Evaluating risk factors for common childhood comorbidities: An All-Wales longitudinal cohort study.

William Midgley

Applied Medical Sciences

Dr Pramodh Vallabhaneni

&

Dr Arron Lacey

2022

Abstract word count: 300 words

Main body word count: 7495 words

Declaration of originality

I certify that this work is original in its entirety and has not been submitted previously for any form of assessment.

The practical work, data analysis and presentation and written work presented are all my own unless otherwise stated.

Signed:

Abstract

Childhood obesity (CO) is a pandemic. The USA&UK are predicted to reach the highest levels of obesity in Europe & USA over the next few decades. CO is a particular concern in Wales; rates being highest of the UK. CO is correlated to unprecedented increases in paediatric metabolic associated fatty liver disease (MAFLD) and type-two diabetes (T2DM). Both produce complications faster in children, and MAFLD in particular, is frequently undetected. CO has also been linked to deprivation, constipation and depression. I aim to investigate the possibility of developing a machine learning model to predict T2DM, MAFLD, constipation and depression, using predictors: obesity, deprivation, year-of-birth, and gender. I linked routinely collected children's general practitioner records with BMI scores in the SAIL databank, to the Welsh Index of Multiple Deprivation using the dataset processing language SQL, before exporting them into the statistical programming language R. I performed statistical tests (unpaired Wilcoxon and chi-squared) to identify relationships between the predictors and comorbidities in 2yrs-6yrs, 6yrs-11yrs, & 11-16yrs; then logistic regression to predict comorbidity risks within and between the three age groups. Children with at least one comorbidity in the eldest group had significantly increased BMI and deprivation scores compared with their peers. My model accurately classified 11-to-16-year-old children with T2DM and MAFLD, proving an excellent predictor of both (AUCs=0.91 & 0.88 respectively); also predicting T2DM, MAFLD, and depression in 11-16yr-olds using their BMI taken 6-11yrs (AUCs=0.95, 0.86 & 0.70); but failed to classify children with constipation (AUC=0.53-0.59 depending on group). Introduction of further risk factors, for example: ethnicity, family history or clinical observations; and a longitudinal retrospective study of older patients with obesity related and inter-related comorbidities, may identify further predictors to improve the model's accuracy further, making it useful in diagnosis and a valuable a population screening tool, to target costly interventions.

Keywords: obesity, overweight, paediatric, BMI, diabetes, T2DM, fatty liver, MAFLD, NAFLD, constipation, depression, mental health, machine learning, logistic regression, big data, SAIL, WIMD, Wales, Welsh, general practitioner, GP

Contents

Acknowledgements

I would like to thank my co-supervisors Dr Pramodh Vallabhaneni for helping me identify significant current health concerns in paediatric medicine and for his support and encouragement throughout the project and Dr Arron Lacey for his for his encouragement and support in using SAIL, as well as submitting the SAIL application. I would also like to thank Hamed Ghanbarialadolat, Ian Farr, Kevin Williams, Huw Collins, Carys Jones, and Hywel Turner Evans, among the rest of the staff at SAIL for their patience and support as I learned to use the database.

This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. I would like to acknowledge all the data providers who make anonymised data available for research.

Abbreviations

AUC – Area Under ROC Curve
BMI – Body Mass Index (weight in kilograms divided by the square of height in metres)
cf. – *confer* (compared with)
CMPW – Child Measurement Programme for Wales
CV – Crossfold-Validation
FN – False Negative
FP – False Positive
GP – General Practitioner
ID – Identification
LgR – Logistic Regression
MAFLD (formally NAFLD) - Metabolic Associated (Non-Alcoholic) Fatty Liver Disease
N.B. – *Nota Bene* (note well)
NPV – Negative Predictive Value
NHS – National Health Service (UK)
PPV – Positive Predictive Value
R – a statistical programming language
ROC – Receiver Operating Characteristic
SAIL – The Secure Anonymised Information Linkage databank
SQL – Structured Query Language
TN – True Negative
TP – True Positive
T1DM – Type 1 Diabetes Mellitus
T2DM – Type 2 Diabetes Mellitus
UK – United Kingdom
WIMD – The Welsh Index of Multiple Deprivation
Yr(s) – Year(s)

1) Introduction

1.1) Obesity and deprivation

Obesity in children has emerged as a serious public health concern in many places across the world, being described as a childhood obesity pandemic (Storz, 2020). Although the rate at which incidence is increasing is now slowing down, rates are expected to continue to rise in Europe and the USA for the next decade or so, with the UK and USA being predicted to reach the highest (Janssen et al., 2020). Childhood obesity is a particular concern in Wales; with rates being higher in Wales than in the whole of the rest of UK. Public Health Wales Child Measurement Programme for Wales (PHW, 2022) latest figures from 2018/19 demonstrated that children in Wales are more likely to be obese (26.9%), than children living in England (22.6%) and in Scotland (22.4%).

UK government statistics (UK Parliament, 2022) found that 14.4% of 4-5-year-old children are obese with a further 13.3% overweight. At age 10-11 25.5% are obese and 15.4% overweight (data from 2020/21 gathered as part of the National Child Measurement Programme, NHS Digital, 2021). These figures are a large increase on the previous year (2019/20), when they found 9.9% of 4-5-year-olds and 21% of 10–11-year-olds were obese. therefore, it is clear that the rates of childhood obesity in Wales are both a significant and a growing concern.

The Public Health Wales Child Measurement Programme latest figures from 2018/19 also confirms that children living in areas of deprivation are significantly more likely to be obese, and this gap in prevalence is increasing.

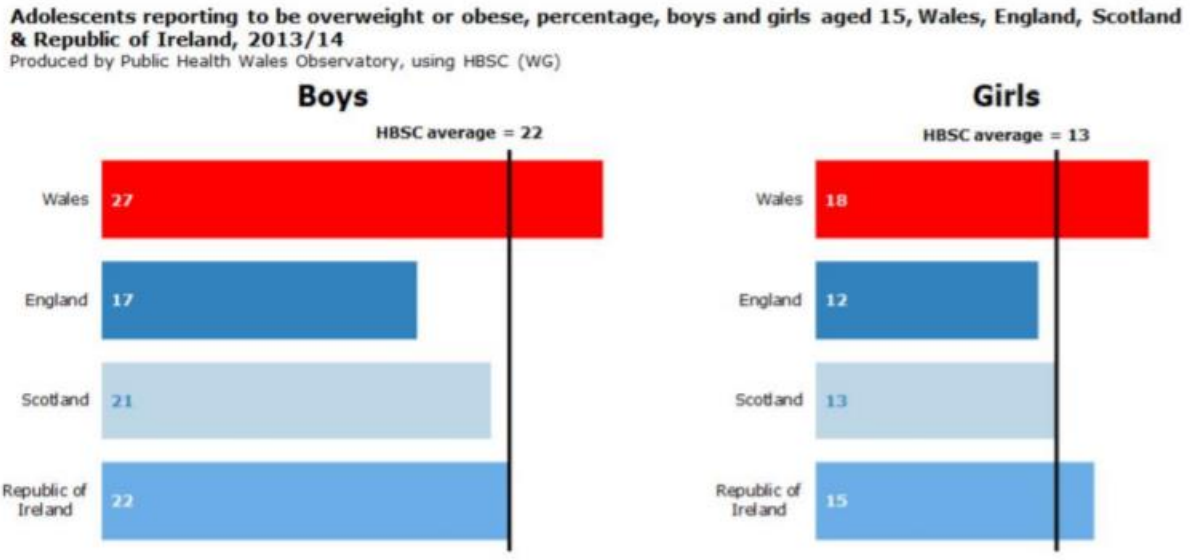

Figure 1a: Percentage 15yr-olds reporting to be overweight or obese by country 2013/14 (Duncan-Jones et al., 2019)

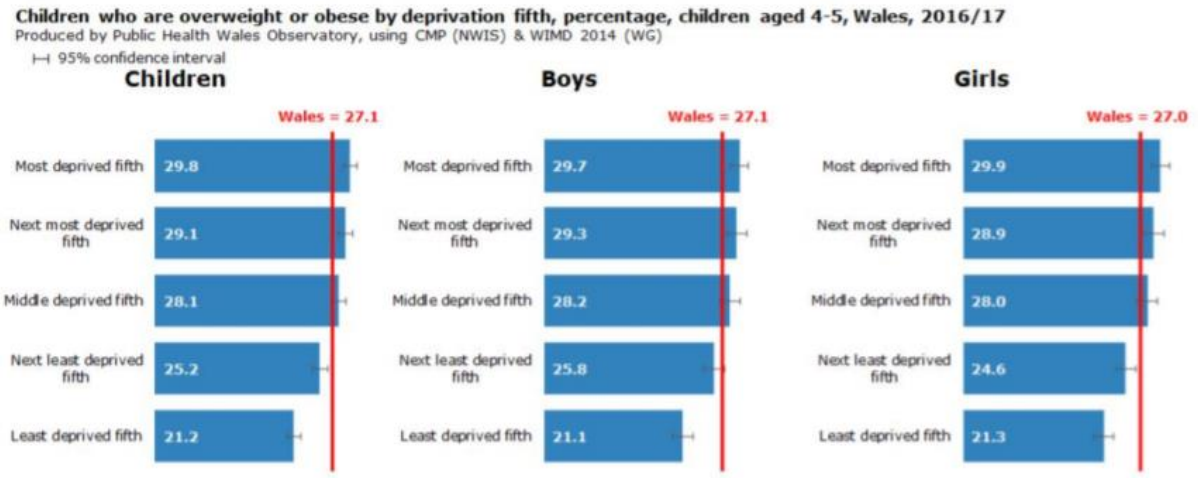

Figure 1b: Percentage 4-5yr-olds overweight or obese by deprivation quintile 2016/17 (Duncan-Jones et al., 2019)

Beynon & Bailey (2019) identified that on average 3.1% of children in Wales were severely obese with figures varying between health boards (2.5% in Powis, to 3.9% in Cwm Taf); Swansea Health Board being 2nd highest with 3.3% of its children classified as severely obese, again suggesting that the degree of severe obesity was linked to socioeconomic deprivation. They also found that levels of severe obesity were significantly higher in boys than in girls (3.6% *cf.* 3.0%), although these figures could underestimate the true extent of the problem because current child measurements in Wales are determined at early childhood (Jarvis et al., 2022). They found that by around 10 years-of-age, approximately one quarter of boys and girls in late childhood were classified at an unhealthy weight and they predicted that this percentage had increased to approximately one-third in both sexes by 18 years. Furthermore, childhood obesity has been found to be difficult to address. Skinner et al. (2018) found that despite substantial clinical and policy efforts targeting obesity, rates in US children had not declined between 1999 and 2016.

Increases in childhood obesity demonstrate a worrying trend and is one of the most significant current challenges to public health. Obesity affects multiple organs and is associated with significant morbidity and ultimately premature death. Furthermore, complications associated with obesity, including dyslipidaemia, hypertension, MAFLD, and psychosocial complications are becoming increasingly prevalent within the paediatric populations (Apperley et al. 2021). Rao et al. (2020) conducted a systematic review and meta-analysis which demonstrated that obesity increases the risk of depression in children and adolescents; and Sanders et al. (2015) found that overweight or obese Australian children and adolescents had more cardio-metabolic and MAFLD risk factors, and were experiencing more negative psychological outcomes (depression, low self-esteem, and lower scores of health-related quality of life), compared to normal-weight peers.

Apperley et al. (2021) note that the number of potential therapies for children and the research to support their use is lacking when compared to research and treatment available for adults

with obesity. They conclude that lifestyle interventions, the current focus of childhood obesity management, is not sufficient in many patients and that pharmacotherapy may be required. This is collaborated by Chadda et al. (2021), stating that pharmacological options for management of both obesity and T2DM in children are limited. NHS Digital (2020), have linked obesity to an increased risk of T2DM, constipation, MAFLD, and mental health issues; another condition which has increased over the last few years (NHS Digital & Thandi, 2020).

A further concern expressed by Rees et al. (2009) is that even though comorbidities of obesity usually present in adulthood, the underlying factors may originate from childhood due to their long-term nature. They emphasise the importance of finding when health consequences and risk factors for these diseases occur, and how early the diseases can be detected if they are to be addressed successfully. Childhood obesity must be tackled early so that it can be managed before the onset of complications and to reduce the prevalence of adult obesity.

As the National Assembly for Wales' Committee said in their booklet 'Childhood Obesity' (Senedd.wales, n.d.) "There is general agreement that childhood obesity is a problem that needs to be tackled…[and]…there isn't a silver bullet to solve the problem. It seems that the best approach is a suite of solutions, covering a number of different policy area"' and in 2016 the UK government introduced 'Childhood Obesity A Plan for Action' (HM Government, 2016).

Current NICE guidelines (2022) state that "Tailored clinical intervention should be considered for children with a BMI at or above the 91st centile, depending on the needs of the individual child and family." And that "assessment of comorbidities (such as hypertension, hyperinsulinaemia, dyslipidaemia, type 2 diabetes, psychosocial dysfunction and exacerbation of conditions such as asthma) should be considered for children with a BMI at or above the 98th centile" (the boundary for very overweight). The guidelines also state that these children should be assessed for psychosocial distress, such as low self-esteem, teasing and bullying. NICE (2013) identified a research focus identifying "a lack of data on effective and cost-effective approaches to weight management for children younger than 6 years, including the views of their parents and families."

1.2) Type two diabetes mellitus

In the 1990s, type two diabetes mellitus (T2DM) rarely occurred in children (Serbis et al. 2021) and was until recently thought of as a disease of older individuals (Piscopo et al., 2005). Although still a rare disease, incidence of T2DM in children and adolescents is mirroring an increase in both prevalence and severity of paediatric obesity since; and in 2005, the mean age of diagnosis of T2DM in young people was 12-14 years (Piscopo et al., 2005; Serbis et al., 2021). Dündar & Akıncı (2022) detected insulin resistance in over 70% of their cohort of overweight and obese children with 3.7% being diagnosed with T2DM.

T2DM is preventable if its major risk factor, obesity, is managed; but Serbis et al. (2021) suggest that families, schools, doctors, health services, and policy makers tend to accept obese as the new 'normal' and predict that no management approaches will be able to protect the next

generation from many years of serious health problems and low-quality life unless prompt action is taken.

There is a more rapid progression of T2DM and related complications in children and young people than in children with type 1 diabetes mellitus (T1DM) or adults with T1DM or T2DM (Barrett et al., 2020), with faster beta-cell decline and a greater risk of complication in young onset T2DM than T1DM (Magliano et al., 2020). Magliano et al. (2020) found, T2DM is higher in adolescent girls than in adolescent boys. Early diagnosis and new treatments for children with T2DM in children are urgently required (Barrett et al., 2020).

1.3) Metabolic associated fatty liver disease

Metabolic associated fatty liver disease (MAFLD, formally known as non-alcoholic fatty liver disease) has become the most common causes of chronic liver disease in the paediatric population (Chen & Pan, 2022; Jia et al., 2022). Its prevalence is higher in obese clinical populations than in the general population, it is greater in males than females, and increases incrementally with BMI category. However, prevalence of MAFLD in both children and adults in the general population is difficult to assess accurately due to a lack of non-invasive diagnostic tests. Furthermore, current markers may underestimate MAFLD prevalence in young obese people and overestimate prevalence in the general population (Anderson et al., 2015). MAFLD in children is mostly associated with insulin resistance and obesity (Jia et al., 2022), with the impact of genetic variants being greater in children who are overweight (Lee et al., 2022). The rising cases of obesity in children and adolescents in recent years have led to increased MAFLD cases (Jia et al., 2022).

