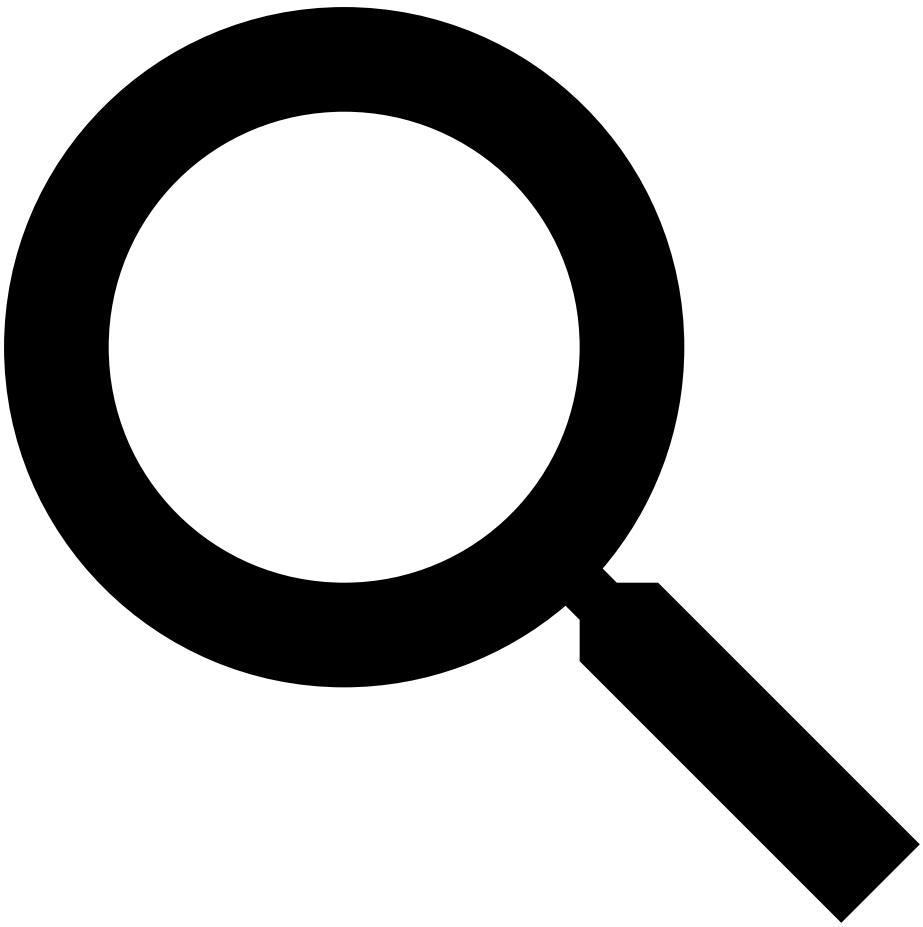
