Diabetes prevalence management and health inequalities

Digital Analysis and Research Team (DART)
Data Science Internship

Author: *Stephen Richer*

Supervisor(s):	*Paul Carroll & Jonny Pearson*
Department:	Digital Analysis and Research Team (DART)
Date:	January 1, 2023

Contents

List of Figures

List of Tables

Acronyms

BMI Body Mass Index.

BP Blood Pressure.

DNA Did Not Attend.

ESNEFT East Suffolk and North Essex NHS Foundation Trust.

GFR Glomerular Filtration Rate.

GP General Practice.

HbA1C Hemoglobin A1C.

ICD-10 International Classification of Diseases.

IMD Index of Multiple Deprivation.

IoD English Indices of Deprivation (2019).

LSOA Lower Layer Super Output Area (2011).

NHS National Health Service.

NICE National Institute for Health and Care Excellence.

ONS Office for National Statistics.

OSM Open Street Map.

QoF Quality and Outcomes Framework (2021-2022).

T1D Type 1 Diabetes Mellitus.

T2D Type 2 Diabetes Mellitus.

Chapter 1

ESNEFT Tools

Python Suite for Demographic and Health Inequality Analysis.

Summary

The utilisation of demographic data is vital for understanding patterns of health inequalities within a population. Demographics are data that describe a population; they include metrics such as age, ethnicity and sex, as well well socio-economic factors such as employment and education. Helpfully, a wide variety of demographic data are available within the public domain. However, these datasets are spread across disparate sources and additional work is required to link and aggregate the data before any demographic analysis can begin.

The collection and aggregation of demographic data is a common component of many analytical workflows. As such we sought to simplify and standardise the process of demographic data collection by developing ESNEFT Tools . ESNEFT Tools is an open-source, Python-based utility for collecting, processing and aggregating a wide a variety of datasets that are frequently utilised for demographic analysis and health inequality research.

In addition to performing data aggregation, ESNEFT Tools includes additional data manipulation methods to more thoroughly investigate patterns of disease prevalence and deprivation. Disease prevalence data, available at General Practice (GP) level from the Quality and Outcomes Framework (2021-2022) (QoF), are combined with GP registration data to estimate disease prevalence per Lower Layer Super Output Area (2011) (LSOA). Similarly, LSOA level deprivation statistics are combined with GP registration data to estimate deprivation per registered GP population. Finally, ESNEFT Tools includes a number of easy-to-use, customisable methods for interactive visualisation of demographic data.

Taken together, ESNEFT Tools implements a simple interface for accessing and manipulating a wide variety of datasets relating to demographic and population inequalities.

Source Code & Documentation: GitHub: ESNEFT Tools

1.1 ESNEFT Tools Overview

The ESNEFT Tools workflow is illustrated in fig. 1.1 and the public data sources collected by the utility are detailed in table 1.1 on the next page.

1.1.1 Workflow

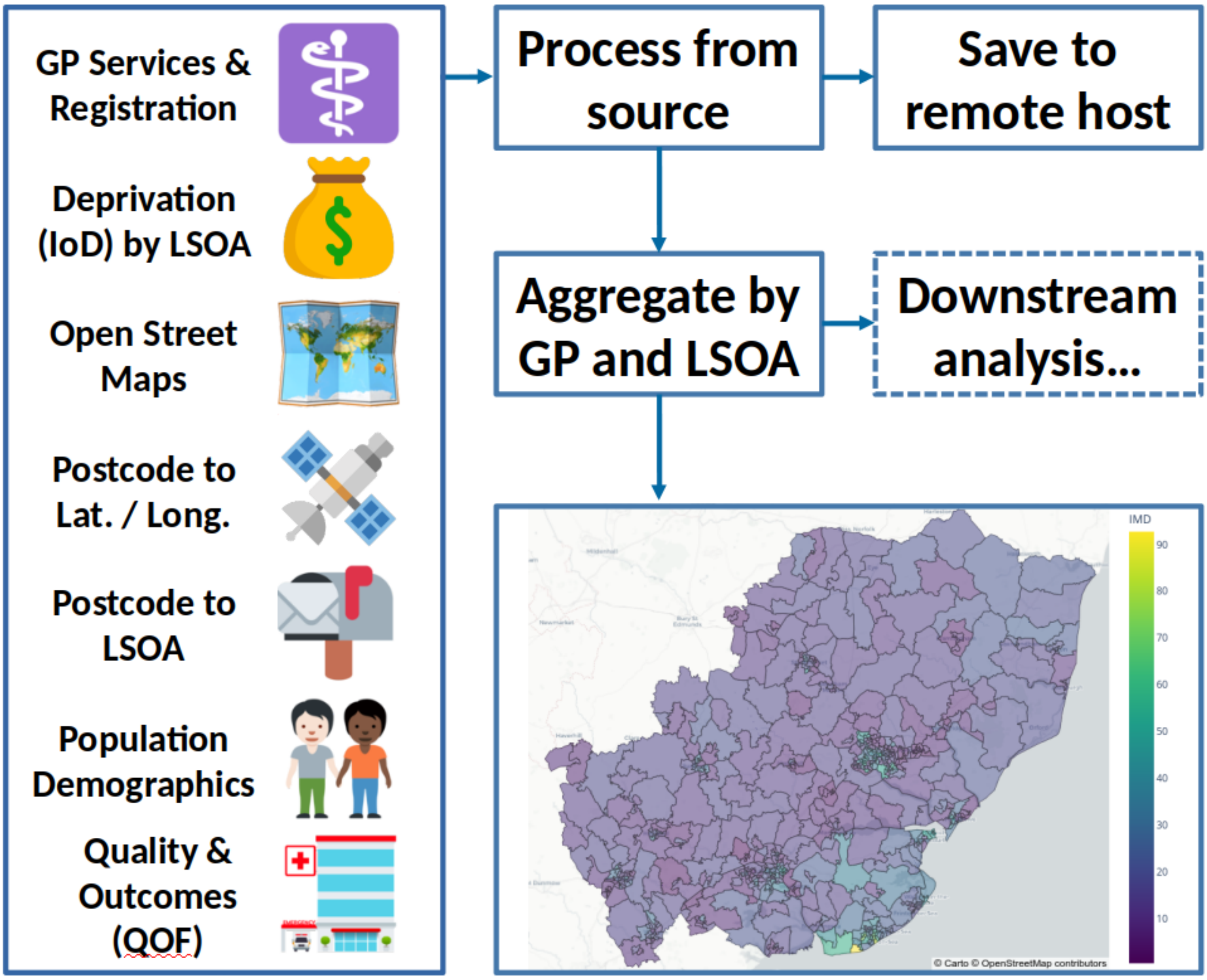

Figure 1.1: **Overview of ESNEFT Tools Workflow.** For each public data source, ESNEFT Tools downloads and processes the source data before saving a local copy on the user's system for quick retrieval later. A pre-processed copy of each dataset is also hosted on the ESNEFT Tools GitHub. Downloading from *host*, rather than *source*, is the recommended way for users to retrieve the data this is described in section 1.2 on the following page. Following data retrieval, the datasets can be aggregated to LSOA level or GP level and can be used for downstream analysis as required. ESNEFT Tools also includes some additional functions for visualisation which are more thoroughly described in the documentation.

1.1.2 Data Sources

The data sources utilised by ESNEFT Tools are described below and source links are valid at the time of writing (January 1, 2023). However, it is likely that, over time, these public sources will become unavailable or will be updated. For these reasons, it is recommended that users obtain static copies of the hosted data using the `.fromHost()` method described below. Alternatively, users may provide a YAML format file containing updated links to the data. An example is provided here and usage is described in the documentation.

Table 1.1: **ESNEFT Tools Data Sources**

Dataset Description	Source
Postcode Lookup to LSOA and Lat/Long (Nov 2022)	ONS via ArcGIS
English Indices of Deprivation by LSOA (2019)	National Statistics (.gov.uk)
LSOA Population Estimates by age and sex (mid 2020)	ONS
Proportion of Ethnicity Minorities by LSOA	NOMIS
Land Hectare Measurements by LSOA	ONS via ArcGIS
GP Registration by LSOA	NHS Digital
GP Site Information	NHS Digital
GP Practitioner Information	NHS Digital
QoF Disease Prevalence Data	NHS Digital
Geographical Boundaries by LSOA	UK Data Service
Open Street Map for ESNEFT	Geofabrik
LSOAs within ESNEFT	ESNEFT

1.2 Setup and Basic Usage

For comprehensive documentation and usage examples, please refer to the complete documentation available on GitHub.

1.2.1 Requirements

ESNEFT Tools requires Python >= 3.7 and up-to-date versions of pip and setuptools. The package is compatible with all main operating systems including Windows, macOS and Unix.

```
pip install --upgrade pip setuptools
```

1.2.2 Installation

ESNEFT Tools , and all additional dependencies can be installed as followed. For instructions on using virtual environments, to avoid dependency conflicts, refer to the documentation.

```
pip install esneft_tools
```

1.2.3 Usage

Once installed, load an interactive Python interactive environment, such as a Jupyter Notebook, and try the following commands.