Therefore, T2DM and MAFLD are both metabolic diseases that are strongly associated with obesity and deprivation, and are increasingly seen in children. There is an urgent clinical need to identify children at risk in order to provide both preventative interventions and treatment to children at risk.

1.4) Depression

Studies have also demonstrated that obesity is associated with risk of depression in children (Lindberg et al., 2020; Rao et al., 2020). Sutaria et al. (2018) found strong evidence that obese female children have a significantly higher risk of depression than healthy weight counterparts, and that this risk persists into adulthood, although this was not true for obese male children who had no significant increase in depression. Furthermore, I have not found longitudinal studies exploring whether obesity in younger children is linked to depression in their teenage years.

1.5) Constipation

Constipation is another emerging global public health concern (Rajindrajith et al., 2016), and is one of the most common chronic disorders of childhood (Nurko & Zimmerman, 2014). In the United States, children with constipation cost the health care system three times as much as children without constipation, and the negative effect on quality of life can persist into adulthood (Nurko & Zimmerman, 2014). However, I could not find much research, nor a clear

consensus in the research I did find, as to whether there is an association between constipation and obesity. For example, Koppen et al. (2016) found no association between functional constipation and overweight or obesity in children in Colombia, but Pashankar & Loening-Baucke (2005) found prevalence of obesity significantly higher in constipated children in their study.

Therefore, ways of identifying children most at risk of developing serious health conditions due to obesity as early as possible is of great importance.

1.6) Aims

In this project I aim to use general practitioner (GP) data from the SAIL databank to determine the following questions:

1. What is the relationship between T2DM, MAFLD, constipation, and depression; and BMI and WIMD of my cohort?
2. Is it possible to develop an effective machine learning classifier to identify children at risk of developing T2DM, MAFLD, constipation, and depression using BMI, deprivation, and sex as predictors?
3. Can predictors in children aged 2-6 and 6-10 be used to identify increased risk of developing comorbidity outcomes in later childhood?

2) Methods

2.1) The Secure Anonymised Information Linkage Databank

The Secure Anonymised Information Linkage (SAIL) Databank is a national databank of primarily General Practitioner (GP) data from all across Wales (Ford et al., 2009; Lyons et al., 2009; SAIL, 2022). It holds billions of anonymised patient records that can be linked to other data including the Welsh Index of Multiple Deprivation (WIMD) (Statswales, 2019). I used patients records from SAIL taken between January 2001 and December 2021.

2.2) The Welsh Index of Multiple Deprivation

WIMD is a government-run summary of deprivation of the whole of Wales based on census data. It divides the country into small areas (lower super output areas, LSOA) and ranks them on their deprivation according to: income, employment, health, education, access to services, community safety, physical environment, and housing. The WIMD scores I will be using is a summary of these factors. To maintain anonymity, I only have access to each child's WIMD as a score of 1-10 where 1 is most deprived. For the purposes of this study, I will use the 2014 revision as it represents a time over which I study.

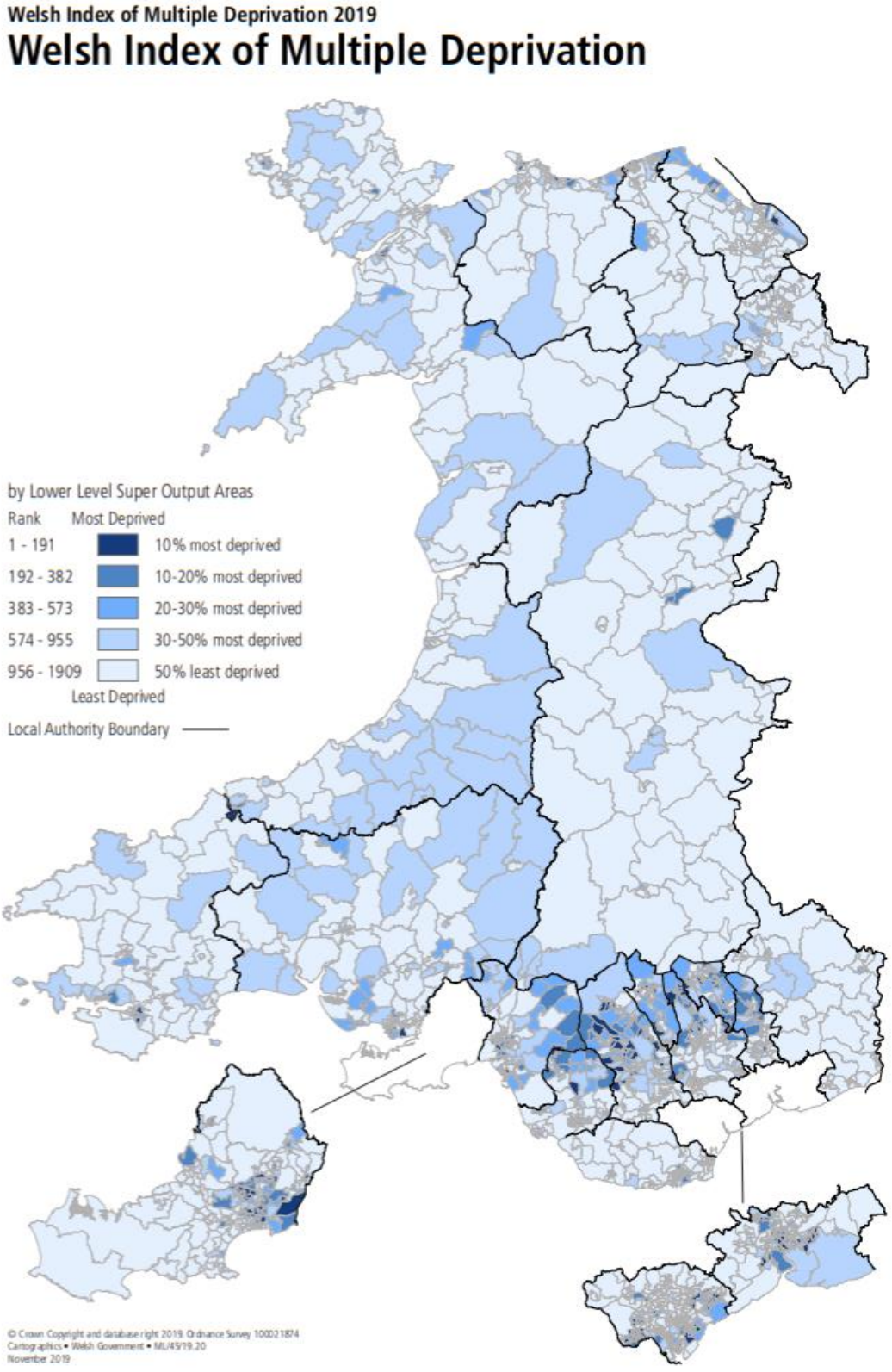

Figure 2: A map of Wales by LSOA indicating deprivation according to WIMD (Statswales, 2019)

2.3) Structured Query Language (SQL)

2.3.1) Joining Tables

After receiving access to the SAIL Databank, I used the data manipulation language SQL to extract relevant information from the many tables provided. Below shows mock tables that exemplify the data available (Tables 1a-f). In Table 1a, each row represents a GP visit, and an "event" represents a test, diagnosis, or prescription and is coded using International Classification of Diseases (ICD) coding (ICD, 2022).

Table 1a: Mock patient GP data

Patient ID	Gender Code	Date of Birth	Event Code	Event Value	Event Date
12345	1	03/06/2014	22K..	19	05/05/2016
12345	1	03/06/2014	C10F.	N/A	09/02/2018
12346	2	27/11/2007	J520.	N/A	15/12/2017
12346	2	27/11/2007	J61y1	N/A	06/10/2015
12346	2	27/11/2007	22K..	18	08/09/2014
12347	2	10/01/2009	22K..	20	20/09/2020
etc.	…	…	…	…	…

Table 1b: Mock event code descriptions

Event Code	Event Description
22K..	BMI
C10F.	Type 2 Diabetes
J520.	Constipation
E112.	Major depressive episode
J61y1	Metabolic associated fatty liver disease
etc.	…

Table 1c: Mock patient WIMD scores

Patient ID	WIMD Quartile	WIMD Quintile	WIMD Decile
12345	3	3	6
12346	1	1	2
12347	4	5	9
etc.	…	…	…

These tables can be linked by comparing common columns (for example patient ID) to compile the information relevant to the study. New tables can also be made to add new columns to pre-existing tables, or only select rows with relevant events.

I made a new table using only rows which had the event code for a body mass index (BMI, a ratio of a person's mass to their height squared) measurement, and the difference between event date and date of birth being between 2 and 16 years. I also only took patients born after 1999 since records taken before 2001 are often unusable.

I did not use diagnosed obesity as a predictor, as far more people with a high BMI have their BMI recorded as such than being diagnosed with obesity as their event. This also allows me to quantify adiposity where obesity diagnoses do not.

I then made the following changes to the table to aid logistic regression.

2.3.2) Removing erroneous data

I only selected records where BMI was 0-100 to remove erroneous data. I used this broad range because it is very hard to separate real extreme values from false ones. A range of 0-100 removes values that would significantly harm the dataset (for example BMI=$1x10^6$) but keeps all possible real BMI values.

2.3.3) Gender code 9

A very small number of patients had a gender code of 9, representing indeterminate sex. Because of their rarity, it is likely that my model will not encounter gender code 9 during training and hence cause an error when it encounters it during testing. Therefore, I removed all patients with indeterminate sex.

2.3.4) BMI normalisation

Because the healthy range for BMI changes as children age and whether they are male or female, it is recommended that adiposity in children is diagnosed using boundaries in Z-score (distance from the mean) or percentile (a child's BMI ranked in a cohort set to a scale of 0-100) (Gov.wales, 2021). However, these methods are unsuitable for machine learning, where more continuous data allows more accurate models (Must & Anderson, 2006). Furthermore, these percentiles and Z-scores were last updated in 2012 are based on cohorts from 1990 (RCPCH, 2022) which no-longer represent the modern cohort (Apperley et al. 2021). Therefore, I used raw BMI (a continuous variable) as my main predictor.

To mitigate this effect, I split my cohort into three age groups and used sex as a covariate. I also tested the use of BMI normalised within my own cohort as an alternative predictor to BMI. To do this, I ranked all the children by their BMI per age and gender and gave them each a percentile ranking.

2.3.5) Creating binary variables

I then joined this new table with the original GP table to include all events (that have occurred between ages 2 and 16) of the pre-selected patients. I then added another column stating if the event is a diagnosis for four relevant comorbidities (T2DM, constipation, MAFLD, and depressive events), and another binary "outcome" (1/0) column for each comorbidity, using a variety of event codes for the conditions.

Table 1d: Mock event comorbidity data

Patient ID	Sex	Date of Birth	WIMD Decile	BMI	Event Code	Event Value	Event Date	Comorb.	T2DM	MAFLD	Const.	Dep.
12345	1	03/06/2014	6	19	22K..	19	05/05/2016	N/A	0	0	0	0
12345	1	03/06/2014	6	19	C10F.	N/A	09/02/2018	T2DM	1	0	0	0
12346	2	27/11/2007	2	18	J520.	N/A	15/12/2017	Const.	0	0	1	0
12346	2	27/11/2007	2	18	J61y1	N/A	06/10/2015	MAFLD	0	1	0	0
12346	2	27/11/2007	2	18	22K..	18	08/09/2014	N/A	0	0	0	0
12347	2	10/01/2010	9	20	22K..	20	20/09/2013	N/A	0	0	0	0
etc.	…	…	…	…	…	…	…	…	…	…	…	…

2.3.6) Separating age groups

I then separated the patients into groups of 2-6-yr-olds, 6-11-yr-olds, and 11-16-yr-olds (where both a BMI recording and event occurred in this time-frame), and compiled the events of the three tables so each row represented one patient, summarising the four binary outcome columns. I took an average of all BMI in each time frame to give a single value for each patient per age group.

Table 1e: Mock compiled comorbidity data

	Patient ID	Sex	Date of Birth	WIMD Decile	BMI	T2DM	MAFLD	Constipation	Depression
2-6	12345	1	03/06/2014	6	19	1	0	0	0
6-11	12347	2	10/01/2009	2	20	0	0	0	0
11-16	12346	2	27/11/2007	9	18	0	1	1	0
	etc.	…	…	…	…	…	…	…	…

2.3.7) Quantifying severity

For constipation and depression, severity may be somewhat quantified using the number of events there are regarding it for each patient. Therefore I counted the number of events regarding this comorbity in the given age group in another column.

2.3.8) Predicting comorbidity from BMI taken at a younger age

Finally, I made more tables following children from the younger age groups into the older age groups using their patient ID. I took their BMIs from the younger age group and lined them up with any comorbidities suffered in a later age group, forming three tables: BMI from 2-6-yr-olds and comorbidity in 6-11-yr-olds (Group 1); BMI from 2-6-yr-olds and comorbidity in 11-16-yr-olds (Group 2); and BMI from 6-11-yr-olds and comorbidity in 11-16-yr-olds (Group 3).

Table 1f: Mock Group 1 data

From ages 2-6					From ages 6-11			
Patient ID	Sex	Date of Birth	WIMD Decile	BMI	T2DM	MAFLD	Constipation	Depression
12345	1	03/06/2014	6	19	0	0	0	1
etc.	…	…	…	…	…	…	…	…

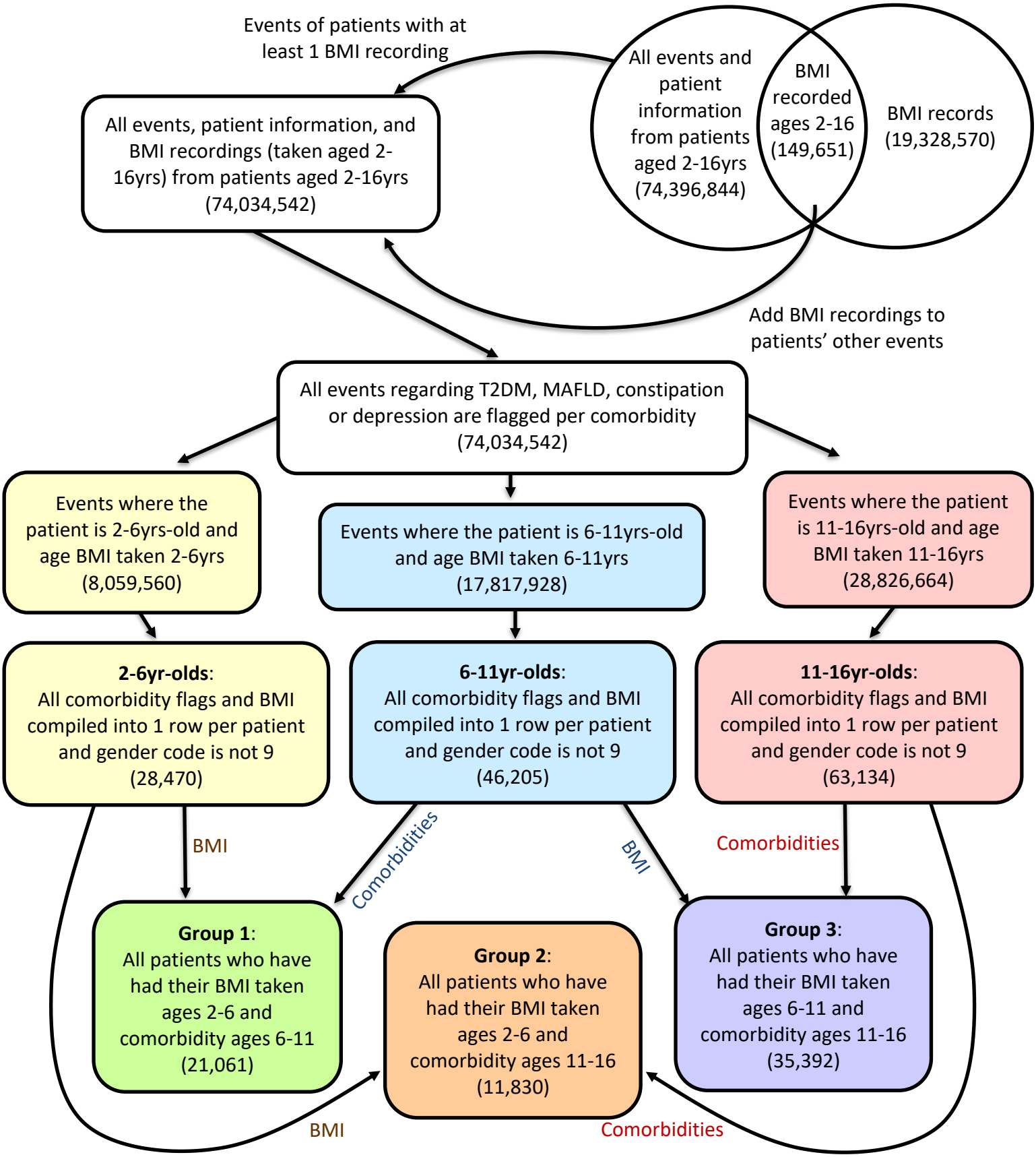

Figure 3: A flow diagram demonstrating process of acquiring patient cohorts

2.4) R - A statistical programming language

I then imported these tables into R to perform statistical analyses.