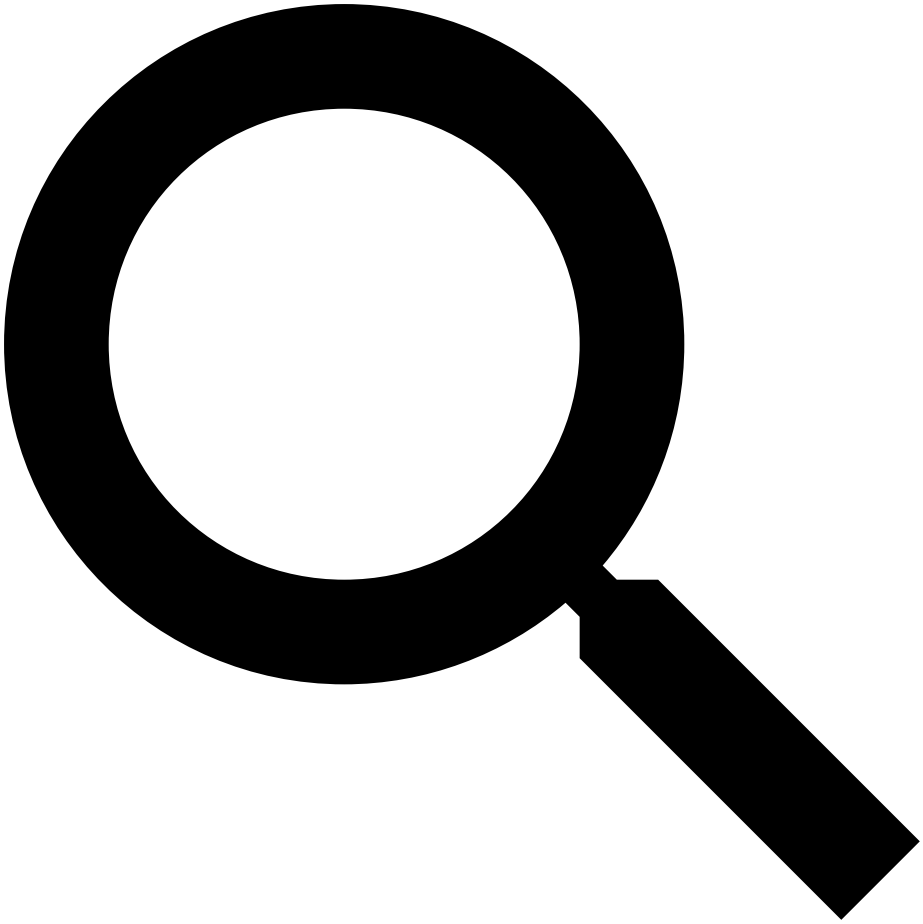
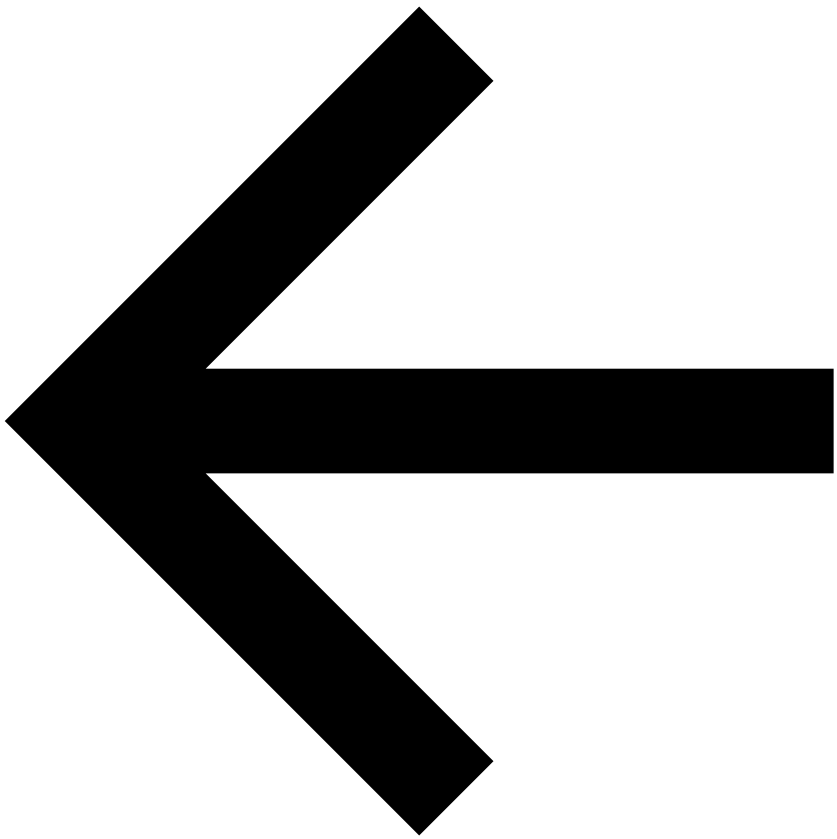
