June 2019

The Alan Turing Institute

Turing-Roche Collaborative Summer Research Programme to progress Data Study Group Findings

Call for Data Analyst Candidates and Community Outreach Call

Contents

Introduction

The Alan Turing Institute is working with Hoffman-La Roche to shape a programme of collaborative research to take place through the summer (July to September 2019), which progresses ideas generated at the Data Study Group of April 2019 http://turing.ac.uk/events/data-study-group-april-2019 and a scoping workshop which took place on 22nd and 23rd May.

This is a precursor to a more expansive programme of research, which may follow through a strategic partnership.

We are seeking data analysts and expert input for the project ***on observational modelling of electronic health records.***

How to apply

Applications must be submitted via the online portal at https://ati.flexigrant.com/. If you have not already done so, all applicants must first register on the system and provide basic details to create a profile. If you have any questions regarding the application form or using the online system, contact the Health Programme team at healthprogramme@turing.ac.uk

Data Analysts

Applications to join the team of data analysts for this summer programme are welcome from across the Turing partner network as well as those who registered interest, or whose supervisors/PIs registered interest, in working on a Turing project with Roche via the recent call that was circulated in May to the Turing community and participants of the Data Study Group challenge .

We are seeking a team of three to four analysts who will be a balance of trainee, junior and senior people. The project lends itself well to PhD stop-the-clock, assistant and associate – level academics.

Please see Appendix A for the role description. If you are available between 1st July and 20th September 2019 to work full time on this project, Turing will buy-out your time from your home institution. If you are a PhD student you will need approval from your university and/or supervisor in order to participate.

Analysts will carry out the practical analysis of data as described, specified by the project management team at Turing. Work includes data curation, cleaning, summarisation, exploration, modelling, reporting, code curation. A stylised dataset description is available below.

The online application form (entitled “Turing-Roche Data Analyst Call”) will ask for your contact details, CV, motivation letter (statement of why you would like to work on the project), examples of previous work, and what your approach to the project would be.

The project will build up during the first four weeks and so there will be two rounds of selection, with applications closing on 20th June (Lead Analyst, team members) or 11th July (team members), with start dates of 1st July and 29th July, respectively.

Expert input

In addition, we are seeking to establish a team of experts from across the data science community who may suggest specific advanced approaches to the challenges, engage in peer review and scientific quality control, or would like to request access to the data to apply cutting-edge methodology. Submissions are welcomed from academics from all UK universities. The submissions will be reviewed on a rolling basis and remain open through the summer until the end of August

A stylised dataset description is available below. The online application form (entitled “Turing-Roche Community Outreach Call”) will allow you to contribute methodological suggestions to the Turing-Roche summer project, for the core analyst team to run. Your contributions will be peer-reviewed, and credited to you. The form alternatively allows you to request access to the data yourself, in order to conduct your own

methodological experiment/s and publish on it, subject to data owner approval and conditions of data sharing contract.

We would particularly like to attract a team of experts who are interested to remain engaged and contribute to a prospective longer-term partnership with Roche, beyond the summer project.

Assessment criteria

Submissions will be assessed based on the following criteria:

Data analysts

- Fit to role description (data analysts)
- Alignment of expertise to required data analysis
- Strength of motivation letter
- Suitability of approach described
- For trainee level roles:
  Potential for positive impact on applicant's training and development
- For senior role:
  Experience in leadership and successful completion of data analytics projects

Shortlisted applicants will additionally be invited to an interview which further assesses the above and also contains a technical assessment component.

Expert input

Actioning of expert input will be prioritized based on the following criteria:

- Clarity of proposed methodological suggestions
- Suitability of proposed methodological suggestions to address main questions
- Logical/empirical correctness of proposed methodological suggestions
- Resource demand (analyst time, algorithm runtime, computational demand, etc) of implementing suggestions

Simplified data description

This is a stylised and simplified description that approximates the actual data with the intention of capturing the main challenges within it. We anticipate that approaches for the simplified picture will also apply to the actual data situation.

The dataset consists of a linked set of observational electronic health record tables for approximately 10,000 US patients with NSCLC (non-small-cell lung cancer) receiving first line therapy.

