⚠️ CONFIDENTIAL DOCUMENT

Classification: CONFIDENTIAL - ACADEMIC RESEARCH
Distribution: Restricted to University of Westminster Academic Staff Only
Purpose: MSc Dissertation Assessment
Handling: This document contains proprietary research methodologies and unpublished findings.
Unauthorised distribution, reproduction, or disclosure is strictly prohibited.

UNIVERSITY OF WESTMINSTER

Faculty of Life Sciences

MSc Artificial Intelligence and Digital Health

Author: Donald Philp
Student Number: 21075797
Supervisor: Dr Tom Oliver
Department: Computer Science and Engineering
Date: 02 September 2025

Document Type: Statistical Analysis Report
Analysis Framework: Two-Stage Hybrid Feature Extraction with SLATE† Architecture
Statistical Confidence: 95% CI with Bootstrap Resampling (n=1,000)

Hypothesis Framework

Research Question: Do UK Biobank studies converge on a small, stable "core" set of features (age, sex, BMI, blood pressure, smoking, etc.) across diverse disease domains, or do different research fields fragment into divergent, silo-specific subsets?

Hypotheses:

- **H_0 (Null):** Feature usage is fragmented across studies with no consistent core pattern
- **H_1 (Alternative):** A statistically significant **core-periphery structure** exists: a small recurring core used across domains vs. silo-specific peripheral features

Scientific Value: Tests whether UK Biobank research genuinely **breaks traditional research silos** and identifies universal health determinants that transcend disease boundaries.

Data Source

UK Biobank Publication Database

This research utilised the official UK Biobank journal research database, accessible through the UK Biobank data showcase: https://biobank.ndph.ox.ac.uk/ukb/schema.cgi?id=19

Dataset Characteristics

- **Data Collection Period:** April 2025
- **Format:** Tab-delimited text file (TSV)
- **Total Publications:** 8,553 research articles
- **Temporal Range:** 2013 – 2025 (partial year)

Publication Timeline Distribution

Year	Publications
2013	5
2014	20
2015	31
2016	94
2017	169
2018	322
2019	511
2020	828
2021	1,209
2022	1,335
2023	1,912
2024	2,113
2025*	4

**2025 data reflects partial year through April 2025*

Project Development Metrics

Metric	Value	Detail
Development Period	27 days	7 August 2025 – 1 September 2025
Total Development Hours	165.9 hours	Active coding and analysis time
Average Daily Effort	6.5 hours/day	Across calendar period
Work Sessions	39 sessions	Average 4.3 hours per session
Peak Work Day	17.0 hours	13 August 2025 (3,402 files modified)
Files Processed	46,676 files	Across 189 directories
Code Efficiency	108.5%	Above standard 6-hour workday benchmark
Project Intensity	High	25 of 27 days actively worked

1. Methodology Overview

The analysis employed a robust, full-stack biomedical research pipeline meticulously designed for the UK Biobank journal dataset, integrating diverse technologies to achieve efficient and accurate transformation of raw unstructured scientific literature into structured, semantically rich data. This full-stack technology leverages Python for core logic, Docker for containerization, NVIDIA CUDA with PyTorch for GPU acceleration, and advanced LLM/NLP frameworks like Hugging Face Transformers, spaCy, and BERTopic for semantic content processing.

Throughout its development, **16 distinct, active components** or "experiments" were successfully integrated, while **11 "experiments" or modules** were explored and subsequently deprecated or remained unused. These deprecated modules were either superseded by more robust approaches or removed due to a strategic pivot towards the core hypothesis.

Active Methodological Stages

Stage	Key Activities	Core Technologies/Libraries	Key Models / LLMs	Validation Approaches
Data Ingestion	PDF acquisition, deduplication, metadata validation	Python, Requests, PyPDF2, hashlib, Docker	N/A (Rule-based)	MD5 checksums, file content checks, functional tests
PDF to MD/JSON Conversion	PDF to Markdown & JSON transformation for NLP	Python, Unstructured, PyMuPDF, Docker (GPU)	Surya-OCR (Marker-PDF)	Output file checks, metadata heuristic validation
Enhanced Document Chunking	Semantic chunking, section classification, missing category embedding, human feedback loop	Python, transformers, Pytorch, scikit-learn, sentence-transformers, BERTopic, Docker (GPU)	BioMistral-7B, Llama-3.1-8B-Instruct, BiomedBERT, SciBERT, BGE-M3	Manual review, human-in-the-loop, clustering metrics
Advanced Markdown Cleaning	Linguistic-model driven text cleaning	Python, SpaCy, scispacy	SciBERT (via en_core_sci_scibert)	Detailed audit reports, quantitative text quality metrics
Core-Periphery Analysis	Feature extraction, statistical hypothesis testing & results	Python, pandas, numpy, scipy, scikit-learn, ahocorasick, Docker	BioMistral-7B (LLM Validation)	Multi-stage matching, LLM-based validation, bootstrapping, Gini coefficient

