Open Call: Predictive Modelling projects in the context of uncertainty and heterogeneity in biomedical data

Background

Launched in June 2021, the [Turing-Roche Strategic Partnership](#) is a 5-year collaboration between The Alan Turing Institute (the UK's National Institute for Data Science and AI) and Roche (global pharmaceutical company).

The partnership's 'North Star' is to **enable the generation of insights to better understand patient and disease heterogeneity and its relevance to clinical outcomes at an unprecedented level of precision in order to improve clinical care.** Publication of methods and algorithms will follow the principles of open science to ensure that they are reproducible and interoperable.

Last year we ran a call and associated workshops on the theme of [Structured Missingness](#), seconding two project teams into Turing for 18-months, which you can read more about [here](#). This year the partnership is exploring the theme of **Predictive Modelling** and is inviting applications for funding.

Research topic

This open call will fund secondments into Turing for teams to undertake their own research proposal, aiming to develop robust predictive models in the face of uncertainty and heterogeneity in biomedical data.

Machine Learning-based predictive models have achieved success in a wide range of real life applications, including a variety of tasks in computational healthcare and biomedicine. This has been fueled by the collection and curation of vast amounts of data, on the quality of which learning models are inherently dependent for their performance. Hence, the power and potential of the models themselves, having been proved time and time again, comes down to their robustness in the face of data uncertainty. Data uncertainty can appear in several forms such as noisy data, biased data, missing data, uncertainty in features, changing data distributions, etc. More information about the types and consideration of uncertain data can be found in Annex 1 at the end of this document.

Three main topics aligned with the North Star of the Turing-Roche partnership and involving predictive modelling development will be open to receive research project proposals. It should be noted that project proposals may encompass, and must specify, at least one theme, but they can cover more than one in the same project:

Stream 1: Generalisation despite data heterogeneity:

Heterogeneous data, which is a hallmark of complex biomedical datasets, presents several of the attributes of data uncertainty. The data itself may be heterogeneous, or it may arise due to patient heterogeneity. An example of inherently heterogeneous data, is cancer histopathology and biopsy data. Cancer, by its very nature, degrades the healthy integrity of cells and organs, and hence the tissue contents themselves become more heterogeneous as they lose their cellular formation. The second type is patient heterogeneity, which arises due to variations in age, sex, ethnicity, economic background, etc. Across certain patient cross-sections, a particular data type might be absent, thus causing structured missingness, in addition to unstructured/randomised missingness

in data. Models should be able to perform robustly across data heterogeneity and should be able to generalise across patient heterogeneity.

<u>Stream 2: Prediction uncertainty in personalised healthcare:</u>

Personalised healthcare utilises predictive models by making predictions of the outcomes of an individual patient under different potential treatments and uses these predictictions to help guide or optimise the treatment decisions for that patient. In order for a physician to use these predictions it is critical to understand the confidence or uncertainty associated with these predictions for the individual patient. This may vary considerably between patients depending upon their individual attributes and their relationship to the distribution of the training dataset. Social determinants of health may also have a causal, or at least associative, effect on risk prediction for patient subgroups. Development of robust methods to quantify uncertainty for individual predictions from a predictive model is the second area of focus.

<u>Stream 3: Algorithmic explainability and fairness for AI in healthcare:</u>

Just because an algorithm yields satisfactory predictive performance with adequate generalisation and specialisation, it does not mean it is fair or unbiased. Data bias and class imbalance can occur due to natural sparsity of available data for some classes, whereas for other classes the data available might be abundant. The models become trained in favour of the larger subset, and do not perform well for the smaller subset. An example is heart disease, which being more common among men, the datasets can have many more male patients than female, giving rise to a similar class imbalance problem by sex. Social determinants of health, like economic background, can also create an unfair bias in predictive models. It is important to ensure algorithmic fairness with respect to such protected attributes like sex, race, etc. It is also important to make the algorithms more interpretable and hence the decision making process more transparent, because this makes the aI systems more "trustworthy". This is especially true since modern machine learning algorithms have high complexity and hence are often opaque to interpretation.

Workshop and collaborative meetings

A workshop will take place at [The Alan Turing Institute](#) on **11th October 2022**.

This **in-person** meeting aims to gather a diverse community of researchers to encourage and facilitate **collaboration**, which is a key pillar of how we expect successful proposal teams to work. During the workshop we'll briefly introduce the partnership and the proposed topics around predictive modelling. We will then have discussion sessions on the topics to scope them out, seek synergies and allow participants to further work on ideas they've brought or develop collaborative project ideas from scratch that they can then submit as funding applications for the open call.

