IMPERIAL

MSc in Statistics

Introduction to Statistics Research Projects

May 2024

Examinations
Well done!

- If you have any comments regarding the examinations, please let the admin team and director know.
- A set of examiners will review the marks of the modules to ensure students are considered fairly across all modules.
- Further details on the liaison panel are described in the handbook.
- Provisional elective module marks will be released towards the end of July.
- Re-sits will take place in September after the projects.

Key Dates

- **Library Training** –
  - 23 May, 11:00-12:30, HXLY 340.
  - Plagiarism
- **Stats Clinic** – Projects
- 24 May, 14:00-17:00, HXLY 342.
- 28 May, 10:00-17:00, HXLY 311.
- **Thesis Writing Training**
  - 01 July, 09:00-11:00 HXLY 340.
  - Writing your research report
  - Oral presentations
- **Poster Presentations**
  - 1 July, 11:00-15:00 HXLY 170QG
  - Opportunity to get feedback from independent academics and friends;
  - Summarise your work;
  - Define objectives of your project;
  - not marked;
  - Prizes for the best posters.
  - Submission deadline: June 24, 9am
- **Report Draft** – 12 August. (suggested)

Stats Clinic
Research Project Skills Training

- **Day 1 (3 hours).**
  - Intro to shell (30 min)
  - File and code structures (20min)
  - Version control (40 min)
  - Interactive GitHub setup (20 min)
  - Workshop (60min)

- **Day 2 (all day).**
  - The Conda IDE
  - LaTeX and Overleaf
  - Reproducibility
    - Seeding
    - Logbook
    - Notebook best practices
  - Config files and parameterization
  - Profiling and optimization
  - Workshop

Key Dates

- **Research Report Submission** – 30 August 2024, 1pm. 90% of project mark.
- **Oral Presentations** – 10-13 September 2024. 10% of project mark.
  - Note that these will be in person.
- **Resits** – after thesis submission.
- **Farewell Party and Awards** - ??? September 2024.
- **Partner Visits** – Arrange and book early. Limited funds available, with approval from supervisor and director. Transport Expenses Policy

Statistics Research Projects
Outline

- Opportunity to work on an exciting, state-of-the-art statistical problem.
  - Developing new theory
  - Creating new methodology
  - Solving an open, application-driven problem

- Each project has its own **novel component**, whether that it the analysis of a dataset with a new method, the development of a new method or theory.

Skills

Statistical	Leadership
Computational	Ownership
Presentation	Independence
Writing	Time Management

. . .