2. Preliminary Results

Bootstrap Validation (n=1,000 iterations)

Metric	Value	Interpretation
Papers Analysed	6,375	74.5% of available UK Biobank papers
Unique Features	3,725	35.5% of total UK Biobank features
Core Stability (Jaccard)	0.970	Extremely stable core across bootstrap samples
Jaccard 95% CI	[0.905, 1.000]	High confidence in core consistency
Gini Coefficient	0.866	Strong concentration (few features dominate)
Top-50 Coverage	96.6%	50 features cover nearly all papers
Top-20 Coverage	91.7%	Remarkable concentration in top features

Statistical Significance

Result: H_1 ACCEPTED - Strong Core-Periphery Structure Confirmed

- **Jaccard similarity of 0.970** → core stable across subsamples
- **Gini coefficient of 0.866** → extreme inequality in feature usage
- **Top-50 features achieve 96.6% coverage** → 50 of 10,489 features dominate research

The Universal Core Features

Field ID	Feature	Domain	Stability
23104	Body Mass Index (BMI)	Anthropometry	100%
23098	Weight	Anthropometry	100%
100022	Alcohol Intake	Lifestyle	100%
100003	Protein Intake	Diet	100%
20116	Smoking Status	Lifestyle	100%
12144	Height	Anthropometry	100%
22009	Genetic Principal Components	Genetics	100%
10721	Life Stressors	Psychology	100%
100002	Energy Intake	Diet	100%
26005	Protein Biomarkers	Biochemistry	100%

Interpretation: These features form the **invariant core** of UK Biobank research, representing fundamental health determinants that transcend specific disease domains. **Domain:** The “Domain” column is author-defined for readability; it is not official UK Biobank metadata.

Executive Summary

Critical Performance Indicators

Metric	Value	Coverage	Implication
UK Biobank Features Utilised	3,725 / 10,489	**35.5%**	Core-periphery confirmed
Papers Analysed	6,376 / 8,553	**74.5%**	Substantial corpus coverage
Successful Extractions	6,221 / 6,375	**97.6%**	High extraction efficacy
Mean Features per Paper	12.5	-	Research depth indicator

▶ **Core-Periphery Structure Confirmed:** Top 50 features (0.48% of total) cover 96.6% of all research
▶ **6,764 features remain unexplored**, representing significant research opportunities
▶ **Gini coefficient of 0.866** indicates extreme concentration
▶ **Bootstrap validation** shows core features are invariant (Jaccard = 0.970 across 1,000 iterations)

3. Re-evaluation

While the results presented here are compelling and reveal an interesting core-periphery structure, I want to express a degree of caution regarding the methodology. I am confident in the results so far, but I intend to re-evaluate and refine my approach to ensure maximal scientific rigor and to feel fully satisfied with the analysis. This re-application of the methodology will strengthen the foundation of these promising findings.

Statistical Glossary for Hypothesis Validation

Core-Periphery Structure: Pattern where small set of features (core) appears across most studies, while large set (periphery) appears rarely

Bootstrap Validation: Resampling technique that tests stability by creating 1,000 random subsamples and checking if the same core emerges

- Jaccard = 0.970 means 97% overlap between cores from different samples
- 95% CI [0.905, 1.000] means we're 95% confident true overlap is between 90.5% and 100%

Gini Coefficient (0.866): Measures inequality in feature usage

- 0 = all features used equally (no core)
- 1 = one feature used by everyone (ultimate core)
- 0.866 = extreme concentration, strong core-periphery structure

Coverage Analysis: How many papers use top features

- Top-50 features → 96.6% paper coverage
- Means 50 of 10,489 features (0.48%) capture nearly all research
- Strong evidence for universal core hypothesis

Why These Tests Matter: Multiple independent statistical tests (Gini, Jaccard, Coverage) all converging on the same conclusion validates that the core-periphery structure is real, not an artifact. This supports the hypothesis that UK Biobank research has identified universal health determinants that transcend traditional medical specialities.

†SLATE (Sentence-Level Annotation & Tagging Engine): Novel computational framework introduced in this work for preserving JSON structural integrity during overlapping text chunk processing. This methodology represents an original contribution to the field of biomedical text mining.