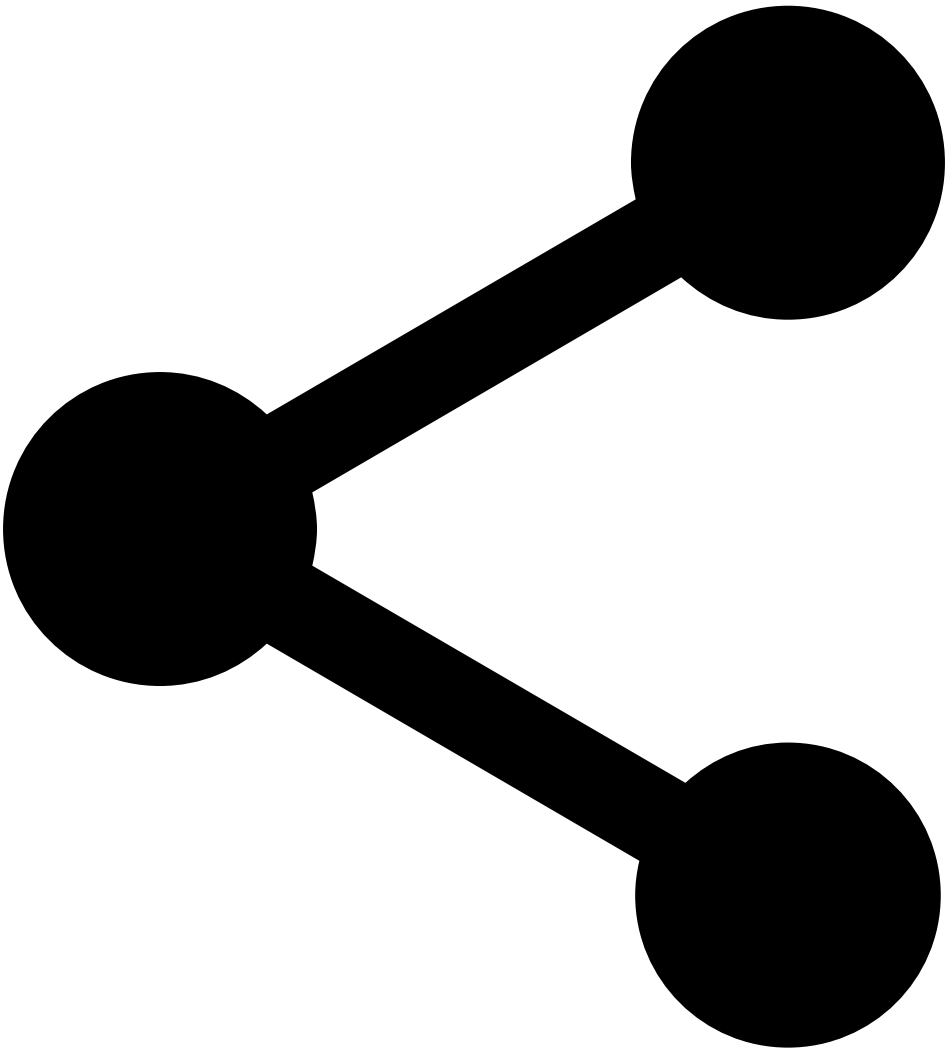
Skip to content