First, I counted the number of children per comorbidity in each of the six groups (Table 3). I plotted histograms of BMI and WIMD for each age group to assess the spread within the dataset (Figures 6-7).

I then plotted boxplots of the BMI and WIMD distribution per comorbidity, as well as the overall cohort, which show the median, interquartile ranges and minimums and maximums (ignoring outliers) per group. To preserve confidentiality, I have not plotted outliers; and I have represented all groups of patients <5 as such. Because I was comparing the means of very small cohorts, I used the non-parametric Wilcoxon test to see if the link between comorbidity and BMI, or deprivation are statistically significant (Figures 8-9).

I also used the χ^2 test to find the significance of patients having more than one comorbidity within each age-group. *N.B.* I did not perform this analysis on children ages 2-6 because there were too few people with comorbidity besides constipation to compare (Table 5).

2.5) Logistic regression

2.5.1) Introduction to Logistic regression

Logistic regression (LgR) is the process of mapping a sigmoid curve onto a set of training data.

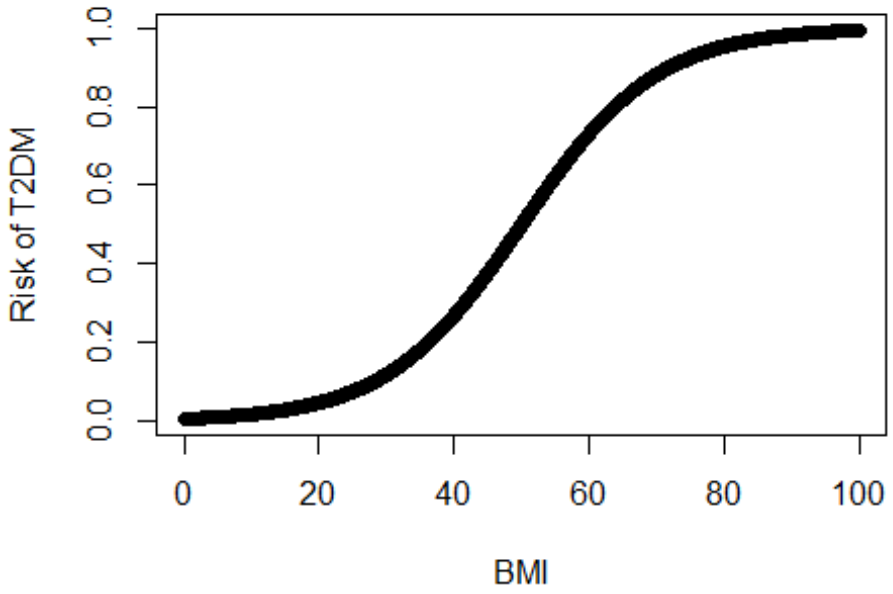

Figure 4: Mock LgR sigmoid curve

The LgR algorithm plots each data-point onto a graph of predictor against risk, where patients having had a GP event regarding the comorbidity are labelled 1, and those who have not are labelled 0. Then the algorithm maps a sigmoid curve onto the graph and calculates a formula from which we can calculate risk for future predictor values. This is called a model.

For example, in Figure 4, a BMI of 35 would give a 0.2 risk of T2DM. From this, we can either use risk on its own, or we can apply a cut off, above which we say that the person is high risk, and below low risk.

I used LgR as the machine learning model of choice because it is trained on a binary outcome (this person has a comorbidity or not) and is most effective when used on continuous or ordinal predictors which BMI and WIMD, my main predictors, are; yet it can also make use of categorical data like gender code. Furthermore, it outputs a true risk; something which would be useful in a clinical setting were nuance is preferred over absolutes. This is not possible with any other binary machine learning algorithm.

2.5.2) Train-test split

Before training the LgR models, I split the datasets into train and test sets. This means that I can use a proportion of the data to train the model on, whilst leaving the rest to test the model.

I varied this train:test ratio to optimise the models. I tried 0.7:0.3, 0.75:0.25, 0.8:0.2, and 0.9:0.1. I do this to weigh up having more data to train the model on(the more training data, the better the model), whilst holding back some to test (no matter how effective a classifier is, if we don't know its accuracy, it severely limits its usefulness). I tried using higher train:test ratios due to the lack of positive (comorbid) training data. After preliminary tests (see appendix 1), I decided to use a 0.75:0.25 ratio. Despite all ratios providing similar accuracies, this the highest accuracy whilst maintaining a somewhat large test set, to give more confidence in any results.

2.5.3) Down-sampling

Generally, machine learning algorithms are trained so that accuracy is maximised, and the model fit the data as well as possible. Most children have not had a GP event regarding any of the relevant comorbidities (approximately 0.01% of children were recorded as having T2DM). Therefore, if I fed the LgR algorithm the whole training dataset, it would classify every person as not having said comorbidity, thereby ensuring a high accuracy (99.99% for the example of T2DM). This does not produce a useful classifier. Therefore, I cut down the training datasets by taking all records of children with the relevant comorbidity, and randomly sampling an equal number of records of children without in a process called down-sampling. This forces the model to classify between the positive and negative cases, rather than assuming a negative result for everyone.

Another option to combat this is to up-sample. This involves taking the positive data-points, looking at their predictor variables, and interpolating possible predictor values for synthetic positive data-points. This allows more real negative data-points to be used during training, and provides more data to train and test on. However, this produces synthetic data on which the

model is being trained and tested which could be unrepresentative of real people. Furthermore, any deviations from the mean of the original positive population will be amplified, but the model will have the certainty as if it was trained on a larger cohort. Rare events are disproportionately represented in small cohorts. Therefore, I used down-sampling.

2.5.4) Combined comorbidities

Because so few children have these comorbidities, I also made models classifying children as having had a GP event regarding any of the specified comorbidities, and those classifying children as having had a GP event regarding either T2DM or MAFLD since these have been shown to be linked (Kosmalski et al., 2022).

2.5.5) Training LgR models

I performed LgR for each of the six groups of children using BMI, sex, and WIMD score as predictors for each of T2DM, MAFLD, constipation, and depression separately. Obesity (which is defined using BMI) is thought to be linked with all four of these comorbidities (except in constipation where there is uncertainty) (Anderson et al., 2015; Lindberg et al., 2020; Pashankar & Loening-Baucke, 2005; Serbis et al., 2021), and deprivation is associated with obesity. It is also known that sex has a significant impact on comorbidity risk (Anderson et al., 2015; Magliano et al., 2020; Sutaria et al., 2018).

Year of birth may represent the general increase in comorbidity rates over time, or it could act an indicator for trends that we cannot account for, however I did not use this as it was not a significant factor in many preliminary models; and although year of birth may be of use for classifying patients born within this time-frame, extrapolating this into the future may not represent the true trend.

I did not use age at which the event occurred in the pooled age groups since this would create a duplicate row every year a patient went to the GP, potentially forming many rows of children who had a comorbidity and attended the GP within the given year, but not regarding their comorbidity, and hence would not be recorded as having any comorbidity in the table.

There were many groups where there were very few patients with these conditions. For 2-6-yr-olds for example, there were no children with MAFLD or depression, and fewer than 5 with T2DM. Having so few positive cases makes it impossible to make an accurate predictor. Therefore, for 2-6-yr-olds, 6-11-yr-olds, and Group 1, I only made models classifying constipation and for Group 2 I made models predicting constipation and depression.

2.6) Alteration of parameters

2.6.1) Normalised BMI

Next, I ran the same models using normalised BMI.

2.6.2) Year of birth

I tested a version of the models using year of birth as an extra predictor to see if this would significantly affect the efficacy of the models.

N.B. I only show these alterations on the most effective classifiers with the largest cohorts where they are relevant. This is the most reliable way of demonstrating their effect on the models.

2.7) Linear Regression

I then performed linear regression using the same predictors to attempt to predict severity of constipation and depression using the number of GP visits regarding the respective comorbidity, since it was not uncommon for children to have many visits a year regarding constipation and depression.

2.8) Testing the models

2.8.1) Predictor significance

Firstly, I can look at the models themselves to see how they have weighted the predictor variables, and hence how useful the predictors are at predicting specific outcomes. The weighting assigned to each predictor is called a coefficient. Each coefficient is labelled as being a significant predictor or not.

2.8.2) Confusion matrices

To test the classifiers, I used test set predictor variables and each model to predict the presence of comorbidity. This outputs the comorbidity risk. I then rounded the output to 0 (negative, no comorbidity) or 1 (positive, comorbidity), and compared the predicted outcome with the true outcome. This produces four groups: true positives (TPs), those accurately predicted as being positive; false negatives (FNs), those predicted negative but are actually positive; true negatives (TNs), those accurately predicted negative; and false positives (FPs), those predicted positive but are actually negative. I then represented these figures in a table called a confusion matrix (see table 2).

Table 2: Confusion matrix

	Observed negative	Observed positive
Predicted negative	TN	FN
Predicted positive	FP	TP

From this, I can calculate the model's sensitivity (the ability of a model to identify positives, TP/(TP+FN)) and specificity (the ability of a model to identify negatives, TN/(TN+FP)) which, can

be a better way to judge a model's usefulness than accuracy ((TN+TP)/(TN+FP+FN+TP)), as well as the positive predictive value (PPV, the likelihood a patient who is predicted positive will be a TP, TP/(TP+FP)), and the negative predictive value (NPV, the likelihood a patient who is predicted negative will be a TN, TN/(TN+FN)).

To optimise the confusion matrices, I calculated the cut-off which corresponded with the maximum sum of sensitivity and specificity for each model and used that cut-off for the confusion matrix.

2.8.3) ROC curves

A confusion matrix only shows the values due to one cut-off. Another measure of a model's usefulness is the area under a receiver operating characteristic (ROC) curve. This plots TP rate (sensitivity) against FP rate (1-specificity) using many cutoffs between 0 and 1, giving a more holistic view of the model's efficacy. A good classifier will show a curve that follows the top left-hand corner of the graph (Figure 5a), maximising TP rate and minimising FP rate, whereas a bad classifier will not deviate from a straight line from the bottom left-hand corner to the top right-hand corner (Figure 5b), meaning that the chance of the classifier correctly classifying the patient is close to 0.5.

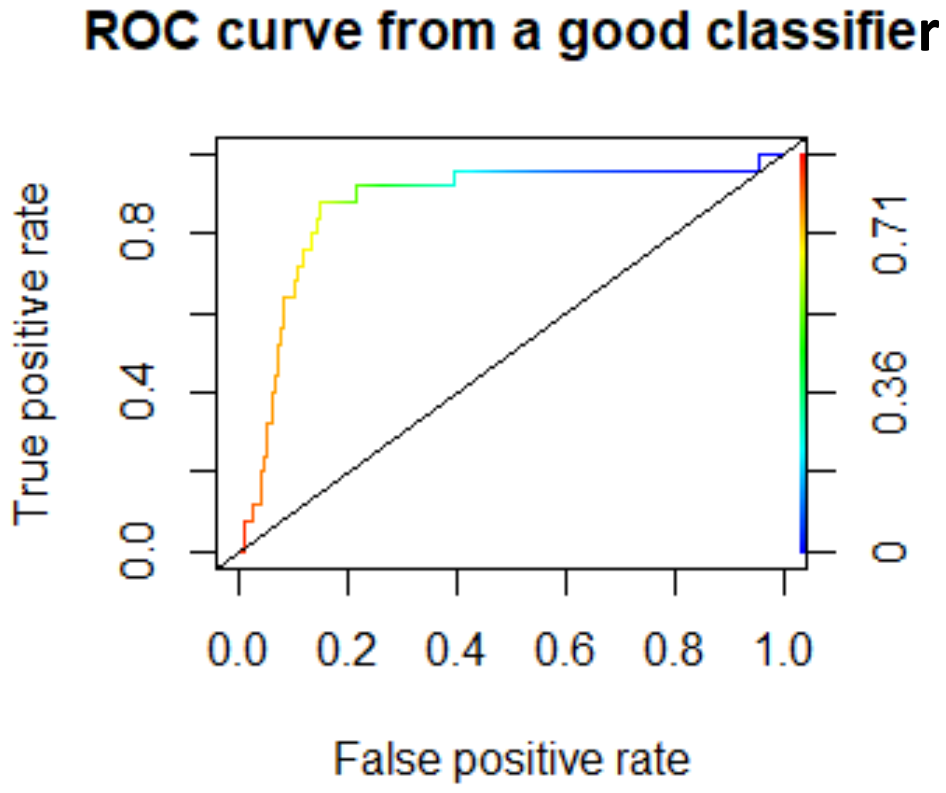

Figure 5a: An example of a ROC curve showing accurate classification

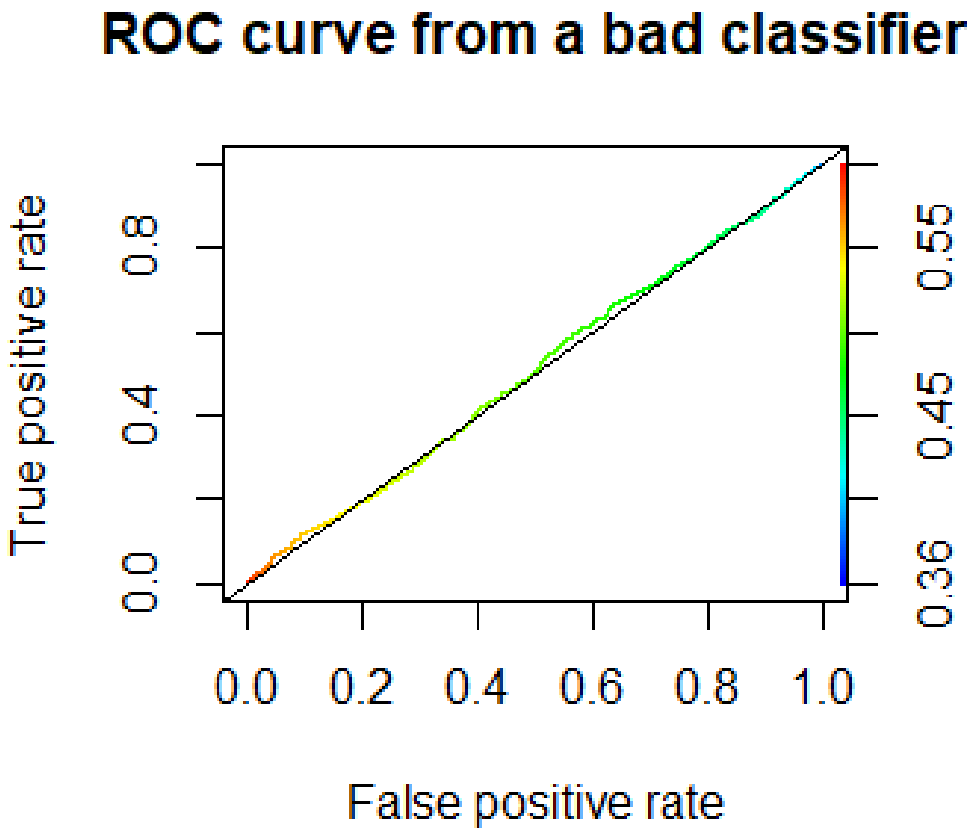

Figure 5b: An example of a ROC curve showing inaccurate classification

The area under the ROC curve (AUC) demonstrates the ability for the classifier to correctly classify the test set. AUC=0.5-0.6 is considered a failed model, 0.6-0.7 is poor, 0.7-0.8 is fair, 0.8-0.9 is good, and 0.9-1.0 is considered excellent (El Khouli et al., 2009). I can also use the colour spectrum on the right of the graph to see which cut-off points correspond to which points on the ROC curve.