For all patients, the following is available:

- Demographics – age, biological gender, self-reported race etc

- high-level type of therapy received (chemo- or immunotherapy; treatment-received)
- primary clinical outcome: time of death (survival time), or last follow-up (censoring)

For some patients, the following may be available:

- Clinical data
- Prior health conditions
- Structured records of GP and hospital visits, with time-stamped clinical (ECOG etc) and lab records

Genomics data

- genomic array, consisting of summaries of germline and somatic genome.
- Does *not* contain full base sequences.

A typical patient will have clinical records at multiple time points before and during treatment, and no genomic record. Only a small sub-set of patients (approximately 1,000), will have typically one genomic record, acquired around the start of cancer therapy.

The data is observational, i.e., follows no particular study design: containedness in the data, and availability of certain records, is subject to confounding and selection bias. No intervention or instrumental variable (e.g., randomised intention-to-treat) is recorded, even if treatment was administered within a controlled study.

Main clinical questions

(A) individual treatment recommendations
Primary questions: given records available at time point of therapy decision, as described above, (A.1) predict individual clinical outcome, and (A.2) inform patient-level treatment decision (chemotherapy vs immunotherapy).
The question is to be considered with respect to the primary outcome of patient survival.

(B) added value of genomic data
in the context of (A), doctors have the option to request additional genomic testing before making a therapy decision. What are good ways to inform the decision to request genomic testing, in order to (B.1) improve the accuracy of an individualised outcome prediction, or (B.2) improve the clinical outcome?

Contractual matters

A Bilateral Agreement will be signed between Turing and Roche for this programme of research. The Turing will then sign an agreement with the analyst's employing institution, which reflects buy-out of their time, reimbursement of expenses and flow-down of data management and intellectual property obligations.

Intellectual property policy

Turing will contract on our standard collaborative research terms, which are in-line with those agreed for the Data Study Group project which the summer programme of research is built on. These include: each party owns the background that they bring into the project; confidential information will be subject to protective measures; any IP arising will be owned by the Turing in order that it may be published in line with academic processes; any code developed would be published under permissive open source licensing.

Publication policy

We encourage researchers to submit their findings to a high-quality, peer-reviewed journal or conference, on an open-access basis (funding for open-access fees will be available on a case-by-case basis). We expect a 'green' open access version of any papers to be published (if allowed by journal/conference - please check http://www.sherpa.ac.uk/romeo/index.php) either as a pre-print on (e.g.) the ArXiv (https://arxiv.org/) or in an institutional repository.

Publications must not contain Roche confidential information and Roche will have the opportunity to review proposed material in advance of publication.

Staying in touch with this work

If you would like to be kept informed of opportunities, but do not have time to contribute over the summer, please send an e-mail to healthprogramme@turing.ac.uk

Contact

If you have any queries, please contact the Health Programme Team: healthprogramme@turing.ac.uk

APPENDIX A: Data Analyst Call - Role Specification

Role:
To conduct analyses on the data as described, specified by the project management team at Turing.

Work includes data curation, cleaning, summarisation, exploration, modelling, reporting, code curation.

Available levels:
Trainee, junior, senior.
Maps to PhD stop-the-clock (trainee), assistant (junior), associate (senior).

Essential:

- Experience with at least one of the following: R language, data analytics stack in python
- Experience with common data scientific analytics workflows in the context of the dataset (as described)
- Familiarity with survival modelling
- Familiarity with toolboxes for tasks and challenges arising in the context of the dataset (as described)
- Familiarity with sub-versioning and git

Desirable:

- Research experience in data science
- Methodological or algorithmic expertise helpful in the context of the dataset (as described)
- Familiarity with causal or counterfactual models
- Familiarity with machine learning, deep learning, or AI
- Familiarity with observational studies and/or electronic health record data

Applicants for senior analyst role will be expected to provide additional evidence for broad methodological overview, project management experience, references/portfolio of prior successful analytics projects.

For trainee analyst role, all requirements are only desirable (not essential). However, applicants are normally expected to be research students at one of the Turing partner universities and have to provide a summary of how the project would positively impact their training and career.