Load Modules and Retrieve Data

The following code will load ESNEFT Tools to your environment and retrieve a hosted copy of each data. The datasets are also saved locally to the cache (`.data.cache`) and so online retrieval is only required the first time. The `data` object is a dictionary containing all of the datasets. This can be passed directly to the aggregation functions, as showed below, or each dataset can be used independently.

```
from esneft_tools import download, process, visualise

# Instantiate data download class.
getData = download.getData(cache='./.data-cache')

# Retrieve all data as dictionary (recommended)
data = getData.fromHost('all')
```

Summarise Data at GP Level

```
GPsummary = process.getGPsummary(**data, iod_cols='IMD')
```

Summarise Data at LSOA Level

```
LSOAsummary = process.getLSOAsummary(**data, iod_cols='IMD')
```

Visualise GPs across England

```
active_gp = GPsummary[GPsummary['Status'] == 'Active']
fig = visualise.scatterGP(active_gp, minCount=250)
```

Choropleth Map of Deprivation by LSOA

```
fig = visualise.choroplethLSOA(
        LSOAsummary, data['geoLSOA'], colour='IMD')
```

Chapter 2

Health Inequalities in Diabetes

Exploring Links with Deprivation Indicators and Diabetes Health Inequality.

Summary

Socioeconomic inequality of diabetes prevalence and access to diabetes treatment is well-documented [Barnard-Kelly and Cherñavvsky, 2020]. In both Type 1 Diabetes Mellitus (T1D) and Type 2 Diabetes Mellitus (T2D), deprived individuals are significantly less likely to receive all NICE recommended care processes. Similarly, deprived individuals are less likely to meet treatment targets for Hemoglobin A1C (HbA1C), Body Mass Index (BMI) and cholesterol. Lifestyle risk factors of T2D, including smoking, obesity and lack of physical activity are all more prevalent in deprived households. Together this can contribute to a 77% increased risk of developing T2D among the most deprived individuals compared to the least deprived [Roper et al., 2001]. Ultimately, a comprehensive understanding of the underlying factors associated with this inequality is critical for addressing these issues.

This work utilises the ESNEFT Tools utility to explore how indicators of deprivation are associated with the prevalence of diabetes across England. A paradigm is presented to explore underlying indicators of deprivation that make up the Index of Multiple Deprivation (IMD). The Education deprivation domain, and specifically the Adult Skills sub-domain, were found to be disproportionality correlated with diabetes prevalence. In addition, the Adult Skills sub-domain is itself comprised of two indicators: Adult Education and English Language Proficiency. This work identifies Adult Education deprivation, which represents the proportion of individuals with with low or no formal qualifications, as being highly correlated with diabetes prevalence. Interestingly, of all the diseases investigated, diabetes stood out as being disproportionality correlated to Education and Adult Skills deprivation.

In addition, this work also investigates patterns of deprivation associated with utilisation of the Ipswich Diabetes Centre within East Suffolk and North Essex NHS Foundation Trust (ESNEFT). As expected, individuals from more deprived Lower Layer Super Output Area (2011)s (LSOAs) were found to be over-represented in the Diabetes Centre cohort. However, patients from deprived back-

grounds were found to have significantly worse health metrics, including higher BMI, HbA1C and Urine Microalbumin upon initial referral to the centre. While this works highlights potential health inequalities associated with diabetes, it also indicates that these inequalities may be modifiable and could lead to improvements in population health.

2.1 Patterns of Deprivation and Disease in England

As discussed in section 1.1 on page 2, ESNEFT Tools includes methods for estimating LSOA level disease prevalence from the Quality and Outcomes Framework (2021-2022) (QoF). By combining disease prevalence with demographic data and the English Indices of Deprivation (2019) (IoD), it is possible to investigate how deprivation is associated with disease prevalence across England. The IMD is a commonly used metric for aggregating the multifaceted components of socio-economic deprivation into a single number. The IMD is calculated by combining numerous indicators of deprivation across seven domains: Income, Employment, Education, Health, Crime, Barriers to Housing and Serves and Environment. Despite it's simplicity, the IMD makes assumptions about the relative importance of each domain of deprivation. For example, Income and Employment are each given much higher weighting (22.5%) than Education or Health (13.5%). Moreover, two regions that have the same IMD score may have very different underlying patterns of deprivation.

To address these limitations, this work proposes a paradigm for studying the underlying domains and indicators of the IMD. This approach can uncover patterns of deprivation that may otherwise be masked by the IMD. Moreover, indicators of deprivation, such as English Language Proficiency, are inherently more interpretable than aggregate scores such as the IMD.

2.1.1 Methods

Detailed documentation, and the code used for this analysis, is available at the GitHub repository associated with this report. In brief, LSOA level data corresponding to population demographics, socio-economic deprivation and disease prevalence were downloaded, as previously described, using ESNEFT Tools . Following this, partial correlation coefficients were calculated between deprivation scores and disease prevalence scores. The partial correlation measures the degree of association between two random variables after controlling for the effect of other controlling random variables. This in particularly important when analysing deprivation domains, which are themselves frequently correlated with one another. For example, areas with high Income deprivation frequently have high Employment deprivation. The use of partial correlation enables the association of each domain to be independently assessed after the removing the confounding effects of the other deprivation domains.

Partial correlation analysis facilitated the identification of the Adult Skills sub-domain as the predominant indicator associated with diabetes prevalence. Adult Skills deprivation was correlated against LSOA-level Blood Pressure (BP) and HbA1C treatment targets, which were inferred from the QoF using ESNEFT Tools . Finally, the relationship between age (median age in LSOA) and deprivation quintile in determining diabetes prevalence was also assessed.

2.1.2 Results

Association of Deprivation and Disease

Of the seven domains of deprivation, Health deprivation was the most strongly associated with disease prevalence across the spectrum of diseases assessed (see fig. 2.1 on the following page). This observation is unsurprising since the Health deprivation domain is already a measure of premature morbidity and disability due to disease. After Health, Education deprivation was the most strongly associated with disease prevalence. This was particularly true for diseases with strong risk factors associated with lifestyle including obesity, smoking and diabetes.

The analysis was repeated using the underlying sub-domains of Education; “Adult Skills” and “Youth Skills” (see fig. 2.2 on page 9). Similar patterns were observed between the two sub-domains but lifestyle-associated diseases were more highly associated with the Adult Skills sub-domain.

Finally, the analysis was repeated using the underlying indicators of the Adult Skills sub-domain; “Adult Education” and “English Language Proficiency” (see fig. 2.3 on page 10). Adult Education corresponds to the proportion of working-age adults with low or no formal qualifications. English Language Proficiency corresponds to the proportion of working-age adults who cannot speak English or cannot speak English well. These indicators are not available through the English Indices of Deprivation (2019) but were obtained separately through Nomis. In general, English Language Proficiency was not positively associated with disease prevalence. However, Adult Education was strongly associated with the prevalence of numerous diseases, particularly lifestyle associated diseases. These results suggest that the Adult Education indicator is primarily responsible for the observed association between Educational Deprivation and disease prevalence. Interestingly, at each level of Education deprivation (Education, Adult Skills and Adult Education), diabetes was the disease most strongly associated with deprivation.

Patterns Associated with Adult Skills Deprivation

For subsequent analyses, the Adult Skills sub-domain was selected to explore relationships between deprivation and diabetes. Although the underlying Adult Education indicator was most strongly associated with diabetes, this indicator is not as readily available and is not provided in the IoD.

Figure 2.4 on page 11 reveals a significant negative correlation between the proportion of patients meeting HbA1C treatment and Adult Skills deprivation ($rho = -0.26$). No meaningful correlation was identified between BP treatment targets and deprivation ($rho = 0.04$).

Finally, fig. 2.5 on page 11 reveals an interaction between Deprivation and Age associated with diabetes prevalence. In general, the median age of the LSOA population is strongly associated with diabetes prevalence - older populations have higher prevalence. However, this association with Age is absent in the most deprived quintile (q5) - diabetes prevalence is high among both young and old populations. Despite this, among older populations, the prevalence of diabetes in the most deprived quintile is similar to other deprivation quintiles. This observation suggests that younger populations may be more affected in deprived areas. As populations age, age-associated risk factors may begin to predominate over deprivation-associated risk factors in determining diabetes risk.