Due to space constraints and ensuring there are productive discussions where everyone can contribute, we will be hosting **35 people** for the workshop. We therefore have a short application process to which you will be able to access through our [FlexiGrant application portal](#). As part of the application you will be asked about your contact details, current research topic and motivation for attending. Deadline for applications will be **26th September 2022**.

More information on the workshop and its application process can be found [here](#).

Please note that **you do not need to attend the workshop to apply for the open call** but be aware collaboration is key throughout the project lifecycle. Each project team seconded under the Predictive Modelling workstream will be expected to communicate with our community through events such as knowledge shares, as well as work with the Partnership Community Manager to engage with one another and to actively collaborate with Roche researchers in this area. Finally, project teams are expected to all communicate and collaborate asynchronously e.g. via Slack discussions.

Apart from the workshops, a series of **open collaborative work sessions** will be held **virtually** on **20th and 28th October 2022** to connect interested researchers outside the workshop and allow for continued collaboration. To attend, please register [here](#) (20th October) and [here](#) (28th October)

Additionally, an **introductory event with a Q&A session** will be held shortly after the open call release, on **16th September 2022.** To attend, register [here](#).

Terms and conditions

Successful applicants and their teams (as applicable) will be engaged **via a secondment** agreement to Turing, which must be signed prior to the project starting, subject to the agreement of your employer and subject to any applicable UK visa requirements.

Although some specific Terms and Conditions may depend on the nature of the project, general terms of secondment will be as per the Turing template secondment agreement, or similar if using a previously agreed template with your university. Those terms will be available to download within the application portal (secondment template).

Apart from those, please note that:

- This funding is flowed down from Turing's strategic partnership with Roche, which forms an important element of total funding for Turing's work on Treatment Heterogeneity.

- The funding will be contracted via **a secondment agreement** aligned to the **IP terms below**.

- Within this partnership, both Turing and Roche are committed to **sharing** algorithms and methods publicly.

- Foreground IP developed through the project **will be jointly owned by Turing and Roche**. The intention is that all IP will be published on an open-source basis under a creative commons licence.

- When project teams are ready to publish results (which may include, but is not limited to, journal articles, code, conference presentations) there will be an approval procedure to follow. See "approvals of publications procedure" in Annex 2 at the end of this document.

Eligibility

To be eligible to apply you must:

● Be part of a **university** or **research** institute. Commercial organisations are not eligible.

● Have **permission from your organisation to apply and for project teams to be seconded into Turing, if successful**. For this, **you will be required to submit an approval letter from your research/finance office** stating that your organisation has approved:

> (i) Above-mentioned terms and conditions
> (ii) Proposed costs
> (iii) Employee secondment and agreed FTE

Please use the provided Submission Approval Letter template. This document can be found in the open call web page (submission approval letter template) and will be available to download as well within the application portal. Please **submit one letter per university if the proposal involves multiparty collaborations**.

Funding

Secondments will be available through two different routes:

A) Main projects (approx. 12-18 months): Designed for multiparty proposals boosting collaboration between Principal Investigators with senior expertise. We expect to fund **2 projects up to £300,000 each** (including 100% overheads).

B) Pilot projects (approx. 6 months): Designed for early career researchers seeking a first leadership experience while pursuing proof-of-concept research. We expect to fund **2 projects up to £75,000 each** (including 100% overheads).

In the application, you must specify which route you are pursuing and provide (i) a complete cost breakdown using the provided costing template and (ii) the overall expected timeline for your project using the provided timeline template. Both documents will be available within the application portal to download.

Please, note that you will be able to apply for both routes and with as many projects as you like, but each researcher will be funded for one project maximum.

Other considerations in terms of funding:

● All salary and on-costs will be paid to the university.

● Overheads should be calculated at the Turing rate of £65,000 per FTE per year, 50% of which will be payable to the university and 50% of which will be retained by Turing. Proposed budget **should include 100% of the overheads**.

● 20% VAT is applicable to these secondments.

- Eligible costs include:

  - Salary of personnel working directly on the project – this could include, for example, PIs, postdoctoral research associates, research assistants, data managers, data scientists or software engineers

  - Travel and subsistence for project researchers (e.g., attending conferences, travelling to/from the Turing/other collaborators)

  - Conference or event attendance fees (where conference/event is directly applicable to the research project)

  - Cloud computing or other high performance computing costs

  - Other costs which are specifically justified for the project e.g., books, meeting room or catering costs, specific laptops (only for researchers 100% FTE seconded to Turing or if specific Turing data is required)

  - Open access publications

Please note, the budget requested may not be funded to the maximum requested amount. Reviewers/leads may work with applicants to build collaborative projects which may involve updating costs for projects.