Best project 2021-22

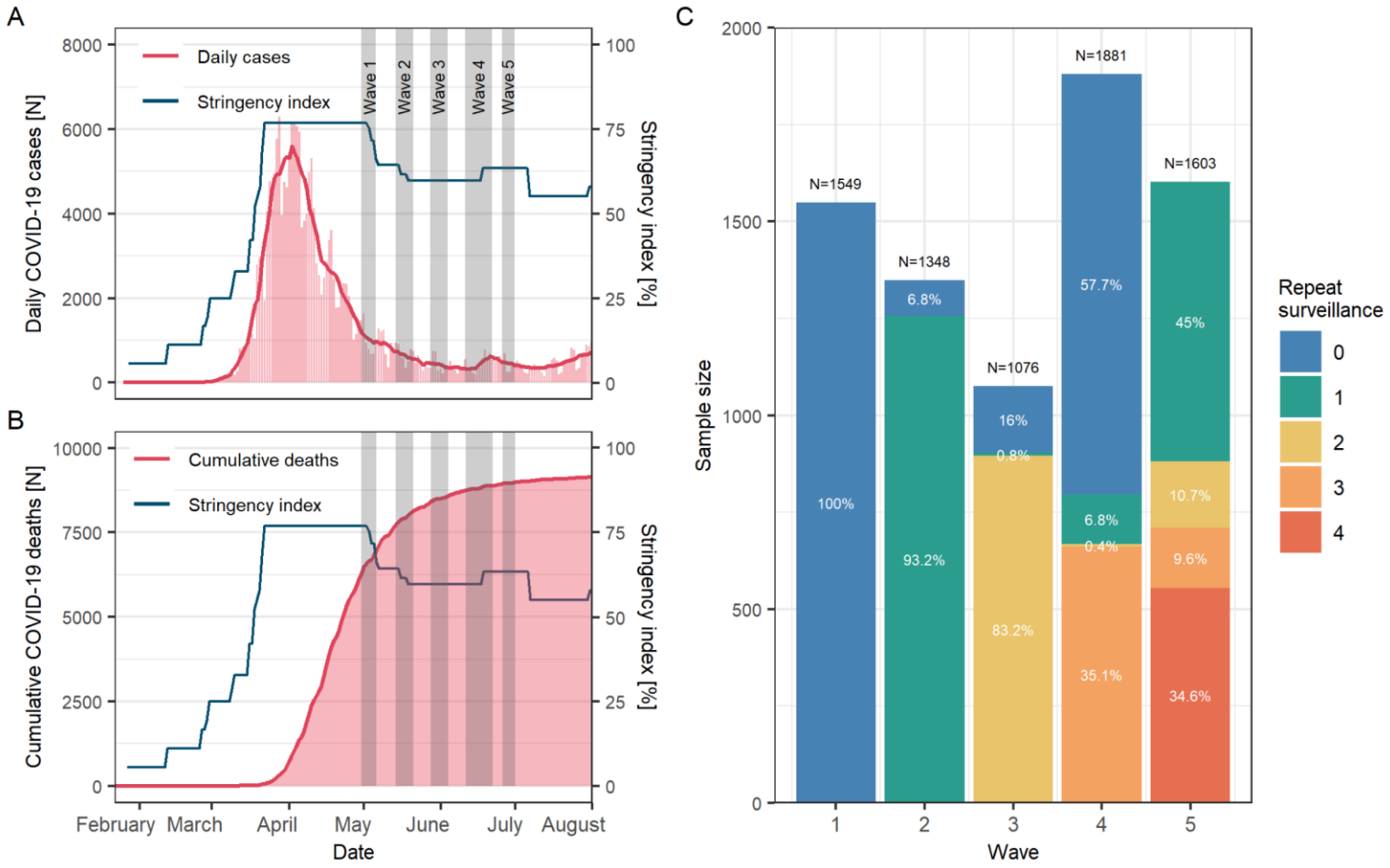
[Submitted on 20 Oct 2022]

Estimating fine age structure and time trends in human contact patterns from coarse contact data: the Bayesian rate consistency model