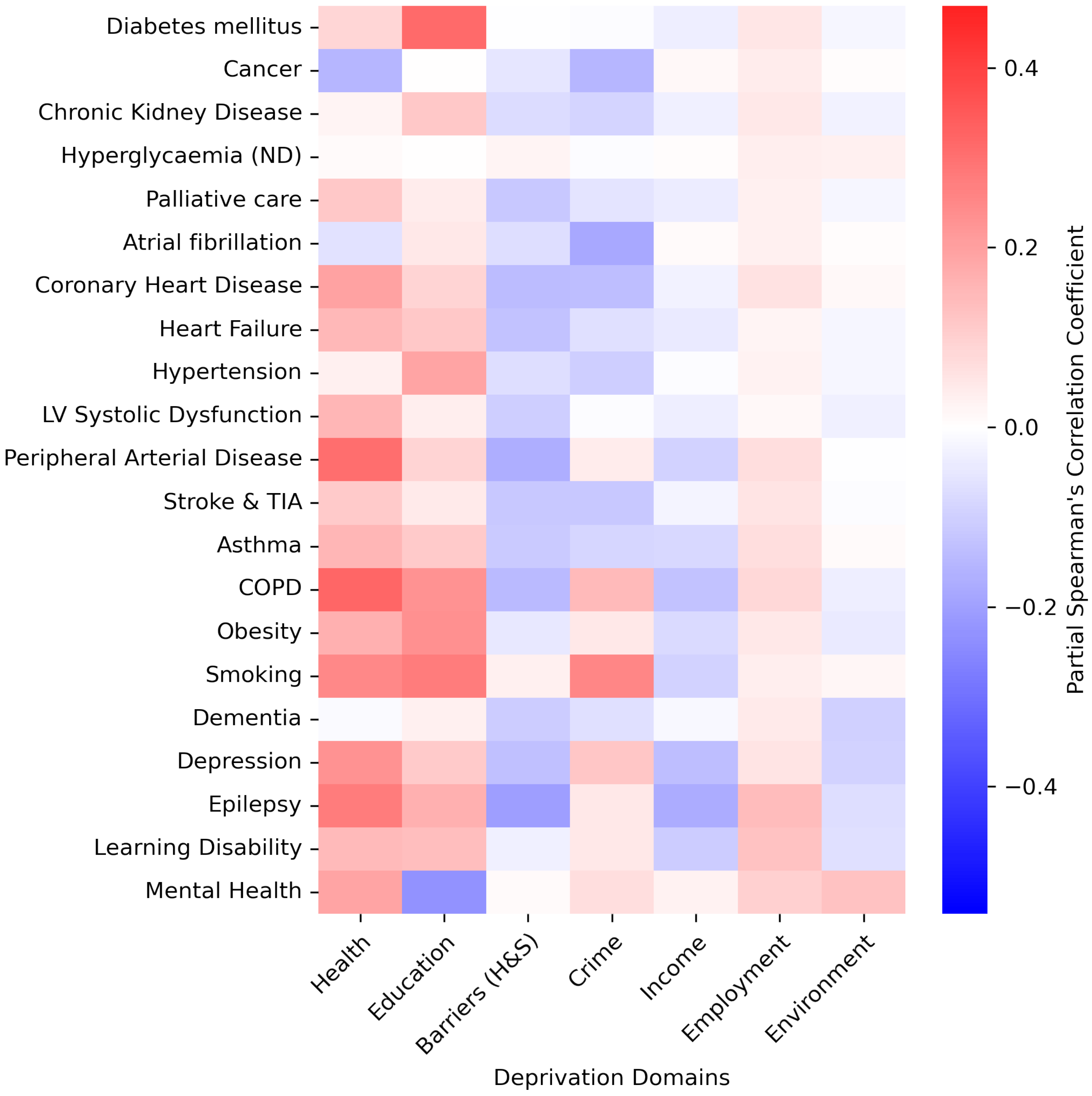

Figure 2.1: **A significant positive correlation was identified between diabetes prevalence and Education deprivation.** Heat map of partial correlation coefficients between estimated disease prevalence and each of the seven primary domains of deprivation. For each domain, partial correlation analysis was performed to correct for confounding factors including demographics (age, sex and ethnicity) as well as the other six deprivation domains.

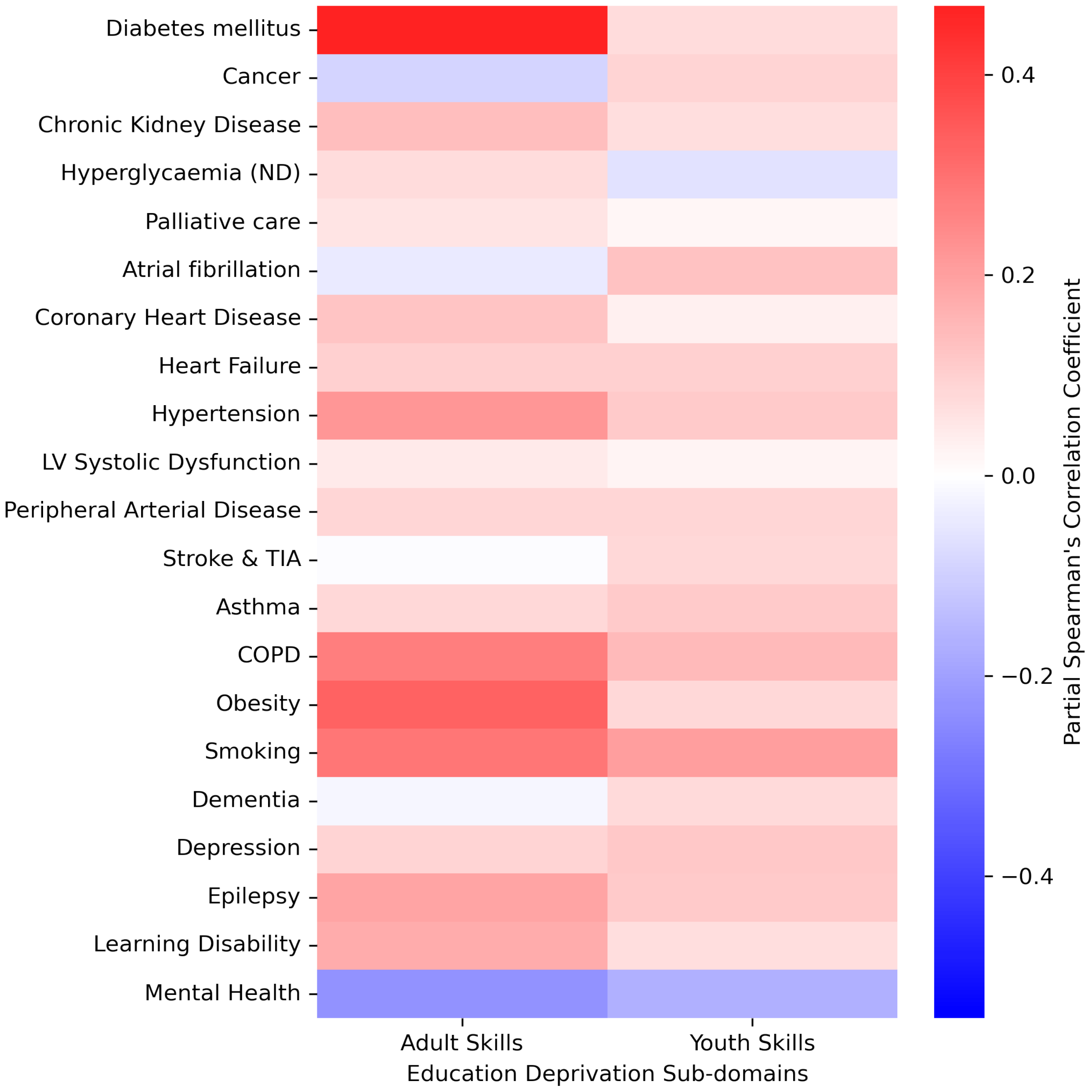

Figure 2.2: **A significant positive correlation was identified between diabetes prevalence and the Adult Skills sub-domain of Education deprivation.** Heat map of partial correlation coefficients between estimated disease prevalence and the sub-domains of Education deprivation. For each sub-domain, partial correlation analysis was performed to correct for confounding factors including demographics (age, sex and ethnicity) and the six non-Education domains of deprivation.