As secondees, researchers can request to receive access passes to use the Turing office space and meeting rooms.

Please note that due to the head terms attached to this partnership, the overhead flow associated with these secondments is an **exception to the usual Turing overhead model for secondments (Turing retains 100% overheads), for this call only.** As stated above in the eligibility section, the funding model described above and specific costs breakdown will require approval from the Principal Investigator's and collaborators' Head of Department and administration teams – **please ensure costs approval is obtained as early as possible.**

How to apply

Submissions for applications will open on **5th September** and will remain open until the submission deadline on **14th November**.

- Applications must be submitted via the [FlexiGrant application portal](). If you have not already done so, all applicants must first register on the system and provide basic details to create a profile.

- Please read the instructions provided and answer every question. When templates are required, please download them, fill them and upload them completed back into the system.

- See Annex 3 at the end of this document for the full application form that will be available on FlexiGrant.

- If you have any questions regarding the application form or using the online system, please contact the Health Programme inbox healthprogramme@turing.ac.uk.

If you are employed by one of the Institute's 13 university partners, please contact your University Liaison Manager – list available here – to make them aware of your application. They can provide support, answer questions and involve you as part of the Turing community at your university from now on.

If you are employed at a university that received a Turing Network Development Award, please contact your Award lead – list available here (scroll to the bottom of the page) – to make them aware of your application.

Review process

Applications will be reviewed by senior members of the partnership from both Turing and Roche. The step-by-step review process will be as follows:

- Basic eligibility check by Research Project Manager

- All eligible applications undergo individual review and scoring by review panel members

- Panel meets to discuss all applications and individual scores. Final outcomes will be decided to be successful, unsuccessful or "revise and resubmit".

  - Note that outcomes of "revise and resubmit" will be provided feedback and the option of discussion meeting with a designated project member in order to revise the proposal and resubmit on FlexiGrant. After resubmission, the panel will reconvene and make final decisions on resubmitted applications.

- Final decisions communicated to all applicants

Selection Criteria

The reviewing panel will evaluate applications considering the following elements:

- Scientific Novelty & Timeliness

  - Does the project contain novel elements, including new method(s) or new observation(s), that lead to new knowledge discovery?

  - Does the overall proposal contribute to scientific progress and how?

  - How is the project addressing the contemporary challenges faced in data science for healthcare, in particular around predictive modelling in the context of uncertainty and heterogeneity in biomedical data?

- Feasibility

    - How can the proposed idea be implemented?

    - Are there any key enablers or blockers to project execution?

    - Is there a well defined and detailed project timeline specifying key project milestones and/or deliverables?

    - Have the applicant(s) clearly described the attainability of the project?

- Level of contribution towards the Partnership's "North Star" of understanding patient and disease heterogeneity

    - Does the project idea contribute towards achieving the Partnership's North Star? For example, how will the idea change/attain/expedite/significantly contribute to the pathway to understanding patient and disease heterogeneity?

- Utilising the Partnership & Benefit of your collaborative team

    - How does the project benefit from both Turing's position as a National Institute and Roche?

    - Why could these specific benefits not have been addressed within a single academic institution? (specifically for main projects)

    - What is the added value of the combined expertise of the co-investigators? (specifically for main projects)

- Value for money

    - Are resources requested appropriate given the scope of the project? Are they clearly defined and justified?

The Panel will be chaired by Chris Holmes, Programme Director for Health and Medical Sciences at The AlanTuring Institute.

After the submission deadline (**14 November 2022**), eligibility checks will be undertaken and completed by **18 November 2022**. The reviewing panel members will complete individual proposal reviews by **5 December 2022**. The reviewing panel members will convene **19 December 2022** to make final decisions as to the outcome of the submitted proposals. Final decisions will be communicated, with feedback, via email before **23rd December 2022**.

The offer acceptance deadline will be end of **1 February 2023**, with secondments and the rest of the documentation to be finalised before **1 March 2023**. Projects are anticipated to start by **March 2023** with pilot projects to be completed and reported by **August 2023** and main projects to be completed and reported by **August 2024**.