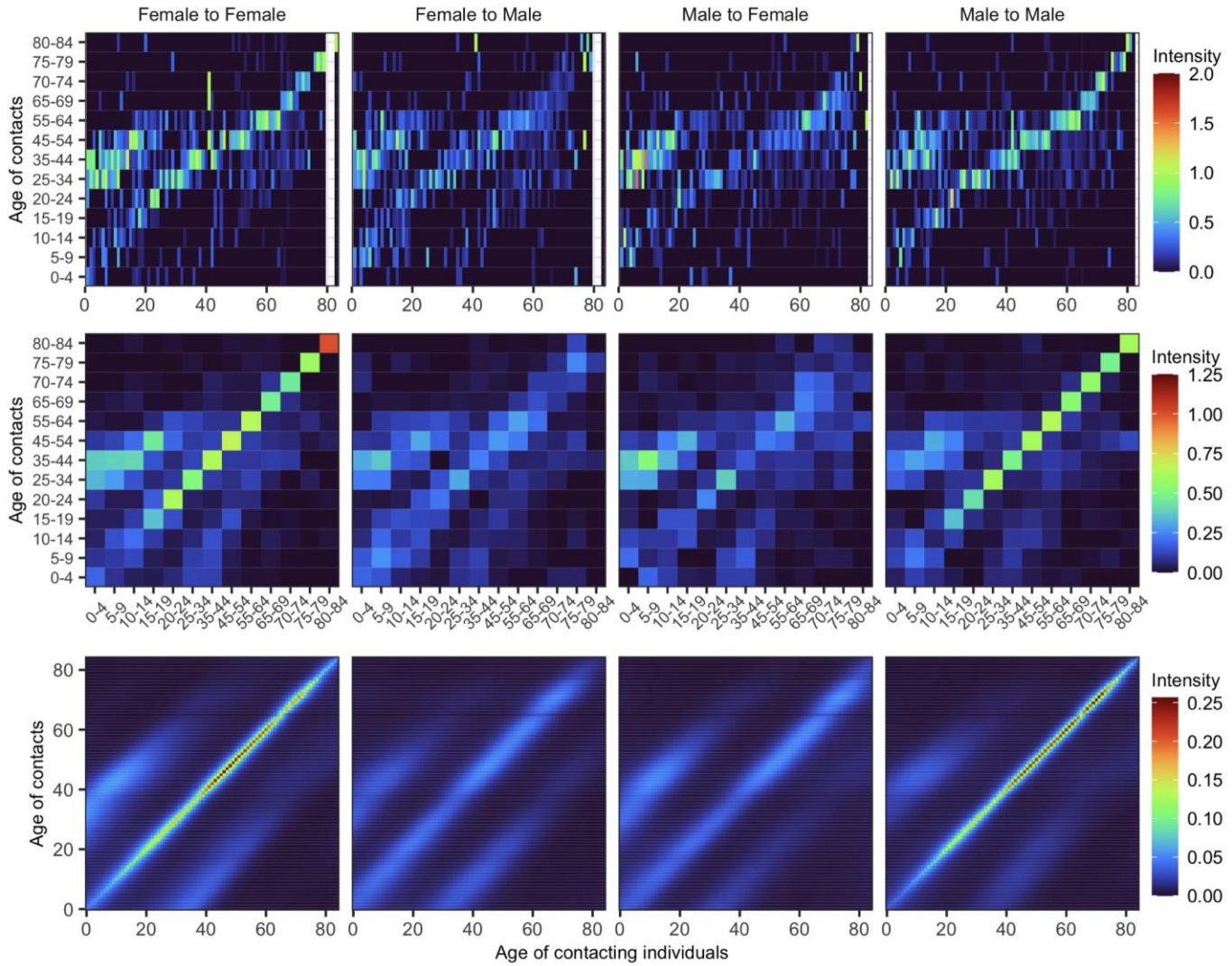
Shozen Dan, Yu Chen, Yining Chen, Melodie Monod, Veronika K. Jaeger, Samir Bhatt, Andre Karch, Oliver Ratmann (on behalf of the Machine Learning & Global Health network)

Since the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), many contact surveys have been conducted to measure changes in human interactions in the face of the pandemic and non-pharmaceutical interventions. These surveys were typically conducted longitudinally, using protocols that differ from those used in the pre-pandemic era. We present a model-based statistical approach that can reconstruct contact patterns at 1-year resolution even when the age of the contacts is reported coarsely by 5 or 10-year age bands. This innovation is rooted in population-level consistency constraints in how contacts between groups must add up, which prompts us to call the approach presented here the Bayesian rate consistency model. The model incorporates computationally efficient Hilbert Space Gaussian process priors to infer the dynamics in age- and gender-structured social contacts and is designed to adjust for reporting fatigue in longitudinal surveys. We demonstrate on simulations the ability to reconstruct contact patterns by gender and 1-year age interval from coarse data with adequate accuracy and within a fully Bayesian framework to quantify uncertainty. We investigate the patterns of social contact data collected in Germany from April to June 2020 across five longitudinal survey waves. We reconstruct the fine age structure in social contacts during the early stages of the pandemic and demonstrate that social contacts rebounded in a structured, non-homogeneous manner. We also show that by July 2020, social contact intensities remained well below pre-pandemic values despite a considerable easing of non-pharmaceutical interventions. This model-based inference approach is open access, computationally tractable enabling full Bayesian uncertainty quantification, and readily applicable to contemporary survey data as long as the exact age of survey participants is reported.