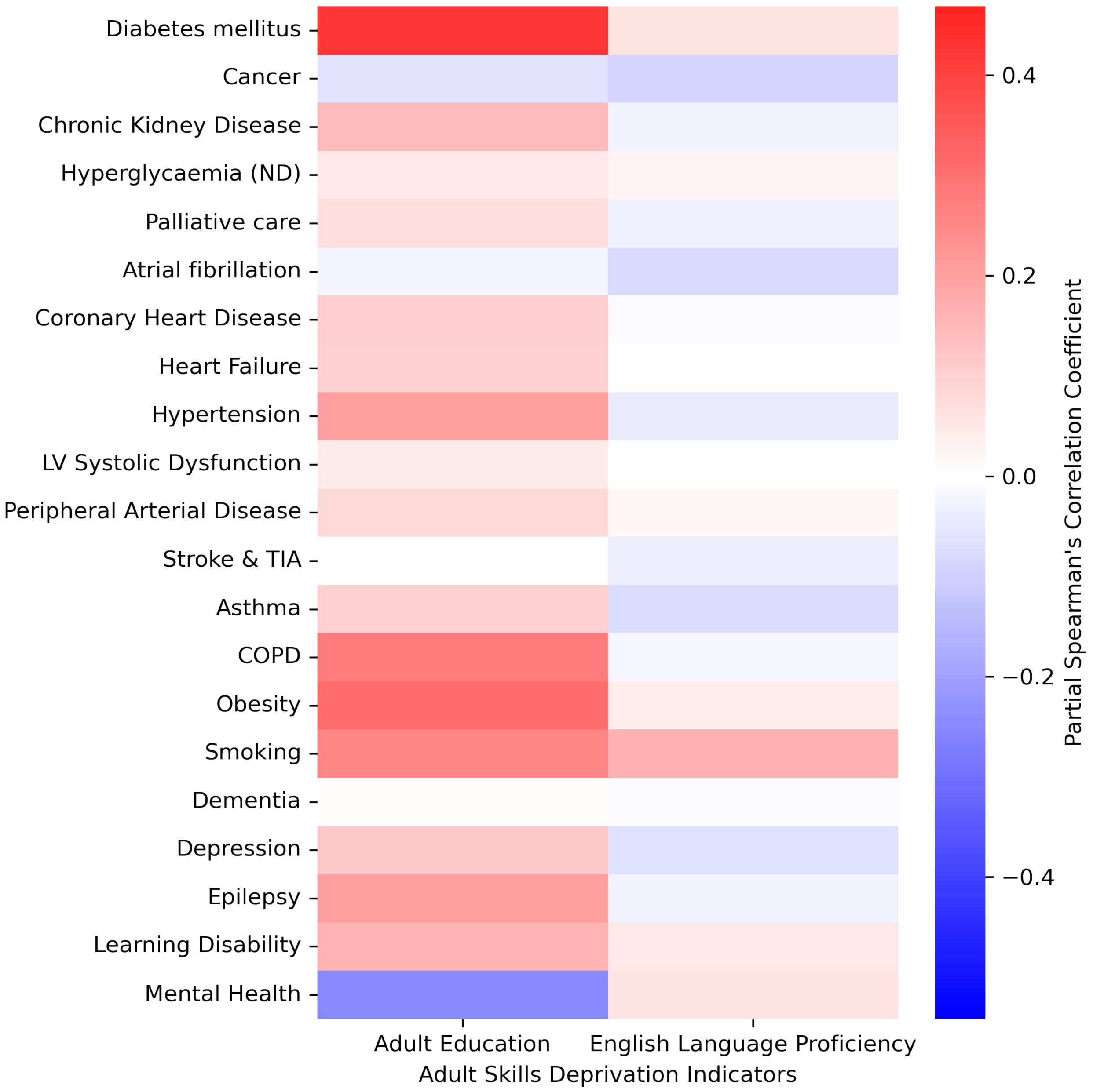

Figure 2.3: **A significant positive correlation was identified between diabetes prevalence and the Adult Education indicator of the Adult Skills deprivation sub-domain.** Heat map of partial correlation coefficients between estimated disease prevalence and the underlying indicators of the Adult Skills deprivation sub-domain. For each indicator, partial correlation analysis was performed to correct for confounding factors including demographics (age, sex and ethnicity) and the six non-Education domains of deprivation.

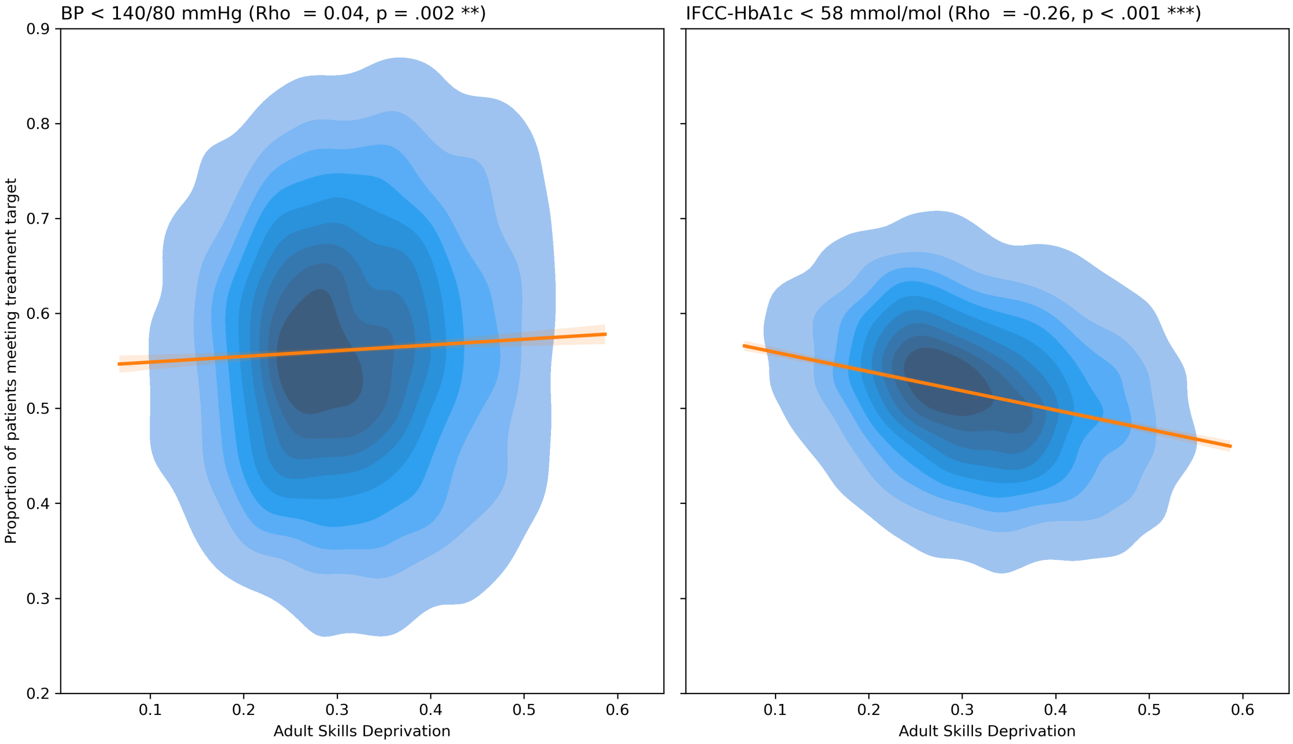

Figure 2.4: **Deprived populations are less likely to meet HbA1C treatment targets.**

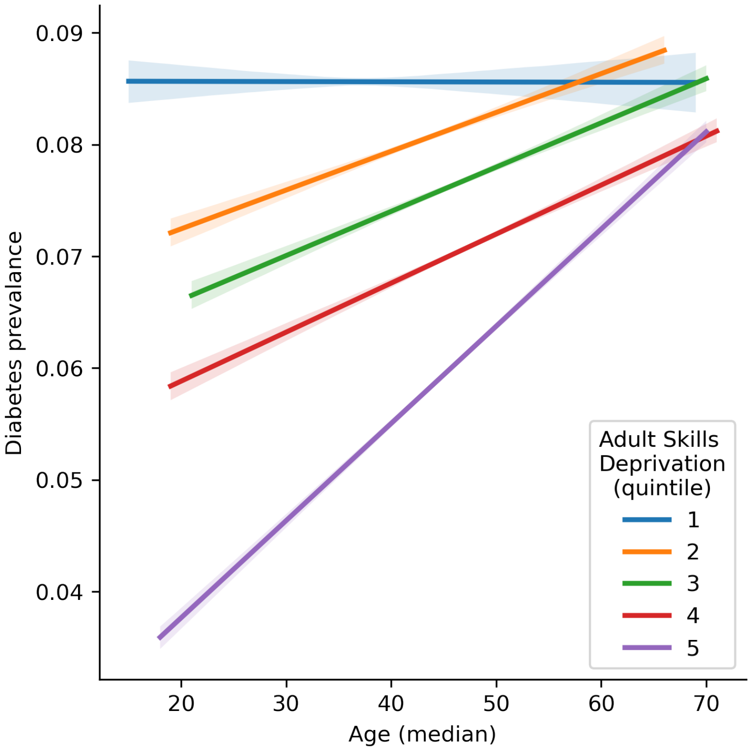

Figure 2.5: **Diabetes Prevalence by Age and Deprivation.** Diabetes prevalence, by LSOA, grouped by Adult Skills deprivation quintile and median age. Prevalence is positively associated with age, except in the most deprived areas. Underlying data points are not shown. Note: 1 = Most Deprived.