2.8.4) Crossfold-Validation

Machine learning models are inherently random. LgR is comparatively reproducible, but there is randomness when splitting for test and train sets, leading to slightly different training sets each time it is run. To make the models more reproducible and quote a more representative efficacy, I performed crossfold-validation (CV). This splits the data, and runs and tests the algorithm multiple times (10 in this case)to find the average test results. Although this does not stop results from varying each time it is run, it significantly reduces this variability, producing a more representative result (*N.B.* due to the way I coded this, each ROC curve and confusion matrix shows one model as an exemplar and hence may not exactly line up with quoted CV results).

2.8.5) Testing linear regression

I calculated the r^2 value of the linear regression model to quantify how well the model fits the data. r^2 is a value between 1 and 0. The higher the r^2, the better the data fits the model.

2.9) Ethical approval and code

This project has been approved by the SAIL Information Governance Review Panel (Project number 1415). Ethical approval is not mandatory for studies using only anonymised data as is the case with this project.

See link to GitHub repository for code in appendix 8

3) Results

3.1) Data exploration and statistical tests

Table 3: Number of children with each comorbidity in each test group

	2-6yrs	6-11yrs	11-16yrs	Group 1	Group 2	Group 3
Overall (to nearest 10)*	28470	46210	63130	21060	11830	35390
None	24889	42411	59329	19200	11236	33171
T2DM	<5*	8	39	<5*	<5*	27
MAFLD	0	6	46	0	11	44
Constipation	3580	3771	3047	1859	514	1855
Depression	0	11	747	<5*	68	338

*I have redacted the exact number of patients to help conserve anonymity as per SAIL policy.

N.B. Comorbidities are not mutually exclusive

Distribution of BMI of children of ages 2-6, 6-11, and 11-16 (figures 6a, b and c respectively), and events regarding children ages 2-6, 6-11, and 11-16 (figures 7d, e and f)

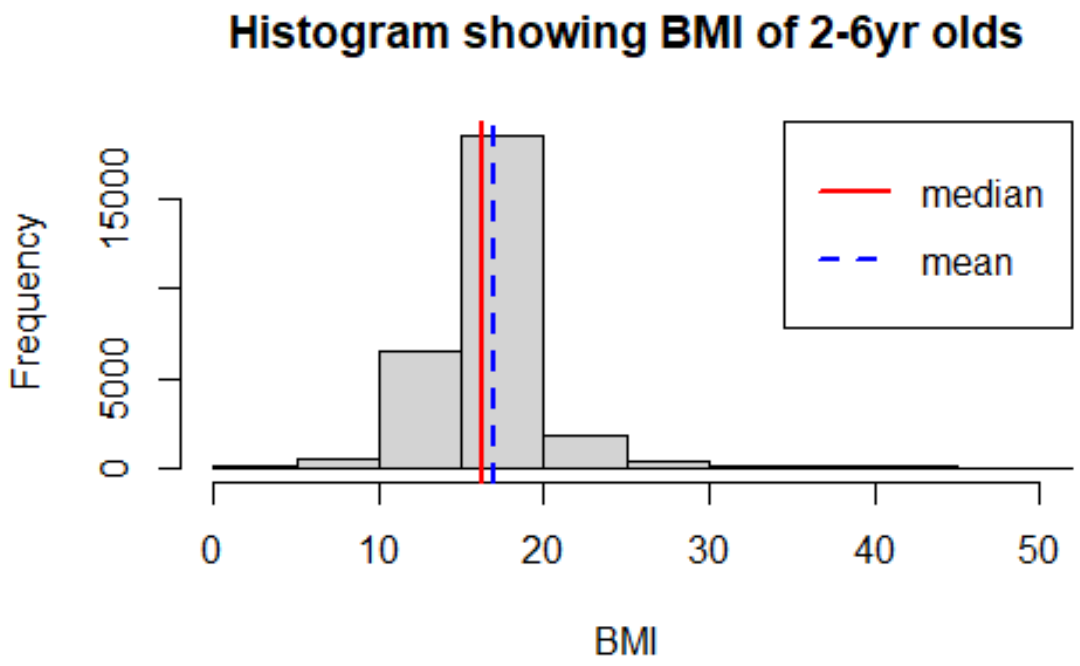

Figure 6a: Histogram showing distribution of BMI in 2-6-yr-olds

Figure 6a shows a normal distribution of BMI in 2-6-yr-olds as mean and median BMIs are very similar.

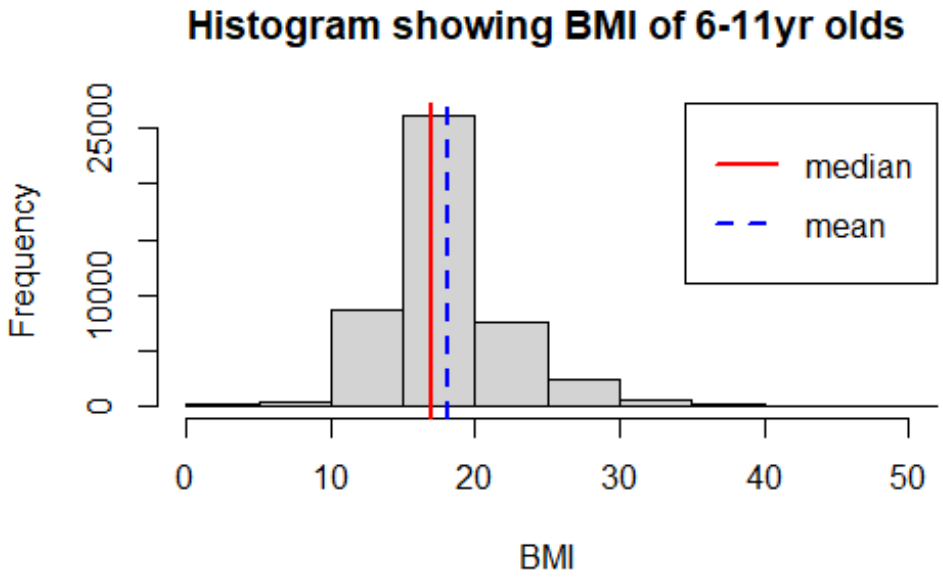

Figure 6b: Histogram showing distribution of BMI in 6-11-yr-olds

Figure 6b shows a mostly normal distribution of BMI in 6-11-yr-olds as mean and median BMIs are similar, but there is a slight skew towards higher BMI.

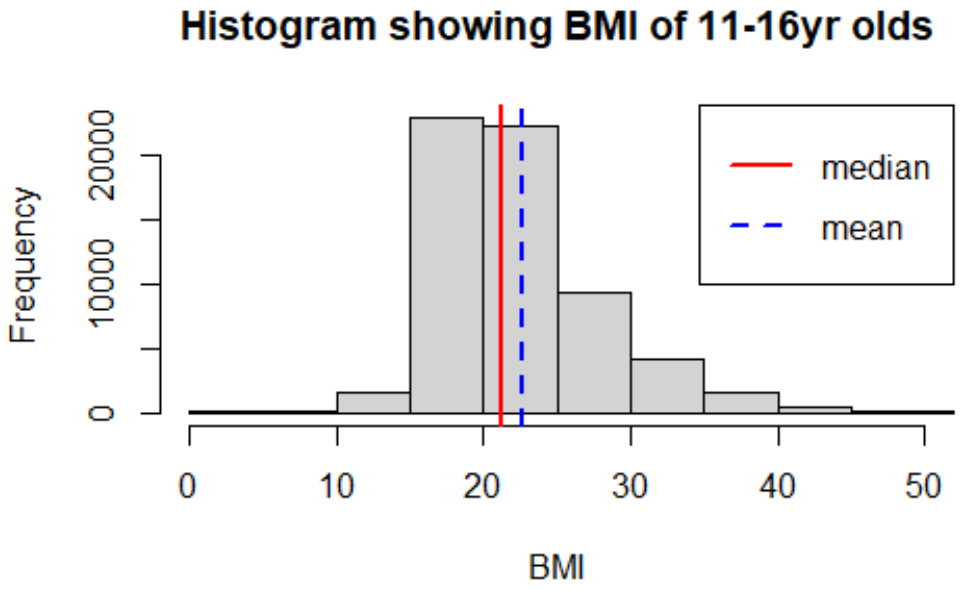

Figure 6c: Histogram showing distribution of BMI in 11-16-yr-olds

Figure 6c shows a normal distribution with a skew towards higher BMI in 11-16-yr-olds, with the mean BMI higher than the median.

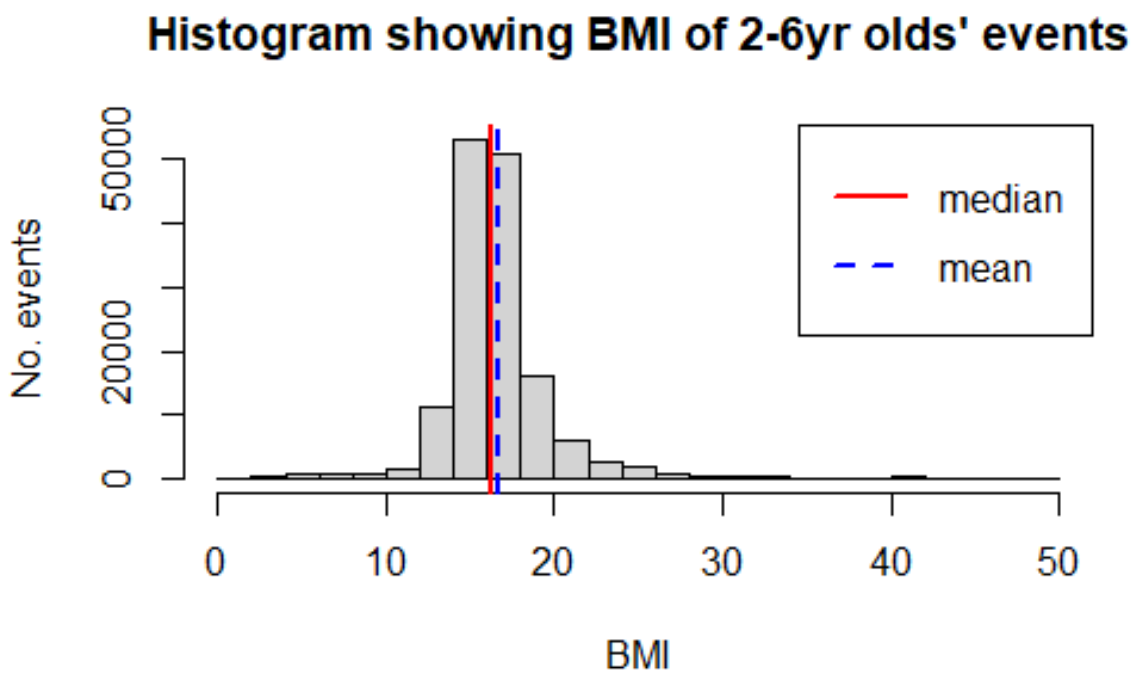

Figure 6d: Histogram showing distribution of BMI in events of 2-6-yr-olds

Figure 6d shows a normal distribution of BMI in 2-6-yr-old events as mean and median BMIs are very similar.

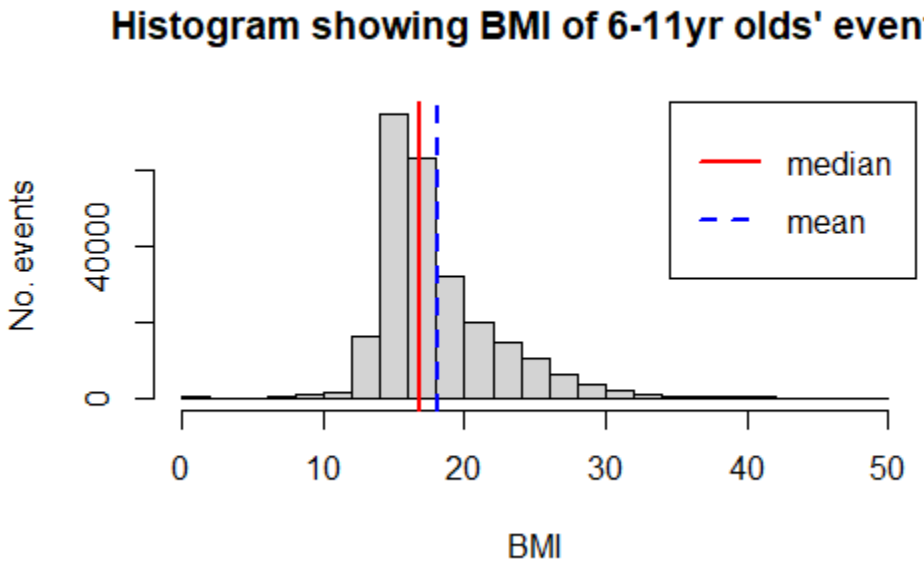

Figure 6e: Histogram showing distribution of BMI in events of 6-11-yr-olds

Figure 6e shows a normal distribution with a profound skew towards higher BMI in 6-11-yr-olds, with the mean BMI higher than the median.