Best project 2021-22



Best project 2021-22



Best Project 2021-22

groups $a \in \mathcal{A}^{trg}$ and population age groups $b \in \mathcal{B}$ to

$$Y^{gh}_{trac} \sim \text{NegBinomial}\left(\sum_{b \in c} \alpha^{gh}_{trab}, \frac{\nu}{1+\nu}\right) \tag{16a}$$

$$\mu^{gh}_{trab} = \alpha^{gh}_{trab}\nu \tag{16b}$$

$$\log \mu^{gh}_{trab} = \log m^{gh}_{tab} + \rho_r + \log(N^g_{tra}) + \log(S^g_{ta}) \tag{16c}$$

$$\log m^{gh}_{tab} = \beta_0 + \tau_t + \boldsymbol{f}^{gh}_t\big(d(a,b)\big) + \log(P^h_b), \quad g = M,\ h = F,\ a, b \in \mathcal{B} \tag{16d}$$

$$\log m^{hg}_{tab} = \beta_0 + \tau_t + \boldsymbol{f}^{gh}_t\big(d(b,a)\big) + \log(P^g_b), \quad g = M,\ h = F,\ a, b \in \mathcal{B} \tag{16e}$$

$$\log m^{gg}_{tab} = \beta_0 + \tau_t + \boldsymbol{f}^{gg}_t\big(d(a,b)\big) + \log(P^g_b), \quad g \in \{M, F\},\ a \leq b, \tag{16f}$$

$$\log m^{gg}_{tab} = \beta_0 + \tau_t + \boldsymbol{f}^{gg}_t\big(d(b,a)\big) + \log(P^g_b), \quad g \in \{M, F\},\ a > b, \tag{16g}$$

and

$$\beta_0 \sim \mathcal{N}(0, 10) \tag{17a}$$

$$\rho_r \sim \mathcal{N}(0, 1) \tag{17b}$$

$$\tau_t \sim \mathcal{N}(0, 1) \tag{17c}$$

$$\nu \sim \text{Exponential}(1) \tag{17d}$$

$$\boldsymbol{f}^{gh}_t\big(d(\boldsymbol{x})\big) | \alpha_{ti}, l_{ti} \sim \text{HSGP}(\boldsymbol{0}, \tilde{\boldsymbol{L}}^{gh,2}_t \otimes \tilde{\boldsymbol{L}}^{gh,1}_t), \quad gh \in \{MF, MM, FF\}, i = 1, 2 \tag{17e}$$

$$\alpha_{ti} \sim \text{Cauchy}^{+}(0, 1), \quad i = 1, 2 \tag{17f}$$

$$l_{ti} \sim \text{InvGamma}(5, 5), \quad i = 1, 2. \tag{17g}$$

Monte Carlo draws from the joint posterior distribution of all parameters were obtained with the probabilistic computing language `Stan` [28] via the `cmdstanr` interface version 0.5.2. Eight chains were run in parallel for 500 warmup iterations and 1000 iterations thereafter. Initial sampling was facilitated by adding the nugget 10^{-13} to α^{gh}_{trab}. We typically observed a small number of divergences in the NUTS algorithm, but these

Student Testimonials

Panayiota Savva (Cyprus)

"Studying my MSc at Imperial College London was an amazing opportunity to get involved in the application of statistical methods to different disciplines, and to work with experts in very exciting projects."

Student Testimonials

Yining Chen (China):

"I enjoyed studying at Imperial College London, especially the high-quality modules and the exciting research projects. The MSc Statistics programme helped me strengthen the statistical theory basis and improve the technical skill effectively."

Student Testimonials

Shozen Dan

"I was very surprised and happy to hear that I had been picked the Best Research Project Prize, especially because my cohort was filled with extremely talented people. I remember working very hard but also having a great time working with Dr. Oliver Ratmann who was a phenomenal mentor. This is huge news to me as I would like to pursue a PhD in statistics after the MSc!"

Supervisors
Setting Expectations

With your supervisor:

Define the research question, hypothesis and aims of the project.

Create a plan for the project.

Have regular group meetings.

Role of the supervisor:

Provide guidance / advice.

Provide feedback on one draft of your research report.

Provide pastoral care.