2.2 Impacts of Deprivation at the Ipswich Diabetes Centre

An analysis was performed to assess the impacts of Adult Skills deprivation with healthcare utilisation and health outcomes at the ESNEFT Ipswich Diabetes Centre.

2.2.1 Methods

In total 268,912 anonymised records, corresponding to 8521 patients, were obtained from the Ipswich Diabetes Centre between approximately 2014 and 2022. For each patient, the full history of recorded events at the Diabetes Centre were provided. Patient demographic data, which was not provided with the Diabetes Centre dataset, were obtained by joining a separate ESNEFT outpatient dataset by anonymised patient identifier. Only 5713 patients were successfully associated with demographic data and could be used for subsequent analysis. Each patient was assigned a deprivation score according to the LSOA associated with their postcode of residence. The deprivation metric selected for this analysis was the Adult Skills sub-domain of Education deprivation. An illustrative example of the Diabetes Centre data is provided in table 2.1. Each record corresponded to a single coded event. The predominant coded events studied in this analysis were BMI, HbA1C, Urine Microalbumin, Diastolic BP or Glomerular Filtration Rate (GFR). In addition, separate coded records were provided for missed appointments (Did Not Attend (DNA)) and the Date of Diagnosis.

Raw event-level records were summarised as patient-level statistics describing health outcomes and service utilisation. For each metric (e.g. BMI), the initial recorded value at the patient's fist appointment, was obtained. In addition, where more than three distinct measurements were available, the mean change between measurements and the mean time between measurements was also calculated. The frequency of missed appointments, the time between diagnosis and referral and the mean time between appointments was also calculated. Finally, a separate inpatient dataset was also joined with the Diabetes Centre data to assess differences in inpatient admissions associated with deprivation.

To assess the association of each metric with deprivation (Adult Skills), the partial correlation was computed after controlling for relevant demographics (age, sex and ethnicity (non-white)).

Table 2.1: **Diabetes Centre Data with Patient Demographics**

ID	Date	CTV3Desc	Value	Sex	Age	Ethnicity	Deprivation
1	12 Oct 2020	HbA1c	82.0	Male	48	White	34.6
1	21 Dec 2020	HbA1c	79.0	Male	48	White	34.6
1	14 Jan 2021	BMI	28.6	Male	48	White	34.6
1	18 Feb 2021	Did Not Attend	-	Male	48	White	34.6
2	01 Feb 1986	Diagnosis Date	-	Female	66	Mixed	21.5
2	03 Mar 2021	HbA1c	49.0	Female	66	Mixed	21.5
2	03 Mar 2021	BMI	33.1	Female	66	Mixed	21.5

Note: This data is fictitious for illustrative purposes. CTV3Code and Recording Unit columns have been removed from this example. Deprivation represents the raw score of the Adult Skills sub-domain of Education. Demographic data was not part of the original Diabetes Centre data and was joined from a separate outpatient dataset.

2.2.2 Results

Partial correlation analysis revealed a significant positive association with initial vitals measurements and deprivation (see fig. 2.6). Specifically, initial measurements of BMI, HbA1C and Urine Microalbumin, were significantly higher among deprived patients. Higher values of these metrics is generally associated with poorer health outcomes in diabetes. This could suggest that patients from deprived backgrounds are being referred to the Diabetes Centre later. Since, there was no association with deprivation and time to referral since diagnosis it is possible that late referral is attributable to a delay in diagnosis. Despite these finding, there was no significant association between deprivation and patient improvements and service utilisation. This may indicate that the services of the Diabetes Centre are equally beneficial to all patients following referral.

Figure 2.7 on the next page shows the absolute values of the significant associations broken down by deprivation quintile. Upon initial referral, patients from the most deprived quintile were found to have an average BMI $2kg/m^2$ higher than the least deprived quintile. Similarly, patients from the most deprived quintile were found to have an average HbA1C of $6mmol/mol$ higher than the least deprived quintile. Total inpatient admissions, measured using data Colchester and Ipswich hospital from between 01 April 2019 and 31 March 2021 were also higher among deprived patients. This is indicative of poorer health but is not directly related to the Diabetes Centre.

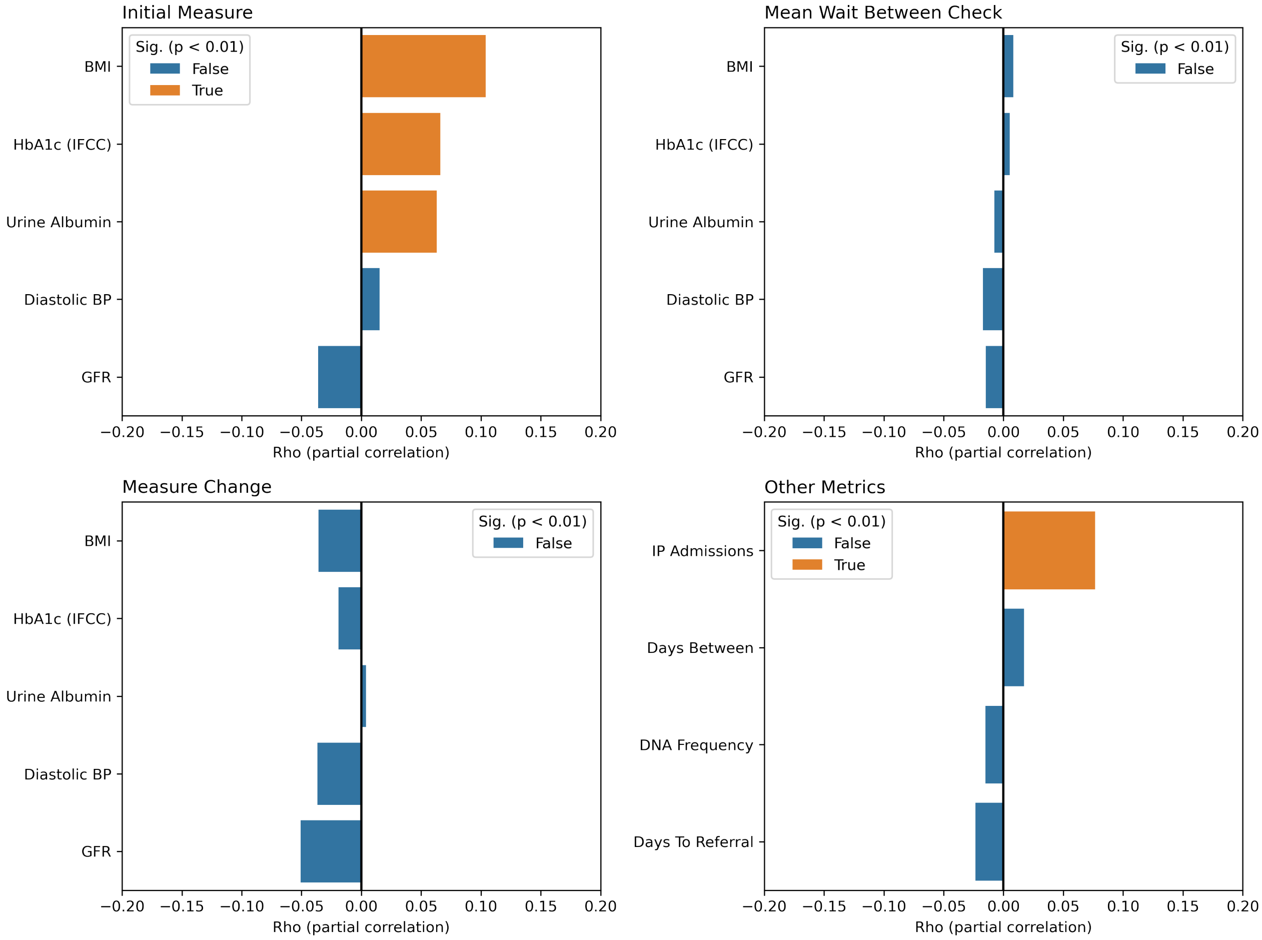

Figure 2.6: **Healthcare Utilisation and Deprivation at the Ipswich Diabetes Centre.**