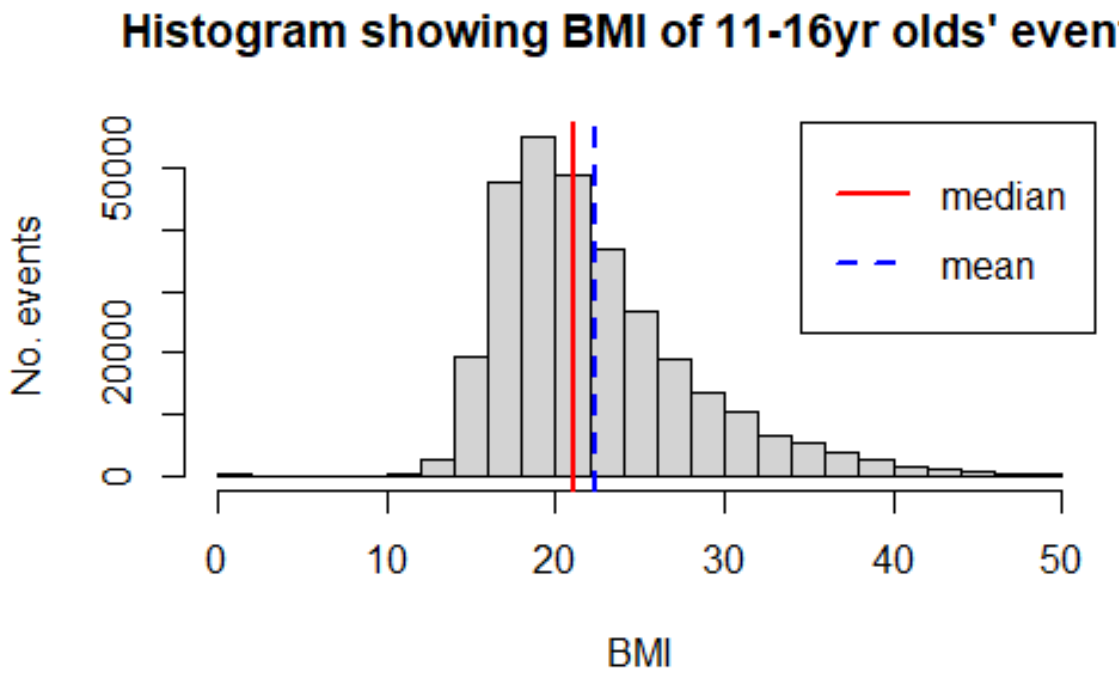

Figure 6f: Histogram showing distribution of BMI in events of 11-16-yr-olds

Figure 6f shows a normal distribution with a profound skew towards higher BMI in 11-16-yr-old events, with the mean BMI higher than the median.

Distribution of WIMD of children of ages 2-6, 6-11, and 11-16 (figures 7a, b and c respectively), and events regarding children ages 2-6, 6-11, and 11-16 (figures 7d, e and f).

N.B. Comparing mean to median is less informative for these since WIMD is ordinal.

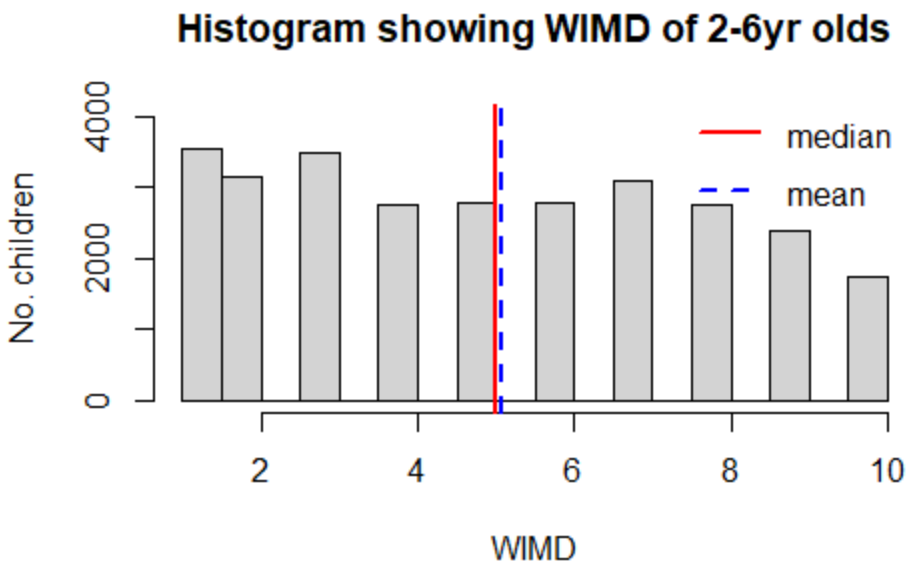

Figure 7a: Histogram showing distribution of deprivation as measured by WIMD in 2-6-yr-olds

Figure 7a shows a mostly uniform distribution of WIMD in 2-6-yr-olds. There is a very small skew towards higher deprivation.

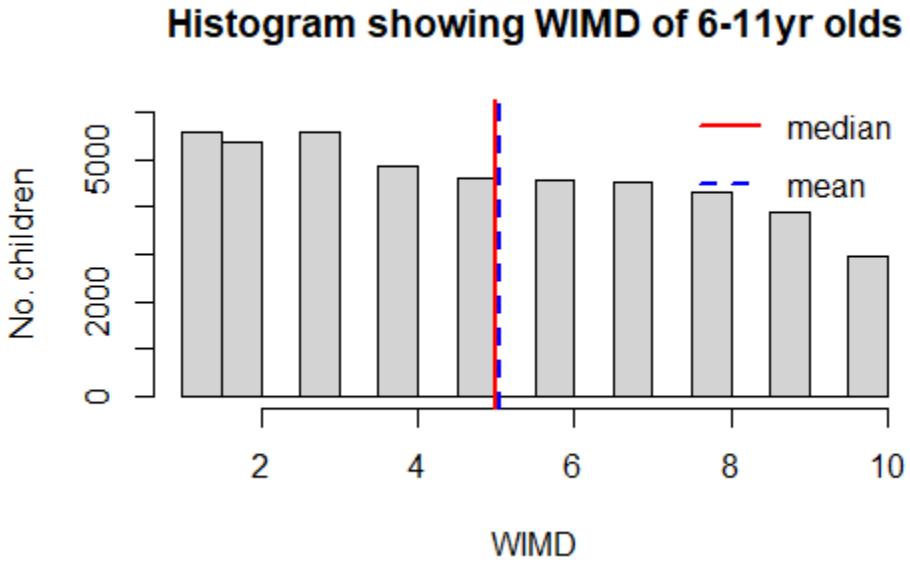

Figure 7b: Histogram showing distribution of deprivation as measured by WIMD in 6-11-yr-olds

Figure 7b shows a mostly uniform distribution of WIMD in 6-11-yr-olds. There is a small skew towards higher deprivation.

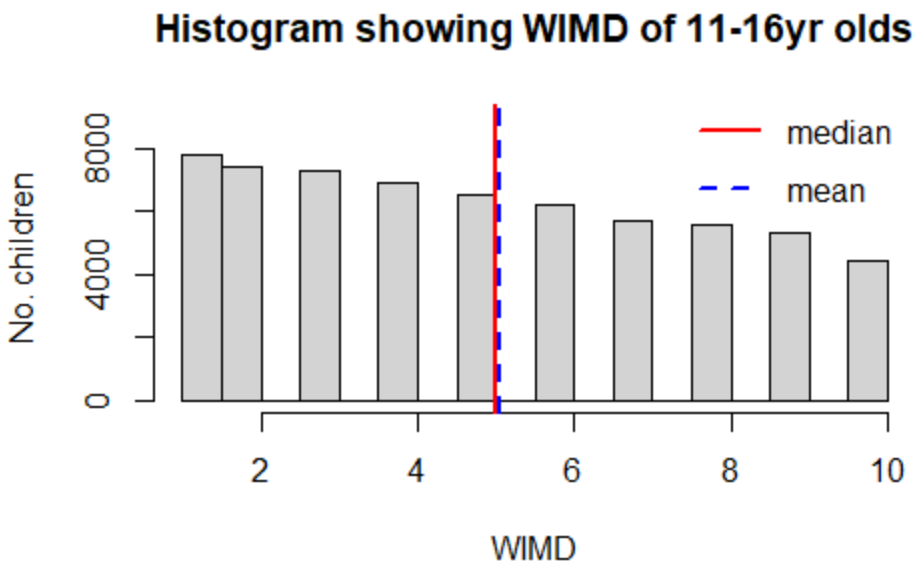

Figure 7c: Histogram showing distribution of deprivation as measured by WIMD in 11-16-yr-olds

Figure 7c shows a mostly uniform distribution of WIMD in 11-16-yr-olds. There is skew towards higher deprivation.

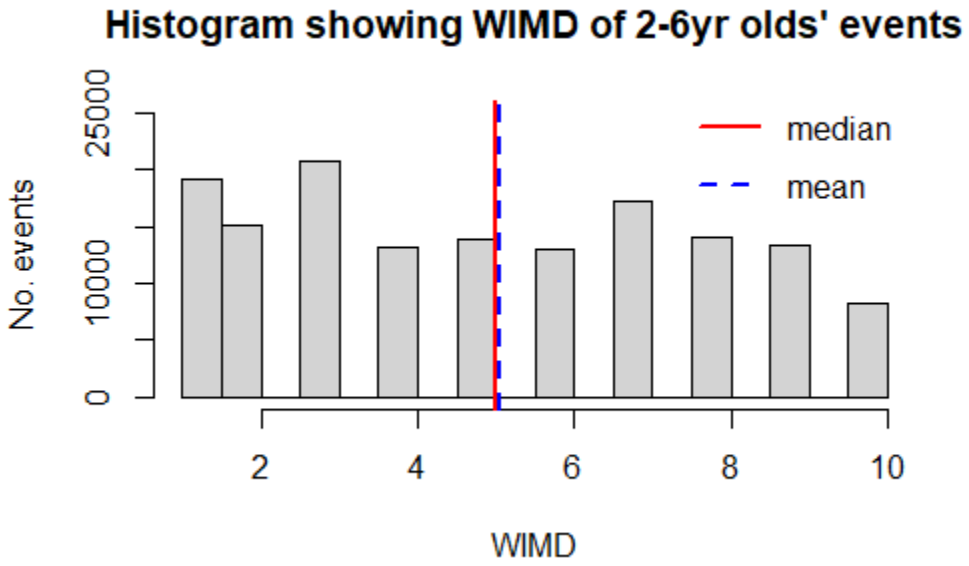

Figure 7d: Histogram showing distribution of deprivation as measured by WIMD in events of 2-6-yr-olds

Figure 7d shows a mostly uniform distribution of WIMD in 2-6-yr-old events.

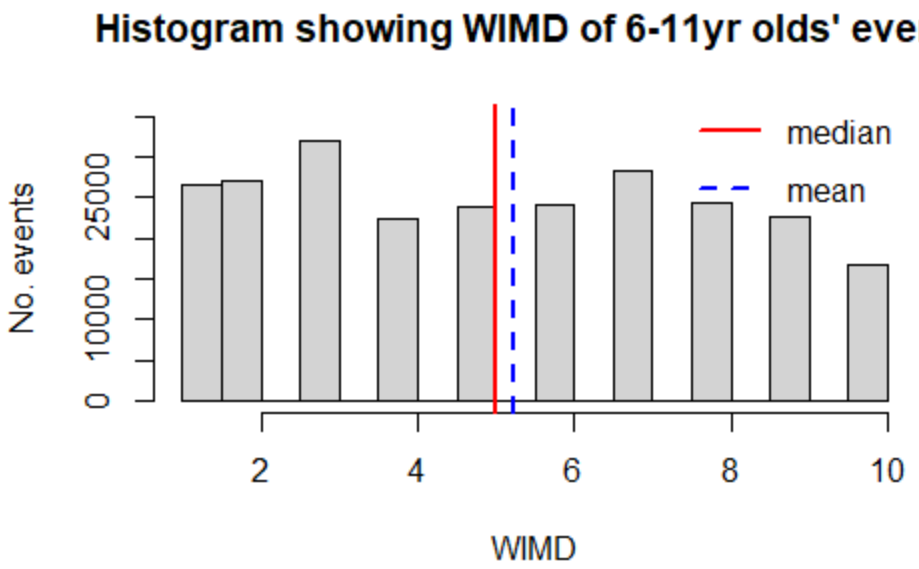

Figure 7e: Histogram showing distribution of deprivation as measured by WIMD in events of 6-11-yr-olds

Figure 7e shows a mostly uniform distribution of WIMD in 6-11-yr-old events.

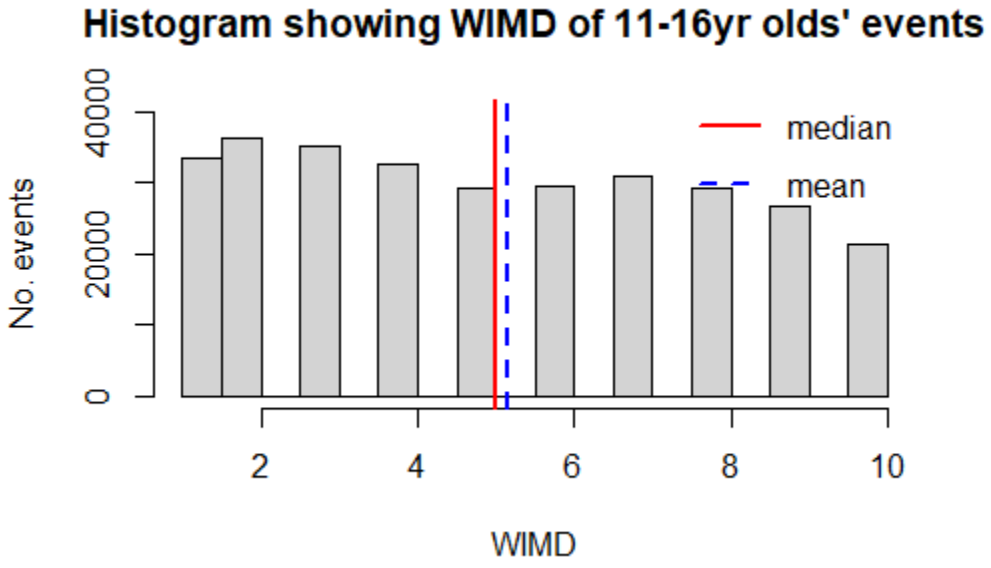

Figure 7f: Histogram showing distribution of deprivation as measured by WIMD in events of 11-16-yr-olds

Figure 7f shows a mostly uniform distribution of WIMD in 6-16-yr-old events. There is a small skew towards higher deprivation.

Table 4: Means and medians of BMI and WIMD per age group of individual patients and overall events

	2-6yrs		6-11yrs		11-16yrs	
	Patients	Events	Patients	Events	Patients	Events
Median BMI	16.2	16.2	17.8	16.8	21.2	21.1
Mean BMI	17.0	16.6	18.1	18.1	22.6	22.4 ($p=2.2x10^{-13}$)
Median WIMD	5	5	5	5	5	5
Mean WIMD	5.08	5.06	5.06	5.22 ($p=2.6x10^{-32}$)	5.04	5.15 ($p=7.2x10^{-20}$)

Events here mean every time a patient attends a GP it is recorded as an event. Event BMI/WIMD is the BMI/WIMD of the patient who attended the GP

p value denotes significance of difference between the mean patient BMI/deprivation, and the mean event BMI/deprivation.

Boxplots show BMI (figures 8a,b and c) and WIMD (figures 9a, b and c) of each comorbidity and overall cohort for children ages 2-6, 6-11, 11-16. Brackets show significance of all significant differences between means.