Summary of key dates

Publication of call for proposals	5th September 2022
Workshop registration opening	5th September 2022
Introduction and Q&A Session	16th September 2022
Workshop registration closing	26th September 2022
Notification for successful applicants (workshop)	27th September 2022
Workshop dates	11th October 2022
1st virtual coworking session	20th October 2022
2nd virtual coworking session	28th October 2022
Deadline for proposals	14th November 2022
Final decision outcome communication	By 23rd December 2022
Offer Acceptance Deadline	1st February 2023
Documentation ready by	1st March 2023
Projects start	1st March 2023
Projects to complete and report by	August 2023 (Pilot) August 2024 (Main)

Offer Acceptance Process

Successful applicants will be notified via email, which will include:

- The offered award for the proposed project

- A draft of the secondment agreement for the organisation to review

- The conditions and instructions for accepting the offer, including any relevant forms to complete.

In order to benefit from the offer, successful applicants will have a period of **forty (40)** days to follow instructions by doing the following:

A. Respond to the email on whether you accept the offer and ensure your Research Office or equivalent was in agreement and included in your answer email

B. Arrange relevant signatory for secondment agreement (either Turing template or alternatively employing organisation template, to be agreed with Turing after your organisation reviews

initial Turing secondment draft). We will need you to send us the relevant contact details within your organisation for us to discuss this.

C. Disclose if you will bring to the project any Background Intellectual Property, Third-Party Intellectual Property or relevant Licence Grants.

D. Disclose if you have in place any Third Party involvement, contributions and/or agreements around the proposed project (please note this is in addition to all named individuals in the application).

E. Complete the Ethical Approval Form (which will be included in the notification email) to be reviewed internally by The Alan Turing Institute. You will be asked for your project description (goal of the project, data sources, data use, data dissemination, research methods), data consent issues, privacy and security issues and other potential harms resulting from the research. This internal approval is aimed to confirm the ethical compliance of your research objective and datasets.

F. Sign the Project Start Confirmation Form (which will be shared with you once all relevant documentation is in place) to confirm the project will start on the date planned.

The offer will lapse if these requirements are not completed by **1 March 2023**. The funding will subsequently be reallocated to other projects in due course. Please note if this completion date is not met purely due to any delay in completing part E) on the Turing side, then an extension will be given.

Post Award information

Awarded projects and therefore the seconded project teams will form the Predictive Modelling workstream of The Turing-Roche partnership. This workstream will operate in parallel with the already established Structure Missingness workstream and opportunities for project teams to connect will be promoted and encouraged.

Collaboration and engagement are critical to the successful delivery of the entire workstream. It is therefore expected that secondees will be active members of the wider workstream, the partnership, and The Turing more broadly. They will be given opportunities to communicate technical topics, both orally and in writing, to colleagues and external partners, for example, by preparing and presenting reports, blog posts, organising and delivering presentations, and taking an active role in regular meetings and discussions.

Each project will be expected to complete short project updates each quarter, in addition, a comprehensive final report outlining the project aims, achievements, research is expected.

Queries

Should you have any queries, please contact healthprogramme@turing.ac.uk. We aim to answer in 3 working days.

Our Values

The Alan Turing Institute is committed to equality diversity and inclusion and to eliminating discrimination. All employees and secondees are expected to embrace, follow and promote our EDI Principles and Our Values.

Annex section

Consideration on types of uncertain data

Changing Distribution: In health datasets, particularly those derived from ongoing large clinical trials, data comes in batches depending on patient cohorts. When such data is collated across several institutions over time, the nature or underlying distribution of the data can change, giving rise to concept drift and persistent outliers. If machine learning models can not cope with changing data, that might lead to falling performance at the minimum, which might escalate through random prediction to catastrophic forgetting at the worst.

Noisy Data: For large multi-institutional health research projects, when data is collated from multiple sources, then there are variations and noise patterns introduced, even when overall experimental conditions remain consistent. One specific example of this is subtle variations of staining patterns for digital histopathology images, when staining is done at different locations or time, even when the same staining type/compound is used. Such types of variations and noise are almost impossible to avoid, and hence the performance of machine learning models must be robust to these.

Biassed Data: Data bias and class imbalance can occur due to natural sparsity of available data for some classes, whereas for other classes the data available might be abundant. One such example is skin cancer, the probability of occurrence of which is higher in paler skin than darker skin, and hence existing datasets have a large number of samples of the former set, and less number from the latter set. Though this represents the patient population faithfully from a statistical point of view, but from a machine learning point of view this causes a biassed learning issue. The models become trained in favour of the larger subset, and do not perform well for the smaller subset, though of course, skin cancer does happen for darker skin types as well.

