease, Schizophrenia and Frontotemporal Dementia; 2) to work on the generalization of the method to allow for a custom selection of the databases to use and phenotypes to predict; 3) to implement an on-line server with a user friendly graphic interface to allow a wide audience to use the method on line without needing an informatics background.

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Machine Learning, programming (Python), human disease genetics and genomics, precision medicine

STARTER REFERENCES

Bean et al. (2017) Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Scientific Reports 7

Bean, D.M.; Al-Chalabi, A.; Dobson, R.J.B.; Iacoangeli, A. A Knowledge-Based Machine Learning Approach to Gene Prioritisation in Amyotrophic Lateral Sclerosis. Genes 2020, 11, 668.

EXPECTED OUTPUTS AND TASKS

1) To test the potential of our machine learning method for the prediction of genetic causes of a wide range of diseases

2) To further develop our analysis software

3) To develop a webserver to allow users to use our method interactively and in real time

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

none

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALE Machine Learning for Health and Bioinformatics;7BBG2014 Bioinformatics, interpretation and data quality assurance in genome analysis;7BBG2016 Advanced Bioinformatics: practical bioinformatics data skills;7PAVAIHA Artificial Intelligence for Health Analytics;

PROJECT ID 32

PRIMARY SUPERVISOR

alfredo iacoangeli

EMAIL ADDRESS

alfredo.iacoangeli@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Daniel Bean

EMAIL ADDRESS

daniel.bean@kcl.ac.uk

PROJECT TITLE

An application of knowledge-based machine learning approach in Drug repurposing

RESEARCH AREA

Machine Learning, Drug repurposing, knowledge graph

PROJECT AIMS

1) To test the potential of our machine learning method for Drug repurposing in a wide range of diseases

2) To further develop our analysis software

3) To develop a webserver to allow users to use our method interactively and in real time

PROJECT OUTLINE

The discovery of therapeutic drugs with novel structures from scratch has the disadvantage of being expensive and time-consuming; whereas the discovery of marketed drugs with potential new uses for treating diseases would be a rapid and effective approach, which is known as Drug repurposing.

Our objective is to exploit the available drug information from public datasets and our current knowledge of drug indications to predict novel drug candidates for diseases.

The student will apply our knowledge-based machine learning (ML) method for this purpose [1-3]. In brief, utilising a set of databases of biological, phenotypic and drug-target information, the tool generates a knowledge graph. ML is then used to identify predictive features. The student will train the model on drug-target interaction from DrugBank, and known drug-indication association from Comparative Toxicogenomics

Database (CTD). Using several sets of known and candidate disease drug targets (genes), the student will generate lists of new candidate drugs. The student will also explore how to integrate this method in a webserver that implements and generalizes our ML method [3].

PROJECT TYPE

Dry Lab (Primary Data)

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Programming (Python), Machine Learning, Drug repurposing, knowledge graph

STARTER REFERENCES

[1] Bean, D.M., Wu, H., Iqbal, E. et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Scientific Reports, 2017, 7(1):16416.

[2] Bean, D.M., Al-Chalabi, A., Dobson, R.J.B., Iacoangeli, A. A Knowledge-Based Machine Learning Approach to Gene Prioritisation in Amyotrophic Lateral Sclerosis. Genes, 2020, 11(6):668.

[3] Hu, J., Lepore R., Dobson, R.J.B., Al-Chalabi, A., Bean, D.M., Iacoangeli, A. DGLinker: flexible knowledge-graph prediction of disease–gene associations, Nucleic Acids Research, 2021, gkab449.

EXPECTED OUTPUTS AND TASKS

1) To test the potential of our machine learning method for Drug repurposing in a wide range of diseases

2) To further develop our analysis software

3) To develop a webserver to allow users to use our method interactively and in real time

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

none

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVMALE Machine Learning for Health and Bioinformatics;7BBG2014 Bioinformatics, interpretation and data quality assurance in genome analysis;

PROJECT ID 33

PRIMARY SUPERVISOR

alfredo iacoangeli

EMAIL ADDRESS

alfredo.iacoangeli@kcl.ac.uk

DEPARTMENT

Biostatistics and Health Informatics

SECONDARY SUPERVISOR

Munishikha Kalia

EMAIL ADDRESS

munishikha.kalia@kcl.ac.uk

PROJECT TITLE

Predicting clinical phenotype of Amyotrophic Lateral Sclerosis using machine leaning and structural information of SOD1 variants

RESEARCH AREA

Amyotrophic Lateral Sclerosis , machine leaning , structural biology, precision medicine, genetics

PROJECT AIMS

1) To test wether molecular dynamics based structural features of SOD1 variants can be used to predict the ALS phenotype

2) To develop a ML based predictor of phenotype severity in ALS based

PROJECT OUTLINE

Amyotrophic Lateral Sclerosis is a fatal neurodegenerative disease, primarily affecting upper and lower motor neurons, that results in progressive weakness and culminates in death from neuromuscular respiratory failure, typically 2–5 years after diagnosis. Superoxide Dismutase 1 (SOD1) was the first gene to be associated with familial ALS. More than 200 SOD1 variants have been reported in ALS patients and new variants are identified in ALS patients on a regular basis. However, neither biological evidence of their role in ALS, nor a characterisation of the molecular mechanisms behind their effect on the phenotype are available for the great majority of them. Moreover, the clinical phenotype of people with SOD1 ALS is highly heterogenic. E.g. some variants are associated with a rapid disease progression and others with slower forms of disease.

The student will engage in an ongoing project which has characterised the structural dynamics of a large set of SOD1 variants. Our results suggest that the differences across variants can explain they heterogenic clinical landscape of ALS. Using a set of machine learning methods and large clinical datasets of ALS patients carrying SOD1 variants, the student will investigate what structural features of SOD1 based on Molecular dynamics (MD) simulations can be exploited to build a predictor of disease severity. The results will be validated using data from clinical trials to evaluate the potential of such a predictor to stratify patients.

Depending on the student interests, there will also be an opportunity gain wet lab experience by validating some of the variants in the lab using Sanger sequencing and PCR.

PROJECT TYPE

Wet Lab

KNOWLEDGE, ATTRIBUTES, SKILLS AND EXPERIENCE

Programming (R or Python), machine learning, precision medicine, Sanger sequencing, PCR

STARTER REFERENCES

https://bmcstructbiol.biomedcentral.com/articles/10.1186/s12900-018-0080-9

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0247841

EXPECTED OUTPUTS AND TASKS

1) To test wether molecular dynamics based structural features of SOD1 variants can be used to predict the ALS phenotype

2) To develop a ML based predictor of phenotype severity in ALS based

RESEARCH SAMPLE SPACE- RE THEIR DATASETS AVAILABLE?

yes

ETHICAL STATUS – IS ETHICAL APPROVAL NEEDED

No

STUDENT REQUIREMENTS

none

IS THE PROJECT SUITABLE FOR A PART-TIME STUDENT?

Yes

NUMBER OF MSC STUDENTS PREFERRED

1

USEFUL OPTIONAL MODULES

7PAVPRMD Prediction Modelling;7PAVMALE Machine Learning for Health and Bioinformatics;7BBG2014 Bioinformatics, interpretation and data quality assurance in genome analysis;7BBG2016 Advanced Bioinformatics: practical bioinformatics data skills;