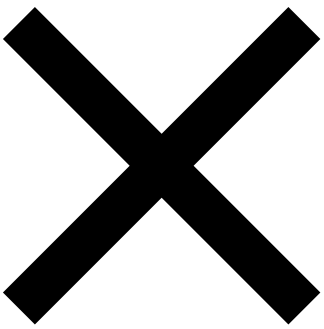
NHS England Data Science
Overview - Data Linkage Hub



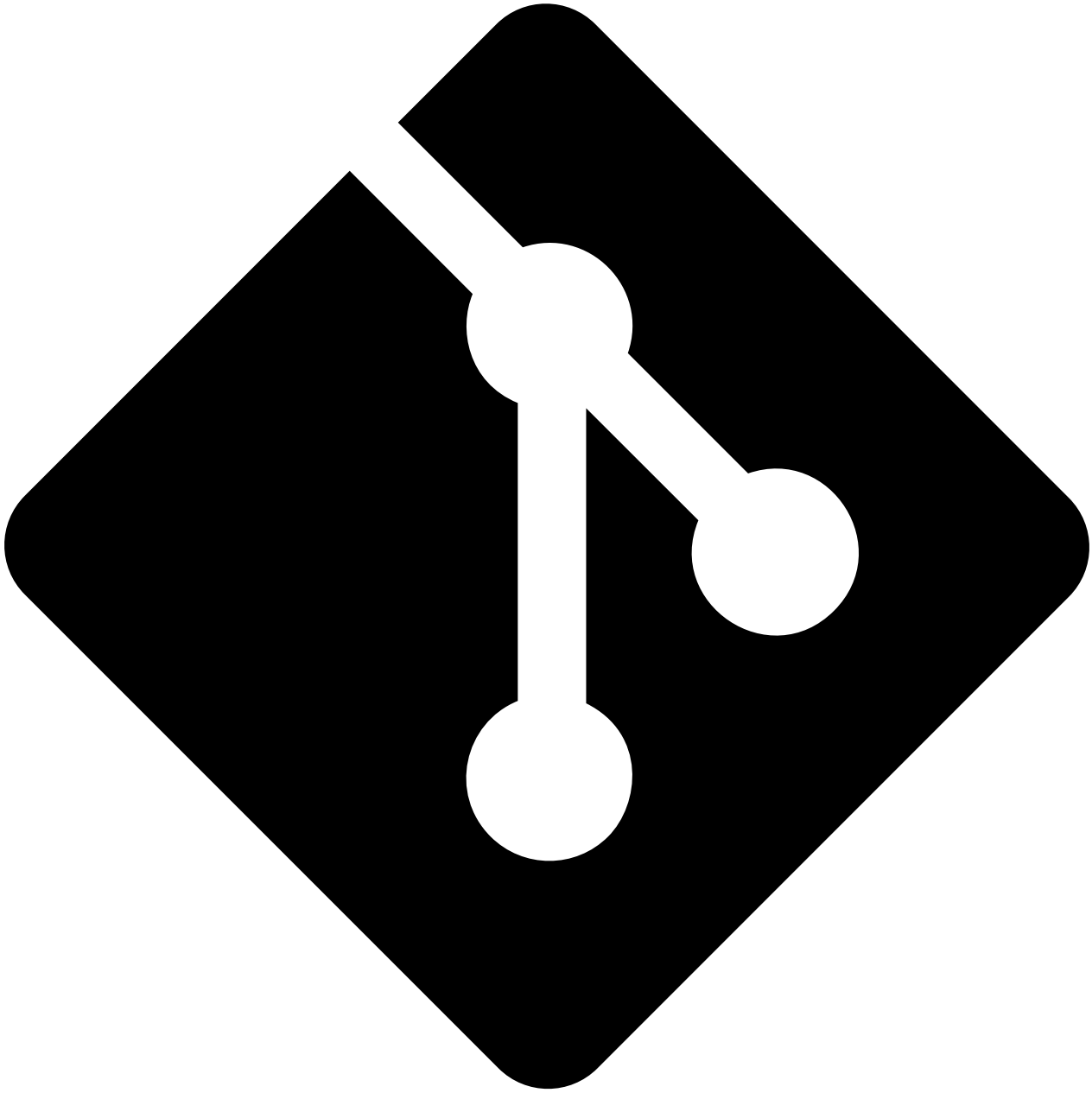








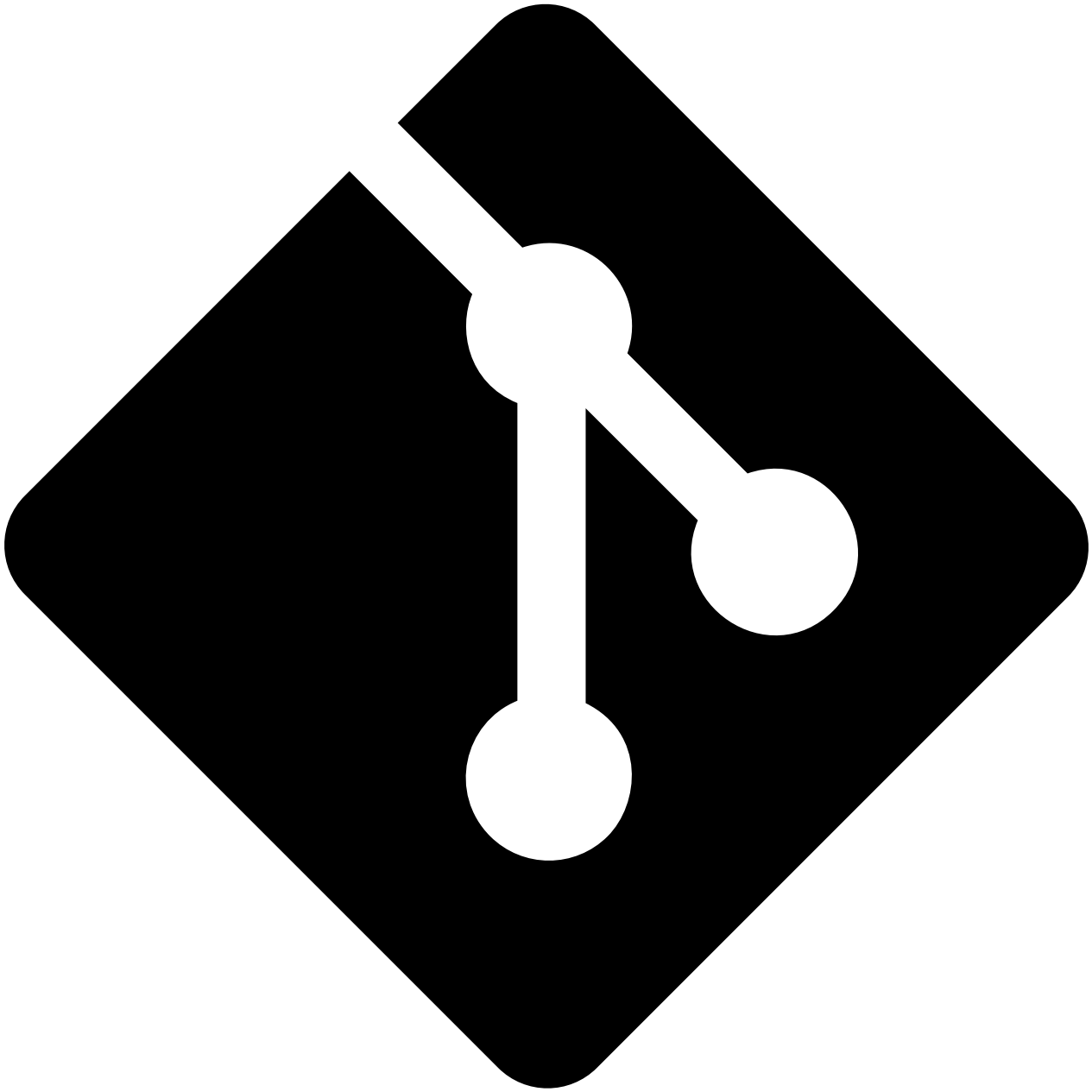
Initializing search



NHS England Data Science Website

- Home
- About the team
- Projects
- PhD Internships
- Data Science MRes
- Articles
- Playbooks
- Useful links
- Site Info

NHS England Data Science



NHS England Data Science Website

Current Projects

- A&E Forecasting Tool
- AI Assurance Research Path
- AI Dictionary
- AI Ethics in Practice
- Better Matching Algorithm
- Cancer high-risk cohorts
- Corporate Services: Pipeline rebuild
- CVD Pathways
- Data Linkage Community of Practice
- Diabetic Retinopathy
- Emerging Privacy Enhancing Technologies
- ePMA Auto Coding
- NHS.UK Automatic Moderation of Ratings & Reviews
- Primary Care Data Generator
- Reproducible Analytical Pipelines Squad
- Reusable Data Validation Process
- Risk stratification models for Population and Person Insights (PaPI)
- Embedded Data Scientists in the National Disease Registration Service
- Clinical Measurement Extractor
- PLOS: Predicting Length of Stay

- ☐

Past Projects

Past Projects

- ☐

2026

2026

- Forecasting Vaccination Demand

- ☐

2025

2025

- NHS.UK Automatic Moderation of Ratings & Reviews
- Reproducible Analytical Pipelines Squad
- Waiting List Minimum Dataset (WLMDS) Proofs of Concept
- Synthetic Clinical Notes

- ☐

2024

2024

- RAG
- Redbox Copilot
- Quality Assurance Framework for Data Linkage
- Tool to Assess Privacy Risk of Text Data - Extended
- Understanding Fariness and Explainability in Multimodal Approaches within Healthcare

- ☐

2023

2023

- AI Models for Shortlisting Interview Candidates
- AI Skunkworks Team
- Deep Learning to Detect Adrenal Lesions in CT Scans
- MPS Handbook
- Enriching Clinical Coding for Neurology Pathways using MedCAT
- Including Mortality in Hypergraphs for Multi-morbidity