X Write or copy-edit your report

X Write or correct your code

Supervisors
Setting Expectations

Meetings: expect on average 1 hour group meeting per week, plus an individual meeting giving feedback on your draft report.

On-campus meetings: During the projects you are expected to be around London, your supervisors will arrange for some (or all) of your meeting to be on-campus.

Special permission is needed for being away from campus.

Draft Report Feedback: Your supervisor will provide comments on one draft of your research report. Recommended mid-august, you should agree a date to deliver your draft document and when they will give feedback to you. (more details in July session)

Supervision Meetings
Setting Expectations

Start of a meeting:

Describe what you discussed last week; Give a brief summary of what you have achieved

During a meeting:

Present the work that you have done since the last meeting: Present tables, figures, RMarkdown and/or a short summary; Take notes

End of meeting:

Summarise the objectives for the following meeting; Agree the time of the next meeting

Just after the meeting:

Record the minutes of the meeting; Share the minutes with your supervisor(s)

Supervision Meetings
Setting Expectations

Your supervisors are looking for:

- Ownership
- Solid working
- Initiative
- Creativity
- Honesty
- Open communication

What do we mean: initiative and ownership?

- The EXTRA steps of work that you will do in addition to what they have asked you to do for the following meeting.
- You should avoid emailing them with questions that can be easily found through an internet search.
- Supervisors are not expected or obliged to respond to your emails all the time. Of course, email them if they are questions that will take your research further.

What do we mean: open communication and honesty?

Be open to them. Don’t hide things or lie to supervisors.

Research Report
Qualitative Description- Distinction

For Distinction level, the level of originality must be significant and not just a straightforward extension of existing work (e.g., easy to implement extensions of existing theory, easy to implement simulations, or straightforward real-data examples, should **not** be deemed as sufficient levels of originality for distinction level grades).

When the thesis is close to a degree classification threshold (i.e. 60 or 70), remember the degree classification algorithm on the MSc in Statistics automatically considers students for an uplift when their thesis mark is in the 68.00-69.99 range and their OWA is above 70.00, and so thesis marks should be marked on the criteria above.

Grades under 70 are below the Distinction threshold, and the criterion for this should be the presence or absence of original work or insights that go beyond straightforward extensions of existing theory/analysis, and the significance of the result(s) obtained. If a mark is just above or below 70, the reason for this should be stated clearly.

Research Report

Qualitative Description- Merit / Pass

Grades under 60 are below the Merit threshold, and the criterion for this should be the presence or absence of an attempt to perform original work or obtain significant novel insights. For example, a project which only reviews the problem and potential analysis methods would fall below the Merit boundary. If a mark is just above or below 60, the justification should be stated clearly.

Grades under 50 are below the Pass threshold and correspond to a Failing mark; this should only be for projects where there is strong doubt as to whether the student is in command of the material. If a mark is just above or below 50, the reason for this should be stated clearly.

Research Report

Qualitative Description- Originality of Research and Writing

Original research is desirable wherever feasible. In some areas a suitable project might be to apply, implement or explain a known result. A report which clarified a known cited result, presenting it more clearly than the published source, might deserve a mark in the 75-84 range, and a higher mark would be justified if there were evidence of a new insight, e.g., if the relationship between known results was clearly discussed.

While original work is not needed for a Pass, some original ideas are needed for a Merit, and they should be very well developed to justify a Distinction. All project topics are of such a level that a Distinction mark could be justified for excellent work on them.

A report which was written in the candidates own words, to the standard of a chapter of an undergraduate textbook, would be regarded as excellent presentation. Of course, verbatim quotes from uncited sources are treated as plagiarism, or at best, as poor academic practice. Please contact the course director if you suspect plagiarism in the thesis.

Research Report
Marking Criteria - Distinction

Grade	Description
85-100	The thesis contains excellent original work, with a significantly novel result. The work is explained excellently, with a very good/comprehensive account of and references to existing work. The conclusions are excellently reasoned and communicated. The quality of writing, plots and presentation is excellent, at the level expected in an academic paper or book.
75-84	The thesis contains good original work, with a significantly novel result. The work is explained very well, with a very good/comprehensive account of and references to existing work. The conclusions are well reasoned and communicated. The quality of writing, plots and presentation is good, close to a publishable standard without much addition.
70-74	The thesis contains good original and novel work. The work is explained well, with a good account of and references to existing work. The conclusions are well reasoned and communicated. The quality of writing, plots and presentation is good, close to a publishable standard without much addition.