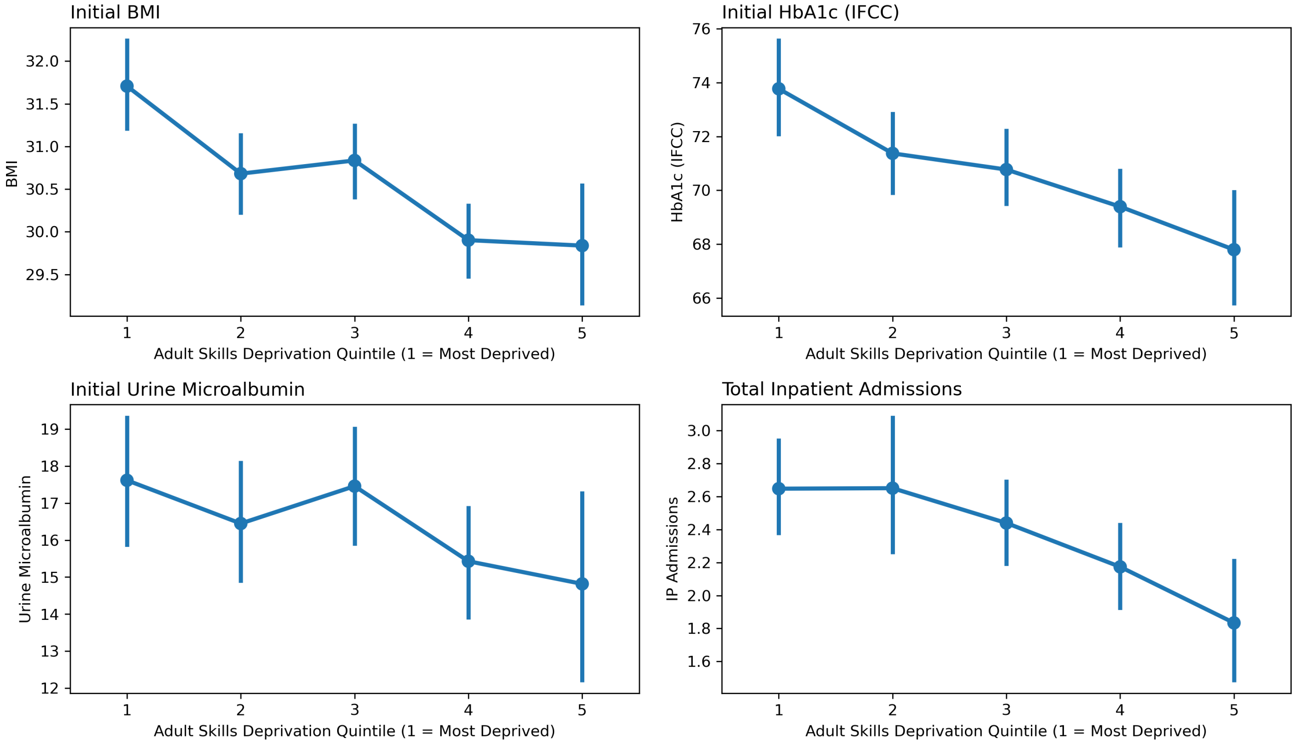

Figure 2.7: **Diabetes Centre Metrics by Deprivation Quintile** Note: 1 = Most Deprived.

2.3 Discussion

This work utilises publicly available datasets of disease prevalence and deprivation, across England, to identify sub-domains of deprivation associated with disease. Adult Skills deprivation, and specifically Adult Education deprivation, was found to be strongly associated with the prevalence of diabetes mellitus at LSOA level. Although Adult Education deprivation is also associated with a variety of other diseases, the association with diabetes was particularly strong. This indicates that deprivation associated health inequality may be worse in diabetes compared to other diseases. It may also suggest that public awareness of diabetes lags behind other diseases with strong lifestyle-associated risk factors.

This work confirms previous findings indicating that deprivation is associated with the attainment of treatment targets [Barnard-Kelly and Cherñavvsky, 2020]. Specifically, Adult Skills deprivation was negatively associated with the proportion of diabetes patients meeting HbA1C treatment targets. However, no substantive associations were found between deprivation and the attainment of BP treatment targets. These results could be informative for determining how inequality could be more effectively addressed. For example, improving HbA1C treatment targets may be more effective at addressing inequality than improving BP treatment targets. Figure 2.5 on page 11 suggests that risk factors associated with deprivation are more predominant among younger populations. Among older populations, the gap in diabetes prevalence between the most and least deprived quintiles was small, suggesting that age is the predominant risk factor. This suggests that deprivation association health inequality in diabetes may be greater among younger populations.

The analysis of the Ipswich Diabetes Centre patients revealed no significant association with deprivation following referral. Specifically, deprivation was not associated with differences in attendance frequency or differences in health outcomes. However, this work revealed that patients from deprived backgrounds had worse initial vital sign measurements at the point of referral. This finding indicates that, while service provision at the Diabetes Centre is equitable, deprivation-associated inequalities may influence initial patient referral. A key next step for this work would be to investigate the sources of referral and assess differences in patient history leading up to referral. It is interesting to note that patients from deprived background were also more likely to have recorded inpatient admissions. For example, it is possible that some patients were referred to the Diabetes Centre, following hospital admission, due to undiagnosed complications of diabetes.

2.3.1 Caveats and Limitations

Public Data

All of the aforementioned results were determined from publicly available data. Public data has large *breadth*; analyses can utilise a wide range of multiple sources of information that describe a population. However, public data has low *depth*. For example, the data used is this analysis is aggregated to LSOA level. This is sufficient to study nationwide trends but insufficient to draw robust conclusions at an individual level. In addition, disease prevalence by LSOA is estimated using General Practice (GP) level prevalence data extracted from the QoF. This approach is inherently imprecise and direct measurements of LSOA level disease prevalence would be preferable but are not publicly available. Finally, the public data utilised by ESNEFT Tools were not all collected at the same points. It is reasonable to assume that population demographics are unlikely to change substantively within the space of a few years. However, with the release of the Census 2021 data, it may eventually be appropriate to update the relevant public sources as required. In addition, a future analysis may seek to repeat this work across multiple time periods. Specifically this analysis might seek to establish whether certain public health interventions can be reliably associated with reductions in deprivation-association health inequality.

Statistical Considerations of the IoD

The IoD is relative, rather than absolute, measures of deprivation. Put simply, they are sufficient to rank areas according to their deprivation but cannot be used to quantify the extent to which one area is more deprived than another. For example, consider area A has a deprivation score of 20 and area B has a deprivation score of 40. We *can* say that area B is more deprived than area A. However, we *cannot* say that area B is twice as deprived as area A. From a statistical perspective the IoD are therefore ordinal, rather than continuous, variables. As such, non-parametric tests, such as the partial Spearman's rank correlation utilised in this work, should be utilised when working with raw deprivation scores. Similarly, when building machine learning models, decision-tree based methods may be preferable over linear / logistic regression or neural networks.

Interpretation of the IoD

The IoD are small area statistics; they cannot identify deprived individuals. Deprived individuals may live in non-deprived areas and affluent individuals may live in deprived areas. As such, the patterns identified in this work cannot distinguish between the *area* and the *individual*. Health inequality may arise from components of deprivation tied to the individual's area or they may arise from circumstances specific to that individual. For example, assume an affluent person moves to a deprived area. Will they be affected by deprivation associated health inequalities? If those inequalities are directly associated with the area, for example due to less healthcare resources, then they will be affected. However, if those inequalities arise from individual circumstance, for example poverty or unemployment, then that individual may be insulated from them.

This work, and any study that utilises the IoD as a measure of deprivation, cannot distinguish between these sources of deprivation. The observed differences in initial referral measurements at the Ipswich Diabetes Centre could be attributable to lack of patient awareness. On they other hand they may be due to insufficient healthcare resourcing within those areas. Identifying the underlying source of health inequality is vital for determining the course of action to address it. Area associated inequalities may require targeted interventions, such as health awareness campaigns, to a specific area. However, individual associated inequalities will required interventions to *all* high-risk individuals irrespective of whether they live in a deprived or affluent area. Future work must be done to obtain individual patient-level assessments of socio-economic deprivation.

Scope of the Analysis

A key limitation of Ipswich Diabetes Centre dataset was that it contained no additional metadata relating to patient history or the underlying reasons of referral. In addition, patient attendance was sporadic and the total number of attendances was highly variable between patients. Given the limited nature of the dataset, a simple analytical approach was selected to minimise the number of assumptions. Specifically three quantitative metrics for each vital health measurement were selected. The ”Initial Measure” corresponds to the earliest recorded value of each health metric. This value is assumed to correspond to the patient's health status at the point of referral and is informative for all patients, irrespective of how frequently they subsequently utilised the service. A caveat of this is that some patients may have been referred from a different Diabetes Centre and so the true initial recorded value is unknown. The average number of days between vital sign measurements was recorded for all patients with atleast three recordings. Finally, the difference between the first and final observed measurements was used to represent the change in health vitals over the patient's timeline. This was recorded for patients with atleast 90 days between these measurements. Although a simple linear regression may have been more appropriate to capture these changes over time, most patients had far too few data points to fit a reasonable model. Despite the evident limitations of these metrics, the ”Initial Measure” remains the most robust and is independent of confounding factors such as appointment frequency and service utilisation.

The results from this work have revealed putative inequalities in the referral process to the Ipswich Diabetes Centre. While much needs to be done to further investigate this, it is hoped that this work will serve as the foundation for a more thorough examination of healthcare pathways leading up to referral that may explain these inequalities.