- Investigating Privacy Risks and Mitigations in Healthcare Language Models
- NHS Synth
- Parkinson's Disease Pathology Prediction
- Process Mining with East Midlands Ambulance Service
- Tool to Assess Privacy Risk of Text Data

- ☐ 2022
  2022
  - Adding a Clinical Focus to Evaluating MM Data Representations
  - Ambulance Handover Delay Predictor
  - Applying & Evaluating a Language Model to Patient Safety Data
  - Bed Allocation
  - Generic Patient Simulator
  - Inequalities in Diabetes from PHM Data
  - Investigating Superpixels in LIME for Explaining Predictions of Facial Images
  - Long Stayer Risk Stratification Baseline Models
  - Nursing Placement Scheduled Optimisation
  - Privacy of Unstructured Data
  - Renal Health Prediction
  - Synthetic Data From Real Data
  - Synthetic Data Generation Pipeline
  - Transforming Healthcare Data With Graph-Based Techniques
- ☐ 2021
  2021
  - AI Deep Dive Workshops
  - CT Alignment & Lesion Detection
  - Creating a Generic Adversarial Attack for Synthetic Data
  - Developing SynthVAE
  - Differential Privacy in a VAE for Synthetic Data Generation
  - Impact of Commercial Data on Predictions
  - Length of Hospital Day Prediction
  - NHS Language Corpus
  - NHS @Home Programme
  - Predicting Negligence Claims
  - SynPath Simulator on Diabetes Pathway
  - Text Analysis using Structural Topic Modelling
  - TxtRayAlign
- ☐ 2020
  2020
  - Data Lens

◦ ■ Main Work Areas
Main Work Areas
- ☐ Predictive Analytics Products
  Predictive Analytics Products
  - A&E Forecasting Tool
  - Ambulance Handover Delay Predictor
  - Bed Allocation
  - Cancer high-risk cohorts
  - Diabetic Retinopathy

- Impact of Commercial Data on Predictions
- Length of Hospital Day Prediction
- Long Stayer Risk Stratification Baseline Models
- Parkinson's Disease Pathology Prediction
- Predicting Negligence Claims
- Renal Health Prediction
- Risk stratification models for Population and Person Insights (PaPI)
- Waiting List Minimum Dataset (WLMDS) Proofs of Concept
- PLOS: Predicting Length of Stay
- Forecasting Vaccination Demand

- ■
  Data Science for Linked/Longitudinal Data
  Data Science for Linked/Longitudinal Data
  - ■
    Data Linkage Hub
    Data Linkage Hub
    - Better Matching Algorithm
    - Data Linkage Community of Practice
    - MPS Handbook
    - Quality Assurance Framework for Data Linkage
  - Inequalities in Diabetes from PHM Data
  - CVD Pathways
- □
  Natural Language Processing Products
  Natural Language Processing Products
  - AI Models for Shortlisting Interview Candidates
  - Data Lens
  - ePMA Auto Coding
  - NHS.UK Automatic Moderation of Ratings & Reviews
  - Redbox Copilot
  - Tool to Assess Privacy Risk of Text Data
  - Tool to Assess Privacy Risk of Text Data - Extended
- □
  Data Science Capability
  Data Science Capability
  - AI Deep Dive Workshops
  - AI Dictionary
  - AI Ethics in Practice
  - AI Skunkworks Team
  - Corporate Services: Pipeline rebuild
  - NHS @Home Programme
  - Reproducible Analytical Pipelines Squad
- □
  Research & Development
  Research & Development
  - □
    Computer Vision retired
    Computer Vision retired
    - CT Alignment & Lesion Detection
    - Deep Learning to Detect Adrenal Lesions in CT Scans
    - Investigating Superpixels in LIME for Explaining Predictions of Facial Images
  - □
    Natural Language Processing

Natural Language Processing
- Adding a Clinical Focus to Evaluating MM Data Representations
- Applying & Evaluating a Language Model to Patient Safety Data
- Including Mortality in Hypergraphs for Multi-morbidity
- Investigating Privacy Risks and Mitigations in Healthcare Language Models
- NHS Language Corpus
- Privacy of Unstructured Data
- Text Analysis using Structural Topic Modelling
- TxtRayAlign
- Understanding Fariness and Explainability in Multimodal Approaches within Healthcare