* <0.05, **<0.01, ***<0.001, ****<0.0001

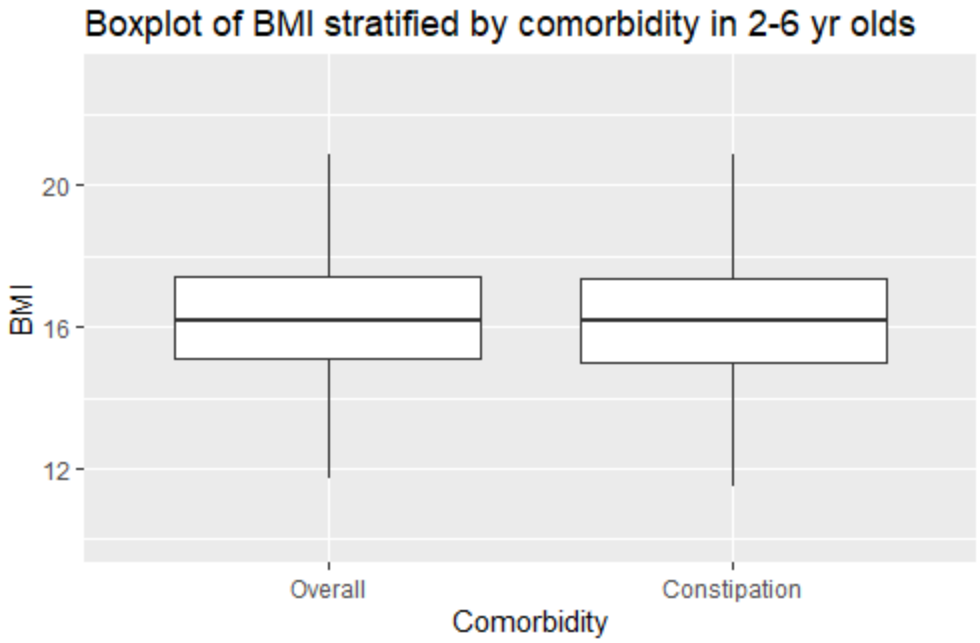

Figure 8a: Boxplot showing BMI of 2-6-yr-olds with constipation, and that of the overall cohort of 2-6-yr-olds

No significant difference in BMI between overall population and those with constipation in 2-6-year-olds.

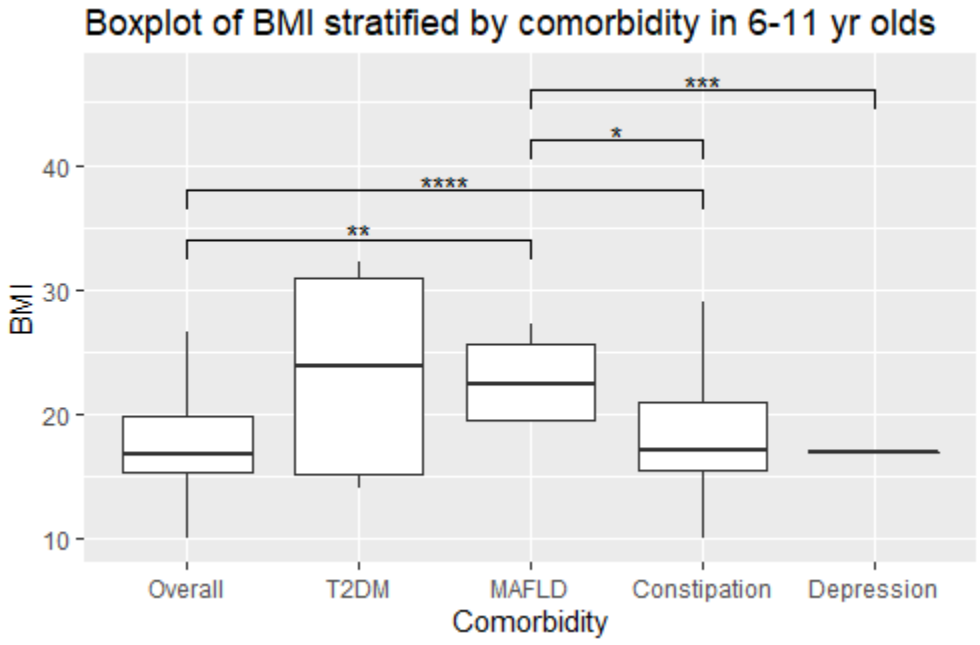

Figure 8b: Boxplot showing BMI of 6-11-yr-olds with each comorbidity, and that of the overall cohort of 6-11-yr-olds

BMI of children with MAFLD or constipation is significantly higher than that of the overall cohort in 6-11-year-olds (both p<0.01). BMI of those with MAFLD is significantly higher than those with constipation and depression (p=0.02, p<0.001 respectively).

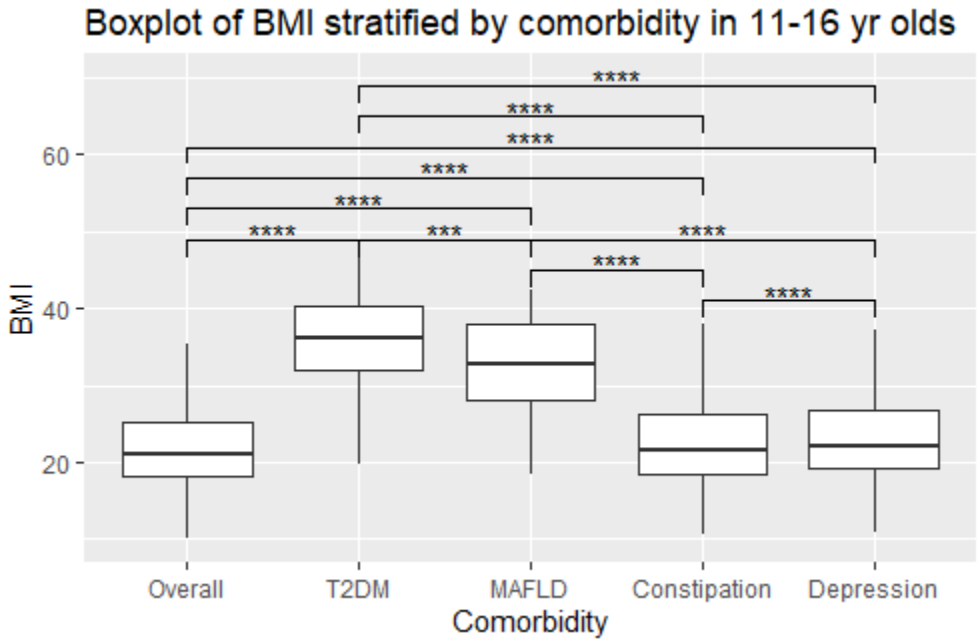

Figure 8c: Boxplot showing BMI of 11-16-yr-olds with each comorbidity, and that of the overall cohort of 11-16-yr-olds

In 11-16-yr-olds, all groups had very significantly different BMI from each other (all p<0.001). Order of groups in ascending BMI are: Overall cohort, constipation, depression, MAFLD, T2DM.

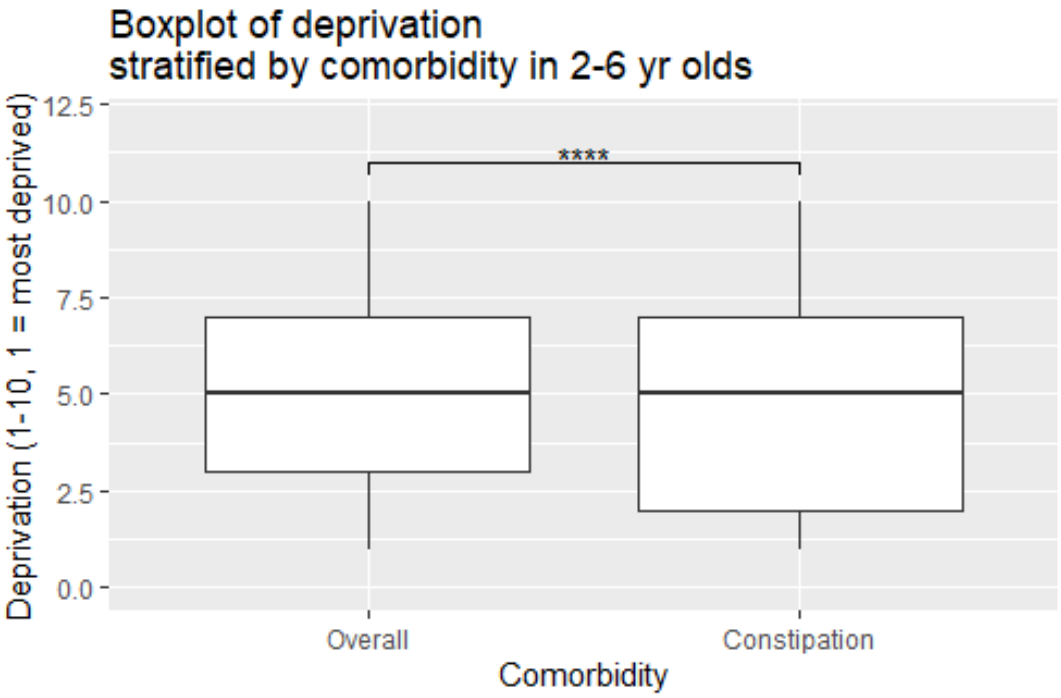

Figure 9a: Boxplot showing WIMD of 2-6-yr-olds with constipation, and that of the overall cohort of 2-6-yr-olds

2-6-year-olds with constipation are very significantly more deprived than the overall cohort ($p<0.0001$).

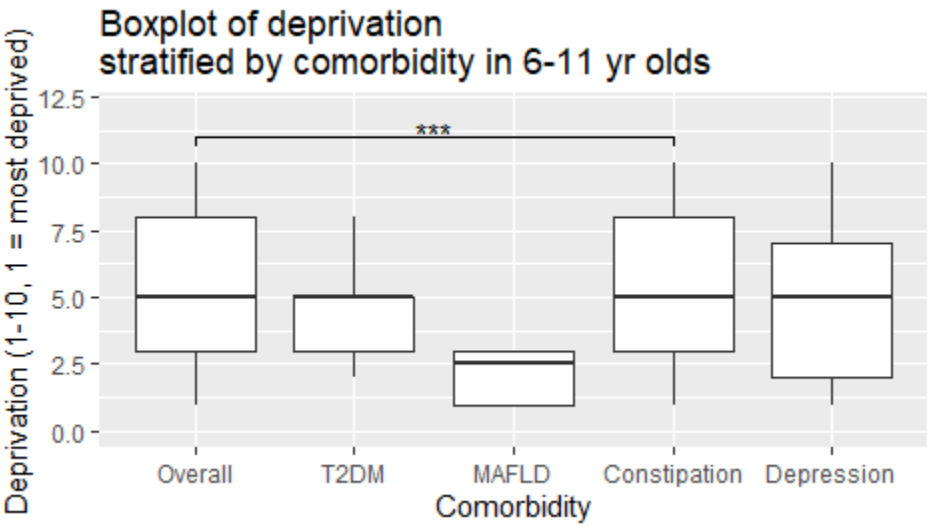

Figure 9b: Boxplot showing WIMD of 6-11-yr-olds with each comorbidity, and that of the overall cohort of 6-11-yr-olds

Only 6-11-year-olds with constipation were very significantly more deprived than the overall cohort ($p<0.001$).

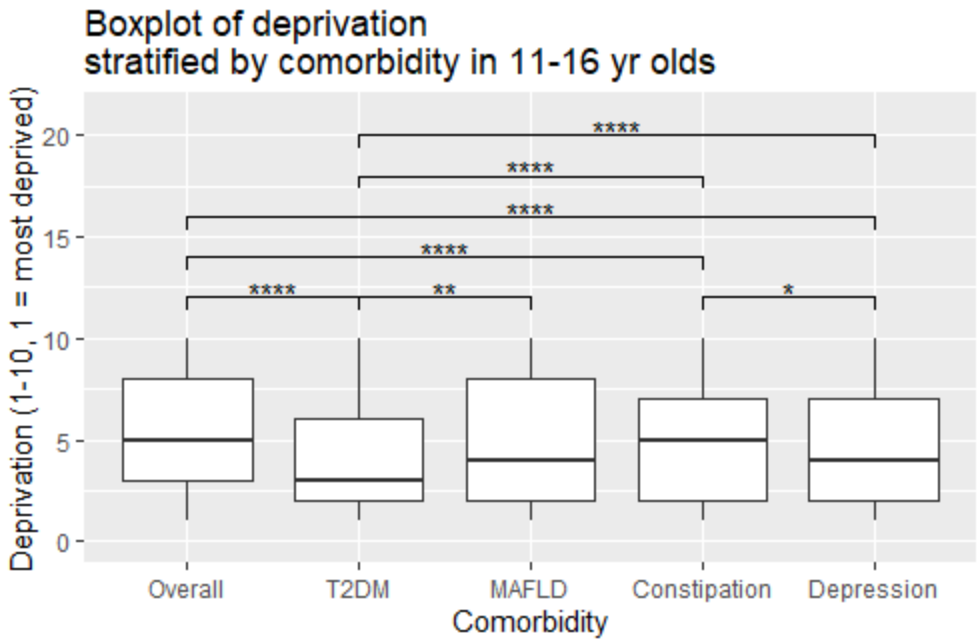

Figure 9c: Boxplot showing WIMD of 11-16-yr-olds with each comorbidity, and that of the overall cohort of 11-16-yr-olds

11-16-year-olds with T2DM, constipation, or depression are very significantly more deprived than the overall cohort of the age group (all $p<0.0001$). Those with T2DM were significantly less deprived than all other groups (all $p<0.01$), and those with depression are significantly more deprived than those with constipation ($p=0.022$).

Table 5: significance of relationship between comorbidities using χ^2 tests

11-16yrs / 6-11yrs	T2DM	MAFLD	Constipation	Depression
T2DM		p=0.87	$p=1.3\times10^{-9}$	$p=1.6\times10^{-7}$
MAFLD	p=0.97		p=0.59	p=0.047
Constipation	p=0.40	p=0.024		p=0.0011
Depression	p=0.97	p=0.97	p=0.32	

Constipation and depression were both very significantly more likely to co-occur with T2DM than random chance in 11-16-year-olds, as was depression with MAFLD and constipation.

The only comorbidities to significantly co-occur in 6-11-year-olds were constipation with MAFLD.

3.2) LgR model tests

Below shows ROC curves; AUC; optimised sensitivity, specificity, PPV, NPV, accuracy; and which predictor variables are significant (in order of significance), using CV for AUC, sensitivity, specificity, PPV, NPV, and accuracy. Exemplar confusion matrices for each model can be found in appendices 2-7.

N.B. Where year of birth is in brackets, this means it is significant in models when using it; the other predictor significances are not affected with and without year of birth.