Chapter 3

Packaging in Python

Exploring Best Practise Approaches in Python Packaging.

Summary

Any Python project can be packaged to facilitate reproducibility and distribution. Appropriate use of packaging can simplify testing and development and encourage good coding practises. Indeed, the wider application of software engineering best practises to data science becomes increasingly important as projects grow larger and more complex. This work has made extensive use of Python packaging to manage distinct coding projects. In doing so it has explored best practise approaches for managing and packaging data science coding libraries and workflows. This short chapter will briefly discuss the putative best practise approaches that were explored throughout this project and how these might be applied and incorporated into future projects.

3.1 Rationale

Packaging provides many of the same benefits as writing functions. In the same way that a function encapsulates a small piece of code for reuse and distribution, packages enable larger collections of functions and code to be reused and distributed together. As such, packaging code inherently encourages good coding development and collaboration between distinct projects. For example, assume an organisation maintains a set of functions for generating consistent and compliant visualisations. If those functions are packaged then users can easily load the functions into their workflows. This approaches simplifies project workflows and, importantly, maintains a single copy of the function's source code. As such, updates to the source code, which may reflect bug fixes or changes to the template, will carry forward to all projects. Without packaging, users must manually copy their function in each script or notebook. This adds code clutter, reducing readability and results in code duplication. As such the functions cannot easily be maintained which will likely introduce bugs or inconsistencies between projects.

3.2 Packaging Guidelines

The Python community have developed many excellent open-source Python project managers including tools such as Hatch and Poetry. Users may wish to consider using these more fully-featured tools as their packages become more sophisticated and particularly if their packages have more complicated dependencies. This work has sought to build minimal However, for users wishing to build lightweight packages, perhaps to facilitate the distribution of a few helpful functions or a simple command-line tool, simpler packaging systems may suffice. This project sought to identify and apply the minimal viable requirements for building a functional package. Specifically, this work follows the official Python guidelines for Packaging available here. The guidelines recommended using a single `pyproject.toml` file to describe package parameters and installation requirements.

In addition, this work employs an additional best-practise for embedding the `__version__` attribute in a Python package. As of Python 3.8, it is possible to extract package metadata, including version, using the built-in `importlib.metadata` library. Given a package called `my_package`, the following code can be included in the `__init__.py` file.

```
from importlib.metadata import version
__version__ = version(my_package)
```

This will expose the `my_package.__version__` attribute, which is a standard convention, and ensures that the version only needs to be recorded in the `pyproject.toml` file.

3.3 Dependencies

In software development a dependency refers to any other coding library that is utilised by the software in development. In Python, dependencies that are not part of the Standard Library must be documented in `pyproject.toml` file. This ensures that those dependencies are also installed along with the Python package of interest. Automatic dependency management is a key advantage of packaging code. It can greatly improve software accessibility, in particular to users that may otherwise struggle to manually install the required dependencies. However, defining the appropriate dependencies and, in particular, defining which *versions* of each dependency are compatible with a package is non-trivial.

One solution is to identify a set of versions compatible with the package and require the package installs those versions only; this is known as dependency pinning. The primary advantage of dependency pinning is that it enables consistent reproduction of a python environment. This ensures that analyses are reproducible and that the package does not become unstable due to deprecation of features within dependencies. On the other hand, dependencies often release updates for good reason, perhaps to improve functionality, fix bugs or patch security issues. Users who pin dependencies will not benefit from these improvements. In addition, it could be argued that a piece of analysis relying on deprecated functionality should not be trusted and the results should be reviewed. Dependency pinning also constrains the generalisability of a Python package and is unlikely to account for all use cases. If users wish to incorporate the package into their own software then they will likely run in to incompatibility issues. As such, strict dependency pinning is not usually a suitable option for developing Python packages. Dependencies should be as dynamic as possible and version constraints should only be specified if they are known to be required. For example, perhaps the package utilises a particular NumPy function only introduced in version 1.2.0. In this case the version constraint would be defined as `numpy >= 1.2.0`.

Despite this, users may reasonably wish to document a record of their Python environment to enable reproduction of a specific piece of analysis. A list of pinned dependencies can be provided in a `requirements.txt` file that is included in the GitHub repository but is independent of the package's installation dependencies. In addition to documenting pinned versions of each dependency the `requirements.txt` file should also include a static version of the package itself. This ensures the Python environment is consistently reproduced even if updates are made to the package itself. A version of the package tied to a specific Git hash (e.g. 6f03bdd) can be installed directly from GitHub by adding the following line to the `requirements.txt` file.

```
git+https://github.com/nhsx/my_package@6f03bdd#egg=my_package
```

3.4 Installation Tests

A simple suite of tests should be deployed to assess whether the relevant package can be installed, imported and run on a variety of systems. The purpose is to identify installation issues and conflicts that may arise when using the package on different operating systems or with different python versions. Regular automated testing also ensures that dependency depreciation issues, that may arise when using dynamic dependencies, can be quickly identified and addressed.

As an example, the DNAttend package presented in this work was designed as an autoML workflow to be applied to various different datasets. DNAttend includes a function a simulating some generic input data. The simulated data is used to assess whether consistent results are produced across different installation environments. Specifically, the simulated data is passed to the workflow and a hash of the final output table is compared against an expected hash. This test is conducted across different systems and Python versions using GitHub actions to verify consistency and functionality. Note that these tests are only designed to assess installations and reproducibility. Separate tests may be required to validate that the observed output is correct. However, the user should not be expected to write unit tests for functions of established libraries such as scikit-learn.

3.5 General Guidelines

The following list describes a non-exhaustive and subjective set of guidelines that an organisation specialising in data analysis may wish to consider when packaging and managing its coding libraries. It is a summary of the ideas explored in this chapter and of approaches considered and adopted when developing packages throughout this project.

Use project templates Project templates (e.g. Cookiecutter) can be employed to vastly simplify and standardise the process of packaging and documentation. For example, simple tests with GitHub actions (see below) can be automatically configured.

Keep package dependencies dynamic Where possible, package dependences should be flexible to avoid incompatibility.

Perform regular tests with GitHub actions A package test may be as simple as checking that it can be installed and imported. Use GitHub actions to run tests using different versions of Python on both Windows and Unix/MacOS. Tests should be scheduled to run every time the code is modified and at least once a week to check for dependency deprecation. The test result should be automatically displayed on the README.

Maintain a requirements.txt file A requirements file is distinct from package dependencies and describes an exact virtual Python environment. It is used primarily to document, and if necessary reproduce, the environment that was used to generate the output of a specific analysis.

Don't package everything! Only package work that you reasonably believe may be used by others. A specific piece of analysis tied to a specific dataset is unlikely be something that can be generalised or reused beyond the purpose for which it was originally designed.

Chapter 4

Experimental: MultiNet and DNAttend

Summary

This short chapter introduces two experimental Python packages built to provide a user-friendly interface with specific methods of healthcare data analysis. The first, DNAttend, is an AutoML framework for building a model to predict patient non-attendance. The second, MultiNet, is a command-line utility for generating multi-morbidity networks and discovering significant associations between pairs of diseases.

Source code: GitHub: DNAttend
Source code: GitHub: MultiNet

4.1 DNAttend

Patient non-attendance is a significant and pervasive burden to healthcare systems. NHS England records approximately 5% of all primary healthcare appointments as Did Not Attend (DNA). In addition to putting patients themselves at risk, DNAs reduce clinical capacity and are estimated to cost the National Health Service (NHS) as much as 216 million annually [Margham et al., 2021]. Supervised machine learning approaches may be capable of identifying hidden patterns that determine an individual's risk of not attending. Accurate prediction of relative DNA risk may facilitate targeted intervention approaches for the most at-risk patients. However, no single model will be effective for all healthcare settings. Tailor made models may be required for a particular hospital, department or General Practice (GP) practise. As such, if the feasibility of these approaches is to be explored, machine learning workflows must be developed that are accessible and simple to train on a diverse range of datasets. This work introduces DNAttend, a user-friendly command-line utility for training and evaluating models to predict patient non-attendance for healthcare appointment data. These models can identify which features most strongly determine the probability of non-attendance and may help inform where to allocate resources for preventing DNAs.

4.1.1 Overview

Patient non-attendance can be treated as a binary classification problem. The objective is to predict is a patient will or will not attend based on a set of input features. The choice of input features will largely depend on the available data. In addition, the predictive power of any particular feature may vary across healthcare settings. Features may include, among others, appointment time, patient age or consultation media. However, DNAttend was built independently of any particular input data. Instead it is intended as a generalised autoML framework for simplifying the training and evaluation of predictive models. This allows other users to focus on proper input data curation and feature selection for their particular dataset.