- Synthetic Data
Synthetic Data
  - Creating a Generic Adversarial Attack for Synthetic Data
  - Developing SynthVAE
  - Emerging Privacy Enhancing Technologies
  - Differential Privacy in a VAE for Synthetic Data Generation
  - Generic Patient Simulator
  - NHS Synth
  - Primary Care Data Generator
  - SynPath Simulator on Diabetes Pathway
  - Synthetic Clinical Notes
  - Synthetic Data From Real Data
  - Synthetic Data Generation Pipeline
- Including Mortality in Hypergraphs for Multi-morbidity
- Nursing Placement Scheduled Optimisation
- Process Mining with East Midlands Ambulance Service
- SDE Service Data Wranglers
SDE Service Data Wranglers
  - Reusable Data Validation Process
- Transforming Healthcare Data With Graph-Based Techniques

  - Our Team's Publications
- PhD Internships
- Data Science MRes
Data Science MRes
  - MRes Projects
MRes Projects
    - Prediction of CVD Onset
    - Supporting Dementia Diagnosis
    - Understanding delays in elective pathways
    - Investigating unknown ethnicity records in NHS emergency care data
    - Predicting GP Staff Turnover
    - Menopause-Related Diagnosis Coding
    - Mental Health Environmental Impacts Leeds
    - Impact of midwife-led continuity of carer on birth outcomes
    - Predicting perinatal depression
    - Poor pregnancy outcomes in women with type 2 diabetes
    - Process Mining on Patient Pathways in Healthcare
    - Relationship between psychotropic medication usage and Talking Therapies treatment outcomes

Overview - Data Linkage Hub

The data linkage hub encompasses all things data linkage, from documenting the existing state of linkage in NHS England in the Person_ID handbook, to exploring

better matching algorithms using probabilistic models and Splink, to creating a Quality Assurance Framework for Data Linkage.

BEST PRACTICE LINKAGE

Diagram representing the four current areas of the data linkage hub: DL Quality Assurance, Better Matching Algorithm, MPS Documentation, and the DL Community of Practice.

Data Linkage is a business-critical process within many government organisations, including NHS England. Being able to link patients across their care journey enables both direct care services and research studies on admin data, which in turn, influences healthcare policies. So taking care of this important service is why the Data Linkage hub was created in the new NHS England.

The role of the Data Linkage hub in NHS England includes:

- identifying points of collaboration with other government departments
- mapping the stakeholders involved in data linkage - both internal and external
- feeding user needs to the Data Linkage vision

Work we do

Info

Click each heading to find out more!

Quality Assurance Framework

If we want to achieve a consistent and high quality approach to linking data, which allows for robust, transparent and auditable results, we also need a framework to operate within. Hence, this workstream aims at creating, testing and implementing in the business process a Quality Assurance Framework for Data Linkage.

Better Matching Algorithm

We're currently working on implementing a probabilistic linkage model using Splink, in order to improve linkage outcomes, and by extension, patient outcomes.