***Table 6a: CV results for models for models predicting* T2DM or MAFLD, T2DM, MAFLD, and depression**

	T2DM or MAFLD		T2DM		MAFLD		Depression		
	11-16 yr-olds	Group 3	11-16 yr-olds	Group 3	11-16 yr-olds	Group 3	11-16 yr-olds	Group 2	Group 3
AUC	0.88	0.89	0.91	0.95	0.88	0.86	0.63	0.72	0.70
Sensitivity	0.82	0.84	0.77	0.72	0.85	0.79	0.77	0.68	0.63
Specificity	0.86	0.77	0.90	0.94	0.79	0.75	0.47	0.65	0.73
PPV	0.01	0.01	0.00	0.01	0.00	0.00	0.02	0.02	0.02
NPV	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
Accuracy	0.86	0.77	0.90	0.94	0.79	0.75	0.47	0.65	0.73
Significant Predictors (in order of significance)	BMI	BMI, Sex, (Year of birth)	BMI	BMI	BMI	(Year of birth), BMI	Sex, (year of birth)	WIMD, BMI	Sex, BMI, (year of birth)

Table 6b: CV results for models for models predicting constipation

Constipation	2-6-yr-olds	6-11-yr-olds	11-16-yr-olds	Group 1	Group 2	Group 3
AUC	0.54	0.53	0.53	0.53	0.59	0.56
Sensitivity	0.53	0.54	0.33	0.53	0.43	0.34
Specificity	0.54	0.53	0.72	0.54	0.75	0.80
PPV	0.14	0.09	0.06	0.11	0.08	0.10
NPV	0.89	0.93	0.96	0.93	0.97	0.96
Accuracy	0.54	0.53	0.70	0.54	0.74	0.77
Significant Predictors (in order of significance)	(Year of birth), Sex, BMI	Sex, (Year of birth), BMI	BMI, (year of birth), WIMD	(Year of birth), BMI	BMI	BMI, (Year of birth)

3.3) Results for specific changes to models

Table 6a: CV results for model predicting any comorbidity in 11-16-yr-olds

In 11-16-yr-olds	Any comorbidity
AUC	0.55
Sensitivity	0.55
Specificity	0.53
PPV	0.07
NPV	0.95
Accuracy	0.53
Significant Predictors (in order of significance)	BMI WIMD (year of birth) sex

Table 6b: CV results for model predicting T2DM or MAFLD in 11-16-yr-olds using BMI percentile

	Raw BMI (default)	Normalised BMI
AUC	0.88	0.88
Sensitivity	0.82	0.86
Specificity	0.86	0.80
PPV	0.01	0.01
NPV	1.00	1.00
Accuracy	0.86	0.80
Significant Predictors (in order of significance)	Raw BMI	Normalised BMI

Table 6c: CV results for model predicting MAFLD in group 3 using year of birth

	MAFLD predicted in group 3 using year of birth	
	- Year of Birth (default)	+ Year of Birth
AUC	0.86	0.88
Sensitivity	0.79	0.80
Specificity	0.75	0.85
PPV	0.00	0.01
NPV	1.00	1.00
Accuracy	0.75	0.85

3.4) Linear regression R-squared

r^2 of Linear regression predicting depression severity in group 3 = 0.11

4) Discussion

4.1) Summary

Diagnoses of T2DM, MAFLD, and depression are more prevalent in 11-16-yr-olds. T2DM, MAFLD, constipation, and depression are all linked with high BMI in 11-16-yr-olds; however, only MAFLD and constipation are significantly associated with high BMI in 6-11-yr-olds. Below the age of 6, there is no significant association between BMI and constipation incidence.

This trend is also apparent in deprivation as measured by WIMD. High deprivation is only significantly associated with T2DM and depression in 11-16-yr-olds, although it is significantly associated with constipation in all three groups, supporting current literature (Cagan Appak et al., 2017).

There were significant associations between comorbidities, with depression being significantly associated with all others in 11-16-yr-olds.

My models using BMI, deprivation, and sex to predict T2DM and MAFLD in 11-16-yr-olds were consistently very good (AUCs 0.86-0.95). Furthermore, T2DM and MAFLD can be predicted in

11-16-yr-olds using both BMI taken within the age group, and from when the children were 6-11yrs.

Interestingly, depression was far more accurately predicted in 11-16-yr-olds using BMI recordings from when they were 2-6yrs and 6-11yrs than from within the same age group (AUCs 0.72 & 0.70 *cf.* 0.63 respectively).

Despite high AUCs, PPVs were low, owing to the overall rarity of these comorbidities. It is important to consider this if models like these are to be used in the clinic to avoid unnecessary concern to children and families.

No models were able to accurately predict constipation (AUCs 0.53-0.59). Normalising BMI stratifying by age and sex did not affect model AUC.

I was unable to predict severity of depression in children using the number of GP visits with a very low r^2 value of 0.11, even using down-sampling.

4.2) Comparison to the literature

My results support the current literature on the effect of high BMI on risk of T2DM (Serbis et al., 2021), MAFLD (Jia et al., 2022; Lee et al., 2022), and depression (Lindberg et al., 2020; Rao et al., 2020). My results show that high BMI is a risk factor for constipation. This contradicts Koppen et al. (2016), but supports Pashankar & Loening-Baucke (2005). My results also support the current literature on the effect of high deprivation on the risk of T2DM (Catherine et al., 2021), constipation (Cagan Appak et al., 2017), and the lack of effect of deprivation of MAFLD (Orkin et al., 2020). I did not find any relevant literature regarding depression and deprivation, although Jolliff et al. (2021) noted higher risk of anxiety among those with lower socio-economic status.

T2DM and MAFLD did not significantly co-occur which contradicts previous research (Kosmalski et al., 2022).

This study produced models predicting T2DM with similarly high AUCs to other studies (Joshi & Dhakal, 2021; Tigga & Garg, 2020) some of which use routinely collected data (Farran et al., 2019; Ravaut et al., 2021; Xue et al., 2020), although my models have lower PPVs than some (Xue et al., 2020). My study also produced models predicting MAFLD with similarly high AUCs to others (Lee et al., 2021; Liu et al, 2021; Ji et al., 2022), although only Ji et al. (2022) used routinely collected data. I did not find any previous research using machine learning to predict depression or constipation. Out of all these studies, only Lee et al. (2021) predicted MAFLD in children, although BMI was not used as a predictor. All other studies were conducted on adults. Furthermore, these studies tended to use far more predictors than I used.

Unlike previous research, I found no significant association between T2DM and MAFLD (Kosmalski et al., 2022), but this may have been an aberration of the low frequency of these conditions in this study and/or the validity of routine GP data collection to ascertain the true frequency of both conditions. Particularly given that, in the paediatric cohort, T2DM is often

coded incorrectly (Rhodes et al., 2007) and MAFLD is particularly difficult to diagnose (Jia et al., 2022).

This study demonstrates that it is possible to develop similarly accurate models to be formed using routinely collected GP data in a paediatric cohort. Furthermore, I have made models predicting depression, a condition not yet predicted.

4.3) Strengths

It is likely that both T2DM and MAFLD will become a serious health concern in the paediatric population. A major strength of this study is the ability to access thousands of routinely collected patient records. This allows me to study these conditions while they are still rarely seen children. Furthermore, this data has been collected for approximately 20 years. This allows me to perform a longitudinal study predicting comorbidity as opposed to just classifying current comorbidity. These models have the potential to be developed further and used in a clinical setting, as well as the overall population without changing current protocols, due to the use of routinely collected data.

The small number of predictors used in this also helps keep these models accessible to many people, and minimises the risk for over-fitting. Furthermore, it has been shown that the use of predictors which require laboratory processing (like blood cholesterol levels) does not improve a classifier's ability to correctly identify patient risk beyond that of using routinely collected variables (like age and BMI) in similar regression models (Kariuki et al., 2017).

4.4) Limitations and potential solutions

The main limitation of this study is the paucity of positive training data in my cohort. These diseases are very rare in a paediatric cohort. More data would allow a larger training and test set which would increase confidence in model coefficients, and produce more reliable results. One way to increase this is to take patients' height and weight from separate events and calculate BMI myself, as well as using precalculated BMI, as there may be patients who have had these but not BMI recorded. Furthermore, I could further optimise the selection criteria to include more relevant patients. To include as many as patients as possible, I used a broad BMI cut-off range to exclude erroneous data. This however may mean that there are extreme erroneous values in the cohorts that may affect model accuracy.

I used GP event codes to flag patients as having a comorbidity or not. If a patient has attended the GP within a timeframe but has not attended regarding their comorbidity within said timeframe, they will be flagged as not having comorbidity. This is a limitation since patients often have conditions whilst being undiagnosed, as is common for MAFLD (Anderson et al., 2015); being diagnosed elsewhere (for example in a paediatric clinic); or incorrectly recorded, as is common with T2DM (Rhodes et al., 2007). Furthermore, conditions like depression and T2DM are highly stigmatised, and GPs may not want to give a formal diagnosis as to reduce risk of further harm to the child. To combat this, it may be useful to include event codes for prescriptions of anti-depressants as these may give a more representative view of children with

depressive episodes, as well as providing a larger set of positive patients, although risking mis-diagnosis due to off-label prescription (prescription for diseases not officially approved).

There are multiple versions of WIMD, depending on the year it was taken. In this study, I used one (2014) to represent all children. There are also children who moved house, hence changing their WIMD. Therefore, it would be more representative to use the corresponding WIMD score for each date of diagnosis, or to take an average until that point. Furthermore, using WIMD as a measure of deprivation is an ecological fallacy since it is based on LSOA and hence assume individuals in an area share deprivation levels (Pickrell et al., 2015). This is exacerbated by the use of the decile scale which limits machine learning where continuous variables are preferred.

4.5) Further research

A similar approach could be used to determine whether BMI could be used as a predictor for other conditions. Asthma and eczema are both examples of chronic childhood conditions which could be tested.

Furthermore, to improve performance of these models, future studies could add more predictor variables, for example blood pressure; family history (these comorbidities are known to run in families (Gov.wales, 2021); level of physical activity; and ethnicity (risk for T2DM is known to increase in certain ethnic groups (Serbis et al. 2021; Piscopo et al., 2005)), all of which are available in SAIL. It may be of use to attempt to create similar models using different machine learning techniques, allowing the model to better fit the data with non-linear trends.

4.6) Conclusion

These models predicting T2DM, MAFLD, and depression show great promise for use in a clinical setting, or for more accurately targeting costly interventions. MAFLD often goes undiagnosed, and there are not yet suitable screening methods, and the best of these are invasive (Jia et al., 2022). Therefore, this model could prove useful for identifying patients at risk of developing MAFLD, allowing us to then only screen these patients with further tests. These models output a true risk which can either be left quantitative or made qualitative using a cut-off. A qualitative cut-off may be useful in population screening, whereas true risk may be of more use in the clinic, where each child will receive their own risk, helping reduce unnecessary concern. Finally, these models allow us to identify patients who are at risk of developing these conditions up to 10 years before current diagnosis, allowing earlier interventions to both prevent and treat potentially serious consequences for both individuals and public health services; something of great value in this sector (Barrett et al., 2020; Rees et al., 2009).

References

Anderson, E. L., Howe, L. D., Jones, H. E., Higgins, J. P., Lawlor, D. A., & Fraser, A. (2015). The prevalence of non-alcoholic fatty liver disease in children and adolescents: a systematic review and meta-analysis. PloS one, 10(10), e0140908.

Apperley, L., Blackburn, J., Erlandson-Parry, K., Gait, L., Laing, P., & Senniappan, S. (2021). Childhood obesity: A review of current and future management options. Clinical Endocrinology, 96(3), 288-301. https://doi.org/10.1111/cen.14625

Barrett, T., Jalaludin, M., Turan, S., Hafez, M., & Shehadeh, N. (2020). Rapid progression of type 2 diabetes and related complications in children and young people—A literature review. Pediatric Diabetes, 21(2), 158-172. https://doi.org/10.1111/pedi.12953

Beynon, C., & Bailey, L. (2019). Prevalence of severe childhood obesity in Wales UK. Journal Of Public Health, 42(4), e435-e439. https://doi.org/10.1093/pubmed/fdz137

Cagan Appak, Y., Yalin Sapmaz, S., Dogan, G., Herdem, A., Ozyurt, B., & Kasirga, E. (2017). Clinical findings, child and mother psychosocial status in functional constipation. *The Turkish Journal Of Gastroenterology*, *28*(6), 465-470. https://doi.org/10.5152/tjg.2017.17216

Catherine, J., Russell, M., & Peter, C. (2021). The impact of race and socioeconomic factors on paediatric diabetes. *Eclinicalmedicine*, *42*, 101186. https://doi.org/10.1016/j.eclinm.2021.101186

Chadda, K., Cheng, T., & Ong, K. (2020). GLP-1 agonists for obesity and type 2 diabetes in children: Systematic review and meta-analysis. Obesity Reviews, 22(6). https://doi.org/10.1111/obr.13177

Chen, B. R., & Pan, C. Q. (2022). Non-invasive assessment of fibrosis and steatosis in pediatric non-alcoholic fatty liver disease. Clinics and Research in Hepatology and Gastroenterology, 46(1), 101755.

Dietz, W., & Bellizzi, M. (1999). Introduction: the use of body mass index to assess obesity in children. The American Journal Of Clinical Nutrition, 70(1), 123S-125S. https://doi.org/10.1093/ajcn/70.1.123s

Duncan-Jones A, Gibbon R, Jones M, Patterson B, Luker M, Hughes R et al. (2019) Obesity in Wales, Phw.nhs.wales. Retrieved 6 January 2022, from: https://phw.nhs.wales/topics/obesity/obesity-in-wales-report-pdf/

Dündar, İ., & Akıncı, A. (2022). Prevalence of type 2 diabetes mellitus, metabolic syndrome, and related morbidities in overweight and obese children. Journal Of Pediatric Endocrinology And Metabolism, 35(4), 435-441. https://doi.org/10.1515/jpem-2021-0271

El Khouli, R., Macura, K., Barker, P., Habba, M., Jacobs, M., & Bluemke, D. (2009). Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. *Journal Of Magnetic Resonance Imaging*, *30*(5), 999-1004. https://doi.org/10.1002/jmri.21947

Farran, B., AlWotayan, R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A., & Thanaraj, T. (2019). Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting

Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. *Frontiers In Endocrinology*, *10*. https://doi.org/10.3389/fendo.2019.00624

Ford, D., Jones, K., Verplancke, J., Lyons, R., John, G., & Brown, G. et al. (2009). The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Services Research*, *9*(1). https://doi.org/10.1186/1472-6963-9-157

Gov.wales. (2019). *An overview of the Healthy Child Wales Programme*. Gov.wales. Retrieved 23 April 2022, from https://gov.wales/sites/default/files/publications/2019-05/an-overview-of-the-healthy-child-wales-programme.pdf.

Gov.wales. (2021). *All Wales Weight Management Pathway 2021 (Children, Young People and Families): Core Components*. Gov.wales. Retrieved 28 April 2022, from https://gov.wales/sites/default/files/publications/2021-06/all-wales-weight-management-pathway-2021-children-young-people-and-families.pdf#.

HM Government. (2016). *Childhood Obesity A Plan for Action*. Assets.publishing.service.gov.uk. Retrieved 29 April 2022, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/546588/Childhood_obesity_2016__2__acc.pdf.

ICD. (2022). International Classification of Diseases (ICD). Who.int. Retrieved 8 May 2022, from https://www.who.int/standards/classifications/classification-of-diseases.