A simplified overview of the DNAttend workflow is provided in fig. 4.1 on the next page. The DNAttend workflow is configured from a single configuration file in JSON format. In brief, the input data is split into a training, validating and testing dataset. Two models are built; a baseline model utilising logistic regression and a gradient boosted decision tree model utilising CatBoost. The CatBoost model hyper-parameters are iteratively tuned, via randomised search, and the validation dataset is used to identify the optimal number of boosting rounds without over-fit ting. Finally, both models are evaluated against the test dataset and the user is provided with multiple test metrics and comparisons of each model. Following selection of the chosen model and set of features, the user can retrain a final model using the full input dataset.

4.1.2 Guidelines and Limitations

DNAttend is publicly available and extensive documentation, with examples using simulated data, is available at it's GitHub repository. While DNAttend serves to automate the technical aspects of model training and evaluation, the user is responsible for ensuring the input data is of high quality. The performance of any supervised machine learning model is arguably more dependent on the quality

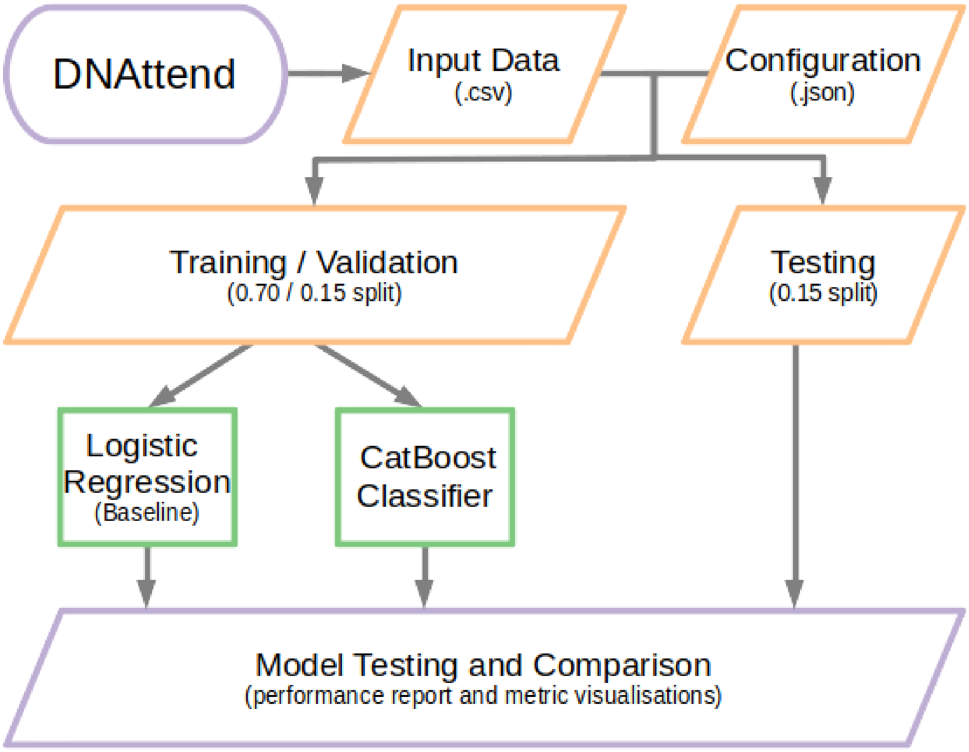

Figure 4.1: **Simplified Overview of DNAttend Workflow.** A thorough description of the DNAttend workflow and functionality is provided in the documentation.

and validity of the input data than of the specific choice of statistical algorithm. Poor quality input data will often produce functional and performant models that may otherwise fail when applied in real-world settings.

Users seeking to build models using DNAttend should be aware that appointment records are often updated emphfollowing an appointment. In particular, some fields may be modified according to whether or the patient did or did not attend their appointment. For example, the consultation media of an appointment (e.g. phone or face-to-face) may be removed if the patient did not attend. This can be highly problematic as information on the patient's attendance status is ”leaked” to other training features. Specifically, machine learning models may associate a missing consultation media field with patient non-attendance. Although such a model would appear to perform very well on the test dataset, it would likely perform poorly in a clinical setting. As such, the user should ensure that the input training and testing data accurately represents appointment information from prior to the appointment.

4.2 MultiNet

Multi-morbidity represents the co-occurrence of two of more chronic medical conditions. A substantial proportion of patients have multi-morbidities and the associated costs of treating such patients are often significantly higher. Multi-morbidities can be studied using network analysis to identify significant disease associations and pathways. The methodology of multi-morbidity network analysis is well established in the research literature [Aguado et al., 2020]. However, few user-friendly tools exist that implement these approaches for use by healthcare researchers. This work introduces MultiNet, a user-friendly command-line utility for generating multi-morbidity networks and discovering significant associations between pairs of diseases.

4.2.1 Overview

MultiNet was written as an experimental multi-morbidity network analysis utility. Given a set of patients and their associated multi-morbidities, MultiNet considers all pairwise pairwise combinations of diseases and computes an odds ratio to quantify the strength of association between them. An odds ratio of 1 indicates that the disease are independent, i.e. there is no evidence that the presence of one disease is predictive of the presence of the other. An odds ratio greater than 1 indicates positive correlation. Specifically, a pair of diseases are present in the same individual more frequently than expected by chance, given the respective prevalences of each disease. The two diseases may then be joined as an edge in a network if the odds ratio exceeds a user-defined threshold (e.g $OR >= 1.2$). The weight of the edge is defined by the *inverse* of the odds-ratio such that strong associations are closer to one-another in the network. If patient demographics are provided then the user can opt to compute a stratified odds-ratio using the Mantel-Haenszel method. Stratification corrects for confounding associations, such as age, but requires discrete categories such as age groups. In most cases age is likely to be the predominant confounding factor. Additional stratification demographics, for example ethnicity, can be provided. However, every additional factor reduces the sample size and may result in less stable statistical estimates.

MultiNet is capable of generating directed networks if the date of diagnosis is provided. For each pair of diseases a z-test is performed to assess whether one disease disproportionality occurs before the other. For example, given two diseases A and B observed in 1000 patients. In 800 patients A occurs before B, and in 200 patients B occurs before A. A two-sided proportions z-test is used to estimate the probability of this assuming no temporal association. If that probability is less than the user-defined value of alpha (e.g. $p <= 0.01$) then those two diseases are connected with a directed edge.

4.2.2 Guidelines and Limitations

MultiNet is publicly available and extensive documentation, with examples using simulated data, is available at it's GitHub repository. The user should carefully consider the limitations of network analysis prior to analysing and interpreting the results. Input data should contain one record per patient; if multiple records are available then the most recent record may be preferable. Stratification by age group is advisable if those data are available. The optimal choice of age groupings may vary according to the dataset of interest. However, Geifman et al., 2013 provides an excellent review of approaches to defining age groupings in the context of studying disease.

Diseases are often be defined according to a coding system such as the International Classification of Diseases (ICD-10) The ICD-10 classifications are hierarchical in nature and the user may wish to aggregate related classification to higher level classifications. Low level, highly specific ICD-10 classifications may be poorly represented in smaller datasets and insufficient sample sizes can lead to unstable estimates of the odds-ratio ratio. Grouping low-level classifications to higher levels will increase sample size and improve the power of statistical estimates at the cost of less specificity.

Finally, when establishing temporal associations it is important to consider that the observed date of diagnosis does not necessarily correspond to the true date of disease onset. Diseases such as hypertension are highly monitored diseases that are routinely checked within primary healthcare. In contrast, other diseases may be associated with delayed diagnosis. This means that the order of diagnosis may not be reliable indicator of order of onset. As such, interpretations of results obtained from MultiNet, or similar analytical tools, should always be reviewed with healthcare professionals with domain specific expertise.

Bibliography

Alba Aguado, Ferran Moratalla-Navarro, Flora López-Simarro, and Victor Moreno. MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity. *Scientific Reports*, 10(1), 2020. ISSN 20452322. doi: 10.1038/s41598-020-59336-1.

Katharine D. Barnard-Kelly and Daniel Cherñavvsky. Social Inequality and Diabetes: A Commentary, 2020. ISSN 18696961.

Nophar Geifman, Raphael Cohen, and Eitan Rubin. Redefining meaningful age groups in the context of disease. *Age*, 35(6), 2013. ISSN 01619152. doi: 10.1007/s11357-013-9510-6.

Tom Margham, Crystal Williams, Jack Steadman, and Sally Hull. Reducing missed appointments in general practice: Evaluation of a quality improvement programme in East London. *British Journal of General Practice*, 71(702), 2021. ISSN 14785242. doi: 10.3399/bjgp20X713909.

N. A. Roper, R. W. Bilous, W. F. Kelly, N. C. Unwin, and V. M. Connolly. Excess mortality in a population with diabetes and the impact of material deprivation: Longitudinal, population based study. *British Medical Journal*, 322(7299), 2001. ISSN 09598146. doi: 10.1136/bmj.322.7299.1389.