Swansea University
Prifysgol Abertawe

Population Data Science
Faculty of Medicine, Health & Life Science

Gwyddor Data Poblogaeth
Y Gyfadran Meddygaeth, Gwyddor Iechyd a Bywyd

# POPULATION DATA SCIENCE SUMMER INTERNSHIP PROGRAMME 2026

Start your career in data science! >

# About our award-winning team

## Who we are

Population Data Science at Swansea University is a cutting-edge research centre based in Swansea University Medical School, Faculty of Medicine, Health & Life Science. Specialising in secure-linked data analysis of health, social and other administrative data sources.

**WE ARE PIONEERING POPULATION DATA SCIENCE**

## What we do

Our award-winning team are focused on delivering a world-class technology infrastructure to tackle some of the biggest challenges facing today's society and enable opportunities - creating a global research ecosystem with the potential to produce educational, environmental, economic, and societal impact.

## Our research projects

Our team is involved with various projects and research programmes, conducting different methodological and applied skills across multiple disciplines to support and achieve our world-leading innovations. These skills include but are not limited to software development, database development and management, epidemiology, statistics, data mining, data visualisation, web development, project management, patient public and stakeholder engagement and Geographical Information Systems.

**Data Science Building**

**Swansea University**

# Why should you apply to our internship programme?

1 Gain work experience with an award-winning team at the forefront of global population data science research.

2 Develop new skills with real-world data analysis to develop and apply your skills in this highly secure, state-of-the-art environment.

3 Explore a career path in Population Data Science and find out what fits you best.

4 Gain confidence with our internship programme. It is a perfect way to help you gain experience transitioning from university to full-time employment. You can gain valuable insights and knowledge, participate in meetings, and perform assigned tasks in a real-world setting.

5 Network and work alongside a team of highly skilled professionals including your supervisors and mentors you meet during your internship, who can be valuable references as you pursue a full-time job.

6 Build a strong CV and give yourself a head start with the internship.

7 Some interns have become full-time employees: Kellie Robinson and Rowan Dash from our 2023 intern cohort, both secured full-time positions in the SERP team at Population Data Science after their internships. Find out more about their journey here.

8 Paid employment for 12 weeks, including annual leave entitlement.

9 The programme culminates in the Showcase Event with an opportunity for interns to present their projects. Check out the highlights from last year's Showcase event in this video.

**Apply now!**

## Keep scrolling to see the project opportunities!

## Project 7

### Title: Automated pre-processing pipelines for multimodal tabular health data in machine learning research

**Summary:** Health data research often involves large, complex tabular datasets containing clinical information, demographic records, and other structured health data. These data are frequently messy and inconsistent, with challenges such as missing values, mixed data types, non-standard coding, and data entry errors. Preparing such heterogeneous data for machine learning workflows is time-consuming, difficult to standardise, and hard to reproduce across projects.

This internship will focus on designing and implementing an automated pipeline for pre-processing tabular health datasets. The purpose of the project is to develop a reliable, reusable tool that streamlines data preparation by handling a range of data types, including numerical, free-text, categorical, and time-series data, and applying consistent, well-documented transformation steps. The resulting pipeline will reduce manual effort, minimise errors, and ensure that tabular health datasets are machine-learning-ready in a consistent and reproducible manner, directly supporting dementia and population health studies conducted within trusted research environments.

### Outcome by the end of 12 weeks

- Developed a modular, automated pre-processing pipeline for heterogeneous tabular health datasets.
- Implemented steps supporting numerical, categorical, text, and timeseries variables.
- Produced clear documentation and user guides to support future use and adaptation of the pipeline.
- Gained experience with real-world datacleaning workflows and machinelearning preparation techniques.
- Contributed maintainable code that can be reused in future research projects.

### Skills/ background suited

**Essential**

- Experience with Python programming.
- Interest or background in end-to-end data science.
- Familiarity with basic machine learning concepts.

**Desirable**

- Experience with healthcare or population-level datasets.
- Knowledge of natural language processing, metadata handling, or timeseries processing.
- Interest in reproducible automated data workflows.

# The centres supporting this year's programme

## SAIL Databank

SAIL Databank is a safe haven for billions of person-based records, which enables researchers to answer important questions for the benefit of society. Researchers can access a broad range of routinely collected data spanning up to 30 years from an entire population.

SAIL Databank achieves this by working collaboratively with data guardians, academics, regulators, members of the public, practitioners, and policymakers from Wales, across the UK and internationally.

The SAIL Databank holds de-identified data in linkable form and, following further safeguards, makes selected data available for analysis in anonymised form. Approved researchers can access the data remotely anywhere in the world, complete with analysis tools. Because the data accessed for research is anonymised, the work is carried out without researchers knowing the identities of the individuals represented.

## SeRP

Secure eResearch Platform (SeRP) is the complete customisable solution for data sharing, linkage and analysis in a safe, secure and controlled environment that's accredited to the highest international standard. It does this in a governed environment allowing data owners to remain in full control at all times.

SeRP is the perfect solution for any organisation that has accumulated a large amount of data, that intends to share that data for the purposes of research and long term benefit to society, and that wants to do so in the most secure and safe way possible to reduce associated risks. SeRP can benefit sectors including healthcare, governments, academic Research, industry, charities & Not-for-profit organisations.

For more information on the benefits of SeRP, visit SeRP Benefits.

# The centres supporting this year's programme

## ADR Wales

ADR Wales (Administrative Data Research Wales) unites specialists in each field from Population Data Science at Swansea University and the Wales Institute of Social and Economic Research and Data (WISERD) at Cardiff University with statisticians, economists and social researchers from Welsh Government.

The partnership is ideally placed to maximise the utility of anonymous and secure data to shape public service delivery, which will ultimately improve the lives of people in Wales. The work carried out by the ADR Wales team covers areas such as early years, education, skills and employability, health and well-being, housing and homelessness, social care, mental health, social justice, climate change and major societal challenges.

## Dementias Platform UK

Dementias Platform UK Data Portal brings together records of over 3 million people in a free-to-access resource.

Researchers can use the portal to identify which cohorts are relevant to them, apply for access to the data and then analyse it in a secure, remote environment complete with data linkage and analysis packages.

The portal brings the records together, ensures they are secure, supports data linkage studies using phenotypic, genomic, and imaging data, and makes them readily available for the benefit of academia, industry, regulators, health care providers, patients, and the public.

# APPLY NOW!

To find out more about the

**Summer Internship Programme**

Visit our website >

Follow us

@popdatasci_su

@PopDataSci_SU

@popdatascisu.bsky.social

Population Data Science
at Swansea University