Research Report
Marking Criteria - Merit

Grade	Description
65-69	The thesis contains several good original ideas that are either generally straightforward to implement and/or not innovating far enough to complete a significant result. The work is still explained well and with good account of and references to existing work. The report is well written and well presented, but not close to a publishable standard without much addition.
60-64	The thesis contains some good original ideas that are generally straightforward to implement and not innovating far enough to complete a significant result. There is fair explanation of the work with some account of and references to existing work. The report is written and presented to a satisfactory standard.

Research Report

Marking Criteria – Pass / Fail

Grade	Description
55-59	The thesis contains an explanation of the problem and the work of others on it, a description of some suitable methods for tackling it, but without much independent work of the candidates own.
50-54	As above but with other shortcomings including incomplete referencing, unclear text or poor presentation.

Grade	Description
40-49	Poor understanding, with many shortcomings including minimal referencing, unclear text, or poor presentation. If significant material is quoted, verbatim, from a cited source, without evidence that the student understands it, no mark higher than 49 should be given.
20-39	Very poor understanding, with many shortcomings including minimal referencing, unclear text, or poor presentation.
0-19	No evidence of understanding, with many shortcomings including scant references, unclear text, and very poor presentation.

MSc in Statistics
Degree Classifications

To qualify for the award of the MSc in Statistics a student must have:

- accumulated credit to the value of no fewer than **90 credits** at level 7
- and no more than 15 credits as a Compensated Pass;

Taught modules = 60 ECTS

Research project = 30 ECTS

Distinction:

The student has achieved an overall weighted average of 70.00% or above across the programme.

The student must normally achieve a distinction (70.00%) mark in the Statistics research project.

Merit:

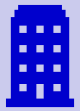
The student has achieved an overall weighted average of above 60.00% but less than 70.00%

The student must normally achieve a merit (60%) mark in the Statistics research project.

Pass:

The student has achieved an overall weighted average of 50.00% but less than 60.00%.

Statistics Research Project
Additional Help



BB page for the MSc projects: Statistics Research Project 2023-2024

Thesis talk: 1st of July (just before the poster presentations)

Library support session: Today!

Maths / Stats computer servers Docs / Stats Clinics / Supervisor

Statistics Research Report
Computing Resources

The department has computing servers/clusters that can be used to run simulations for your projects.

The computing facilities for the Statistics section are described here:
https://sysnews.ma.ic.ac.uk/stats

There is also a dedicated page for MSc projects users:
https://sysnews.ma.ic.ac.uk/stats/MSc_compute_servers.html

More information about the Statistics HPC cluster:
https://sysnews.ma.ic.ac.uk/stats/latest-news_070122.html

Statistics Research Report
Using the Cluster

The website also has instructions for how to use the cluster:
https://sysnews.ma.ic.ac.uk/cluster-job-queueing/quickstart.html
https://sysnews.ma.ic.ac.uk/compute-cluster/

There is also information and tips on accessing and/or using these facilities remotely from outside the College:
https://sysnews.ma.ic.ac.uk/accessing_Maths_systems_from_outside_college.html

We have also written a step-by-step guide to submitting a job that may be helpful, which can be found in the Research Project folder in the MSc in Statistics Blackboard page:
https://bb.imperial.ac.uk/bbcswebdav/pid-2825456-dt-content-rid-18329298_1/xid-18329298_1

Last piece of advice:

Open communication with your supervisors is important

When you take some time off do inform your supervisors;
They should also inform you when they will be unavailable for a meeting;
Do take some time off - You do deserve a break!

Write early and often.

If you haven't already:

Contact your supervisors to start your regular meetings

IMPERIAL

Any questions?

IMPERIAL

Thank you for joining!
Enjoy your projects!