Community of Practice

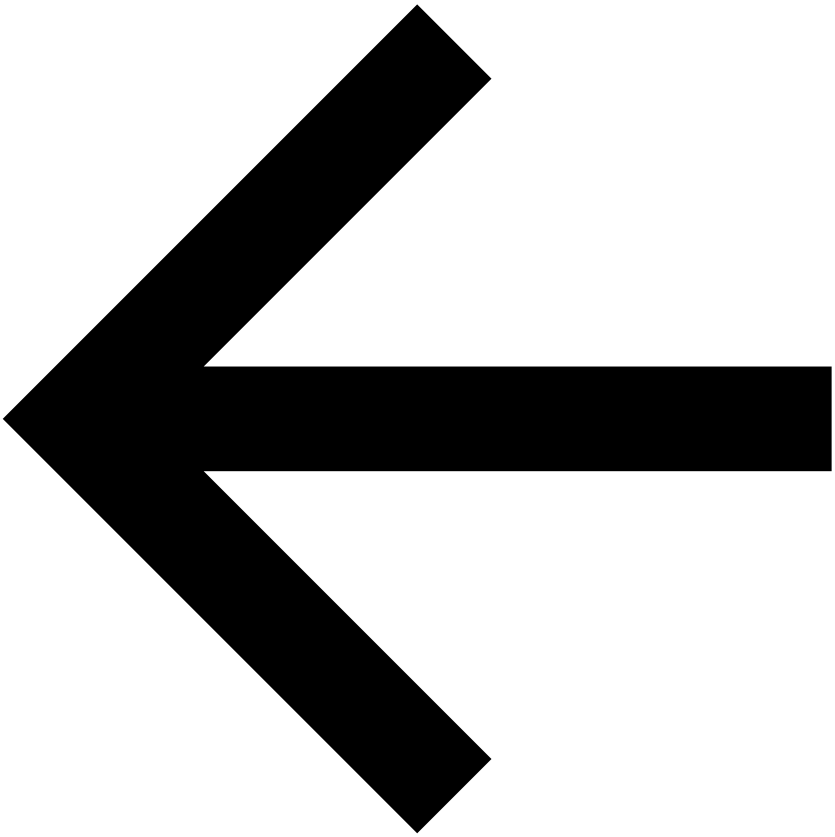
We are fostering a community of practice in NHS England to help people do the best linkage they can, and encourage them to be connected with the cross-government Data Linkage Champions network. The community of practice is open to any data linkage stakeholders in NHS England - to join the community of practice go here.

MPS Documentation

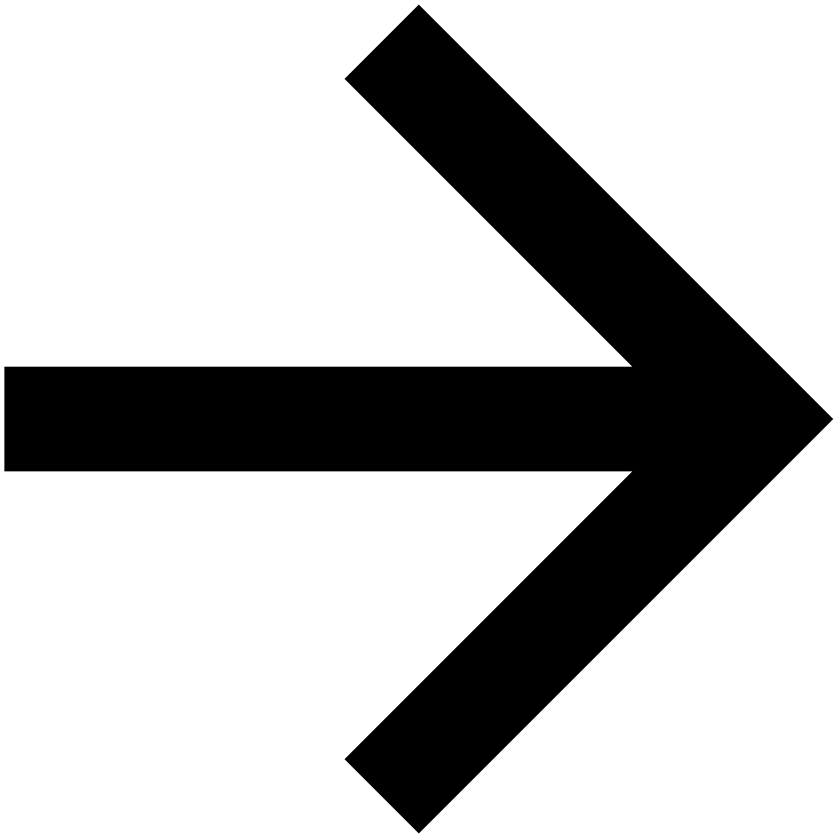
We have been documenting how the Person_ID is generated via the Master Person Service, to make the current process of linking data in the NHS more transparent and easy to understand.

Output	Link
MPS Diagnostics	Github

Output	**Link**
Person_ID Handbook	NHS England Website
Quality Assurance Framework	Work in Progress Link
Community of Practice (*internal only*)	Teams Channel



Previous Forecasting Vaccination Demand
Next Better Matching Algorithm



Contact Us

Developed by NHS England Data Science Profession.

View the team on GitHub