Uncertain Features: Another source of data uncertainty is when the variables or features being used in the predictive model are measured with a high degree of noise or uncertainty. One particular example of this may be when the input feature is itself the result of a prior predictive model and so comes with a prediction confidence interval of its own.

Missing Data: In health datasets, certain fields of data might be missing for individual patients or even certain categories of patients, and the missing data can build up to a substantial portion of a dataset, so much so that it can sometimes be larger than the available data in particular cross-sections or chunks. Even when the main data type is available, other related but important data might be missing, like metadata, labels and annotations. In other cases, getting good quality labels and annotations for very large datasets from human experts/clinicians might be cost prohibitive, especially for biomedical applications.

Annex 2: Approval of publications procedure:

i) A proposed publication shall be submitted to both Turing and Roche prior to submission in order to be able to make comments on the proposed publications

ii) Comments will be provided within 30 days

iii) All reasonable comments will be incorporated (unless the changes will adversely compromise the science)

vi) No confidential or non-public IP will be contained in the proposed publication, without prior written consent

v) The Reviewing Party can request a 90-day delay to submission to publish, which may include reasonable additional delay requests to prevent IP or confidential information being compromised or lost

vi) The Partnership is committed to open science practices and we ask that you publish your paper either in an open access journal, or make the final version of it openly available e.g through a postprint (green open access)

Page 1: Applicant details

Lead applicant details

Name and surname

Organisation

Role

Email Address

Phone Number

CV upload (max 2xA4 pages)

Co-applicant(s) details

Name and surname

Organisation

Role

Email Address

Phone Number

CV upload (max 2xA4 pages)

Page 2: Project Details

Please, state whether you are applying for main or pilot projects. Remember that main projects will be able to get a maximal funding of £300.000 (including 100% overheads) for 18 months whereas pilot projects will be able to get a maximal funding of £75.000 (including 100% overheads) for 6 months. Researchers will be able to apply for both routes if interested in doing so but only one project per researcher will be funded. Pilot projects may be able to seek further funding if evaluation of the final report considers additional research is of interest.

Project title (50 words)

Project Abstract (250 words)

Outline of research question and scientific approach (800 words)

Impact of your project (200 words)

Relevance of your project to the partnership's 'North Star' (200 words)

Describe any collaborative practices for co-creation in design and/or delivery with other funded projects in the Predictive Modelling workstream (200 words)

Project Outputs and Deliverables (300 words)

Risks to Project Delivery and Mitigations (300 words)

Data readiness (source) (200 words): Does the research team have, or have a plan for, access to any datasets and/or facilities essential to the quick commencement and ultimate success of the project? Which ones? Does the team have an alternative data source, open data set, or synthetic data set, which could be used for the research if the originally planned data sources are not accessible? Considering the theme of robust predictive modelling, we encourage you to consider multiple data sources for developing and validating models, including open access data sources.

Data readiness (legal and ethical approvals) (200 words): For the data source(s) you are planning to use, what legal or ethical approval (if relevant) is required to allow the team access? Do you have such approvals? Please note that should your application be successful you may need to provide a copy of relevant ethical documentation if applicable, e.g. Institutional Review Board approval. Data sources, use and dissemination as well as data consent, privacy or security issues will also need to be disclosed.

Proposed timeline (upload): Please, provide a timeline for the consecution of the proposed project objectives. Remember that main projects are expected to run for 12 - 18 months whereas pilot projects will have an approximate duration of 6 months. Please use the template provided.

Proposed budget (upload): Please provide a breakdown of costs for your application. Remember that main projects can be funded up to £300,000 and pilot projects can be funded up to £75,000. When proposing your budget, note that it must include 100% of overheads. Please use the template provided as the Turing overhead model is already hardcoded into the template along with instructions

Justifications of costs (300 words)

Team's roles and expertise (200 words)

References (free text, do not contribute to word count)

Page 3: Eligibility

Please confirm your eligibility below

I confirm that I am employed by a **university** (not limited to the Turing university partner network) or a **research** institute.

I confirm that the appropriate permissions are in place to allow the project to be undertaken under the terms outlined in the call, should this application be successful.

Please upload the Submission Approval Letter from your Finance / Research support office confirming that (i) proposed costs are correct, (ii) proposed FTE are correct and (iii) university will be willing to host and accept the secondment, including the terms outlined in the open call document. Please use the template provided. Please obtain the same for all universities on multiparty applications (Upload document box).

Please declare any conflict of interest here: (Free text box here).