

# Natural language processing of histopathology reports

## Patient Notification

Version 2.1 (16/06/2023)

### What is the study about?

[Histopathology](#) is a medical specialty in which tissue samples removed from patients during surgery or by biopsy, are examined by pathologists to diagnose disease and guide further treatment. Pathologists describe their findings in a [histopathology report](#). Histopathology reports have a wealth of information of medical and scientific value.

This project seeks to bring together large numbers of histopathology reports. These will be analysed using 'Natural Language Processing' (NLP). NLP is a computer method that can accurately extract specific information from text. By extracting information about different pathologies (e.g., different types of cancer) from histopathology reports, we will be able to study characteristics of disease that could lead to improved diagnosis and treatment.

### What data will be collected and how will it be used?

All histopathology reports from participating hospitals will be collected alongside basic characteristics of the patients (e.g., age, gender, and ethnicity). The data will first be processed to remove information that can identify individuals (e.g., names, date of births, etc.). Reports will be labelled with a code to enable information from the same patient to be linked.

Less than 1 in 100 histopathology reports contain the patient's name. In the tests we have done, we have successfully removed all personal identifiers from the reports ensuring that individuals cannot be identified in any manner. However, we cannot give an absolute guarantee that we would be able to achieve this in all cases in the future. We take privacy and security very seriously and therefore, once processed, the data will be transferred to a secure computer system that meets stringent NHS security standards. Only named researchers with training in analysing sensitive data will be permitted access to the data, which will be monitored.

A small proportion of the reports will be read by a researcher and specific features labelled. Should the researcher come across any data that could be used to identify specific individuals, these will be immediately deleted. The reason why they were not removed will be investigated and steps taken to avoid it happening again. The remaining data will then be analysed by 'Natural Language Processing' (NLP).

The study is planned for three years in the first instance but may be extended (with appropriate approval), dependent on progress. Once the study is complete, the original (full text) histopathology reports will be deleted.

### What will happen with the results?

It is hoped that this study will provide large amounts of accurate data on the patterns of disease over time and particular disease characteristics, for example, the best markers (special pathology tests) to identify certain types of cancer.

Findings from this study will be shared with other doctors, scientists, researchers, and the public through presentations at conferences and through publications in freely accessible journals following review by experts in the field. In addition, summary data, which will not contain information that can identify specific patients, will be made accessible online. These will be available for other doctors and scientists to help them with diagnosing and studying disease. It will not be possible to identify any patient from the published results or summary data.

#### What do patients and the public say about this research study?

To determine what patients and the public think about this research project, an anonymous survey was conducted. A total of 57 responses were received. Approximately two thirds of the respondents reported having had a biopsy or undergone surgery to remove tissue that was reviewed by histopathology. This confirms that the respondents had appropriate lived experiences relevant to the study.

All survey respondents felt that the research was important. Furthermore, the vast majority (95%) of respondents felt that it was acceptable to analyse histopathology reports in the described way for the described task. No respondents felt that the proposed research was unacceptable.

#### What if someone does not want their data included in the study?

Patients who do not want their data included in the study can opt out via two mechanisms:

- All information from patients that have declined use of their data for research via the [NHS National Data Opt Out programme](#) will be removed.
- Patients can ensure their data are not included in this specific study by contacting [UCLH.cellpath-research@nhs.net](mailto:UCLH.cellpath-research@nhs.net) with the subject line '*NLP research opt-out*' including their name, date of birth and NHS number. Alternatively, patients can call the UCLH Department of Cellular Pathology on 020 3456 8402.

The care and treatment of those opting out of the study will not be affected.

#### Who is in charge of the study?

The study is being led by Dr Adam Levine from the Research Department of Pathology, University College London, University College London Hospitals NHS Foundation Trust, and Royal Free London NHS Foundation Trust.

#### Who has reviewed and approved this study?

The study is sponsored and supported by University College London (reference 152999). This study is registered with the NHS Health Research Authority (IRAS reference 293404). The study has been reviewed and approved by the HRA Research Ethics Committee (reference 23/LO/0253). The HRA has given section 251 support for this activity following advice from the Confidential Advisory Group (CAG) (reference 23/CAG/0032).