Jarvis, S., Giles, H., Jarvis, P., New, K. (2022). The weight status of children in late childhood within south East Wales and predictions for their future health. Journal of Public Health, fdac040, https://doi.org/10.1093/pubmed/fdac040

Ji, W., Xue, M., Zhang, Y., Yao, H., & Wang, Y. (2022). A Machine Learning Based Framework to Identify and Classify Non-alcoholic Fatty Liver Disease in a Large-Scale Population. *Frontiers In Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.846118

Jia, S., Zhao, Y., Liu, J., Guo, X., Chen, M., Zhou, S., & Zhou, J. (2022). Magnetic Resonance Imaging-Proton Density Fat Fraction vs. Transient Elastography-Controlled Attenuation Parameter in Diagnosing Non-alcoholic Fatty Liver Disease in Children and Adolescents: A Meta-Analysis of Diagnostic Accuracy. Frontiers In Pediatrics, 9. https://doi.org/10.3389/fped.2021.784221

Jolliff, A., Zhao, Q., Eickhoff, J., & Moreno, M. (2021). Depression, Anxiety, and Daily Activity Among Adolescents Before and During the COVID-19 Pandemic: Cross-sectional Survey Study. *JMIR Formative Research*, *5*(12), e30702. https://doi.org/10.2196/30702

Joshi, R., & Dhakal, C. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International Journal Of Environmental Research And Public Health*, *18*(14), 7346. https://doi.org/10.3390/ijerph18147346

Kariuki, J., Stuart-Shor, E., Leveille, S., Gona, P., Cromwell, J., & Hayman, L. (2017). Validation of the nonlaboratory-based Framingham cardiovascular disease risk assessment algorithm in the Atherosclerosis Risk in Communities dataset. *Journal Of Cardiovascular Medicine*, *18*(12), 936-945. https://doi.org/10.2459/jcm.0000000000000583

Kosmalski M, Ziółkowska S, Czarny P, Szemraj J, Pietras T. (2022) The Coexistence of Nonalcoholic Fatty Liver Disease and Type 2 Diabetes Mellitus. Journal of Clinical Medicine. 11(5):1375. https://doi.org/10.3390/jcm11051375

Koppen, I., Velasco-Benítez, C., Benninga, M., Di Lorenzo, C., & Saps, M. (2016). Is There an Association between Functional Constipation and Excessive Bodyweight in Children?. The Journal Of Pediatrics, 171, 178-182.e1. https://doi.org/10.1016/j.jpeds.2015.12.033

Lee, K. J., Moon, J. S., Kim, N. Y., & Ko, J. S. (2022). Effects of PNPLA3, TM6SF2 and SAMM50 on the development and severity of non-alcoholic fatty liver disease in children. Pediatric Obesity, 17(2), e12852.

Lee, L., Yen, J., Lu, H., & Liao, Y. (2021). Prediction of Nonalcoholic Fatty Liver Disease by Anthropometric Indices and Bioelectrical Impedance Analysis in Children. *Childhood Obesity*, *17*(8), 551-558. https://doi.org/10.1089/chi.2021.0054

Lindberg, L., Hagman, E., Danielsson, P., Marcus, C., & Persson, M. (2020). Anxiety and depression in children and adolescents with obesity: a nationwide study in Sweden. BMC Medicine, 18(1). https://doi.org/10.1186/s12916-020-1498-z

Liu, Y., Liu, X., Cen, C., Li, X., Liu, J., & Ming, Z. et al. (2021). Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. *Hepatobiliary &Amp; Pancreatic Diseases International*, *20*(5), 409-415. https://doi.org/10.1016/j.hbpd.2021.08.004

Lyons, R., Jones, K., John, G., Brooks, C., Verplancke, J., & Ford, D. et al. (2009). The SAIL databank: linking multiple health and social care datasets. *BMC Medical Informatics And Decision Making*, *9*(1). https://doi.org/10.1186/1472-6947-9-3

Magliano, D., Sacre, J., Harding, J., Gregg, E., Zimmet, P., & Shaw, J. (2020). Young-onset type 2 diabetes mellitus — implications for morbidity and mortality. Nature Reviews Endocrinology, 16(6), 321-331. https://doi.org/10.1038/s41574-020-0334-z

Must, A., & Anderson, S. (2006). Body mass index in children and adolescents: considerations for population-based applications. *International Journal Of Obesity*, *30*(4), 590-594. https://doi.org/10.1038/sj.ijo.0803300

Senedd.wales. *Childhood Obesity*. Senedd.wales. Retrieved 24 April 2022, from https://senedd.wales/media/ujodr5oo/child_obesity_booklet-english.pdf.

NHS Digital (2020). Part 3: Adult overweight and obesity, Digital.nhs.uk. Retrieved 10 April 2022, from https://digital.nhs.uk/data-and-

information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/england-2020/part-3-adult-obesity-copy.

NHS Digital, & Thandi, S. (2020). *Mental Health of Children and Young People in England, 2020: Wave 1 follow up to the 2017 survey - NHS Digital*. NHS Digital. Retrieved 24 April 2022, from https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2020-wave-1-follow-up.

NHS Digital. (2021). *National Child Measurement Programme - NHS Digital*. NHS Digital. Retrieved 24 April 2022, from https://digital.nhs.uk/services/national-child-measurement-programme/.

NICE. (2013). *Overview | Weight management: lifestyle services for overweight or obese children and young people | Guidance | NICE*. Nice.org.uk. Retrieved 10 April 2022, from https://www.nice.org.uk/guidance/ph47.

NICE guidelines (2022) Recommendations Obesity: identification, assessment and management Guidance, Nice.org.uk. Retrieved 10 April 2022, from https://www.nice.org.uk/guidance/cg189/chapter/1-recommendations.

Nicholas, M., Keown-Stoneman, C., Maguire, J., & Drucker, A. (2022). Association Between Atopic Dermatitis and Height, Body Mass Index, and Weight in Children. *JAMA Dermatology*, *158*(1), 26. https://doi.org/10.1001/jamadermatol.2021.4529

Nurko, S., & Zimmerman, L. A. (2014). Evaluation and treatment of constipation in children and adolescents. American family physician, 90(2), 82-90. https://doi.org/

Orkin, S., Brokamp, C., Yodoshi, T., Trout, A., Liu, C., & Meryum, S. et al. (2020). Community Socioeconomic Deprivation and Nonalcoholic Fatty Liver Disease Severity. *Journal Of Pediatric Gastroenterology &Amp; Nutrition*, *70*(3), 364-370. https://doi.org/10.1097/mpg.0000000000002527

Pashankar, D., & Loening-Baucke, V. (2005). Increased Prevalence of Obesity in Children With Functional Constipation Evaluated in an Academic Medical Center. Pediatrics, 116(3), e377-e380. https://doi.org/10.1542/peds.2005-0490

PHW. (2022). *Child Measurement Programme - Public Health Wales*. Phw.nhs.wales. Retrieved 24 April 2022, from https://phw.nhs.wales/services-and-teams/child-measurement-programme/.

Pickrell, W., Lacey, A., Bodger, O., Demmler, J., Thomas, R., & Lyons, R. et al. (2015). Epilepsy and deprivation, a data linkage study. *Epilepsia*, *56*(4), 585-591. https://doi.org/10.1111/epi.12942

Piscopo, M., Rigamonti, A., Chiesa, G., Bettini, S., Azzinari, A., & Bonfanti, R. et al. (2005). Type 2 Diabetes mellitus in Childhood. *Diabetes In Childhood And Adolescence*, 347-361. https://doi.org/10.1159/000085807

Rajindrajith S, Devanarayana NM, Crispus Perera BJ, Benninga MA., (2016) Childhood constipation as an emerging public health problem. World J Gastroenterol 2016; 22(30): 6864-6875 [PMID: 27570423 DOI: 10.3748/wjg.v22.i30.6864]

Rao, W., Zong, Q., Zhang, J., An, F., Jackson, T., & Ungvari, G. et al. (2020). Obesity increases the risk of depression in children and adolescents: Results from a systematic review and meta-analysis. Journal Of Affective Disorders, 267, 78-85. https://doi.org/10.1016/j.jad.2020.01.154

Ravaut, M., Harish, V., Sadeghi, H., Leung, K., Volkovs, M., & Kornas, K. et al. (2021). Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. *JAMA Network Open*, *4*(5), e2111315. https://doi.org/10.1001/jamanetworkopen.2021.11315

RCPCH. (2022). *UK-WHO growth charts - 2-18 years*. RCPCH. Retrieved 23 April 2022, from https://www.rcpch.ac.uk/resources/uk-who-growth-charts-2-18-years.

Rees, A., Thomas, N., Brophy, S., Knox, G., & Williams, R. (2009). Cross sectional study of childhood obesity and prevalence of risk factors for cardiovascular disease and diabetes in children aged 11–13. BMC Public Health, 9(1). https://doi.org/10.1186/1471-2458-9-86

SAIL. (2022). SAIL Databank - The Secure Anonymised Information Linkage Databank. Saildatabank.com. Retrieved 9 January 2022, from: https://saildatabank.com/.

Sanders, R., Han, A., Baker, J., & Cobley, S. (2015). Childhood obesity and its physical and psychological co-morbidities: a systematic review of Australian children and adolescents. European Journal Of Pediatrics, 174(6), 715-746. https://doi.org/10.1007/s00431-015-2551-3

Serbis, A., Giapros, V., Kotanidou, E., Galli-Tsinopoulou, A., & Siomou, E. (2021). Diagnosis, treatment and prevention of type 2 diabetes mellitus in children and adolescents. World Journal of Diabetes, 12(4), 344-365. https://doi.org/10.4239/wjd.v12.i4.344

Sutaria, S., Devakumar, D., Yasuda, S., Das, S., & Saxena, S. (2018). Is obesity associated with depression in children? Systematic review and meta-analysis. Archives Of Disease In Childhood, 104(1), 64-74. https://doi.org/10.1136/archdischild-2017-314608

Skinner, AC, Ravanbakht, SN, Skelton, JA, Perrin, EM, Armstrong, SC.(2018) Prevalence of obesity and severe obesity in US children, 1999–2016. Pediatrics 141:e20173459. https://doi.org/10.1542/peds.2017-3459.

Statswales. (2019). Welsh Index of Multiple Deprivation. Statswales.gov.wales. Retrieved 9 January 2022, from https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation.

Storz, M. A. (2020). The COVID-19 pandemic: an unprecedented tragedy in the battle against childhood obesity. Clinical and experimental pediatrics, 63(12), 477.

Tigga, N., & Garg, S. (2020). Predicting Type 2 Diabetes Using Logistic Regression. *Lecture Notes In Electrical Engineering*, 491-500. https://doi.org/10.1007/978-981-15-5546-6_42

UK Parliament. (2022). *Obesity Statistics*. commonslibrary.parliament.uk/. Retrieved 24 April 2022, from https://commonslibrary.parliament.uk/research-briefings/sn03336/.

Xue, M., Su, Y., Li, C., Wang, S., & Yao, H. (2020). Identification of Potential Type II Diabetes in a Large-Scale Chinese Population Using a Systematic Machine Learning Framework. *Journal Of Diabetes Research*, *2020*, 1-12. https://doi.org/10.1155/2020/6873891

Appendices

Appendix 1: Varying train:test split

Table A1: CV results for model predicting T2DM or MAFLD in 11-16-yr-olds with varying train:test splits

	0.7:0.3	0.75:0.25 (default)	0.8:0.2	0.9:0.1
AUC	0.88	0.88	0.89	0.90
Sensitivity	0.82	0.82	0.86	0.78
Specificity	0.85	0.86	0.83	0.89
PPV	0.01	0.01	0.01	0.01
NPV	1.00	1.00	1.00	1.00
Accuracy	0.85	0.86	0.83	0.88

Appendix 2: Exemplar confusion matrix for 2-6-yr-olds

Table A2: Confusion matrix for predicting constipation in 2-6-yr-olds

Constipation	Observed negative	Observed positive
Predicted negative	2452	289
Predicted positive	3805	572

Appendix 3: Exemplar confusion matrix for 6-11-yr-olds

Table A3: Confusion matrix for predicting constipation in 6-11-yr-olds

Constipation	Observed negative	Observed positive
Predicted negative	7180	598
Predicted positive	3393	310

Appendix 4: Exemplar confusion matrices for 11-16-yr-olds

Numbers are rounded to the nearest 10 to mask individuals as per SAIL policy

Table A4a: Confusion matrix for predicting T2DM or MAFLD in 11-16-yr-olds

T2DM or MAFLD	Observed negative	Observed positive
Predicted negative	13340	0
Predicted positive	2420	30

Table A4b: Confusion matrix for predicting T2DM in 11-16-yr-olds

T2DM	Observed negative	Observed positive
Predicted negative	15060	0
Predicted positive	720	0

Table A4c: Confusion matrix for predicting MAFLD in 11-16-yr-olds

MAFLD	Observed negative	Observed positive
Predicted negative	13300	0
Predicted positive	2470	10

Table A4d: Confusion matrix for predicting constipation in 11-16-yr-olds

Constipation	Observed negative	Observed positive
Predicted negative	8770	400
Predicted positive	6240	370

Table A4e: Confusion matrix for predicting depression in 11-16-yr-olds

Depression	Observed negative	Observed positive
Predicted negative	7480	50
Predicted positive	80	120

Appendix 5: Exemplar confusion matrix for Group 1

Table A5: Confusion matrix for predicting constipation in Group 1

Constipation	Observed negative	Observed positive
Predicted negative	2343	187
Predicted positive	2462	274

Appendix 6: Exemplar confusion matrices for Group 2

Table A6a: Confusion matrix for predicting constipation in group 2

Constipation	Observed negative	Observed positive
Predicted negative	2479	86
Predicted positive	353	40

Table A6b: Confusion matrix for predicting depression in group 2

Depression	Observed negative	Observed positive
Predicted negative	2232	6
Predicted positive	712	8

Appendix 7: Exemplar confusion matrices for Group 3

Numbers are rounded to the nearest 10 to mask individuals as per SAIL policy

Table A7a: Confusion matrix for predicting T2DM or MAFLD in group 3

T2DM or MAFLD	Observed negative	Observed positive
Predicted negative	6190	0
Predicted positive	2650	10

Table A7b: Confusion matrix for predicting T2DM in group 3

T2DM	Observed negative	Observed positive
Predicted negative	8350	0
Predicted positive	490	0

Table A7c: Confusion matrix for predicting MAFLD in group 3

MAFLD	Observed negative	Observed positive
Predicted negative	6190	0
Predicted positive	2650	10

Table A7d: Confusion matrix for predicting constipation in group 3

Constipation	Observed negative	Observed positive
Predicted negative	7480	330
Predicted positive	900	130

Table A7e: Confusion matrix for predicting depression in group 3

Depression	Observed negative	Observed positive
Predicted negative	5740	30
Predicted positive	3010	70

Appendix 8: Link to code

https://github.com/whmidgley/Evaluating-risk-factors-for-common-childhood-comorbidities-An-all-Wales-longitudinal-cohort-study.

Appendix 9: SAIL Databank Information Governance Research Panel Application

1. Applicant Details

Project Lead

PROJECT LEAD

Arron Lacey

JOB TITLE

Honorary Lecturer

ORGANISATION

Swansea University

EMAIL ADDRESS

a.s.lacey@swansea.ac.uk

Collaborating Organisations

Key contacts from any collaborating organisations:

Data Access Applicants

All applicants who will require access to the data

NAME

Will Midgley

EMAIL ADDRESS

1900481@swansea.ac.uk

JOB TITLE

Student SUMS

ORGANISATION

Swansea University

2.Key Information

PROVIDE DETAILS ON WHO IS COMMISSIONING THE PROJECT

Swansea University 3rd Year Dissertation Project

PROVIDE A PROSPECTIVE START DATE FOR THE WORK INVOLVING SAIL

17/01/2022

PROVIDE ANTICIPATED END DATE OF THE PROJECT

17/01/2023

Permissions

Relevant permissions which have been obtained or that are being sought:

Independent Peer Review

PLEASE PROVIDE INFORMATION ON THE RELEVANT PERMISSIONS YOU HAVE OBTAINED OR THAT ARE BEING SOUGHT

Not required

IF YOU HAVE TICKED 'NOT REQUIRED' PLEASE SPECIFY THE REASONS:

This is a student project to which peer review publication will not be intended for this project

Research Ethics

PLEASE PROVIDE INFORMATION ON THE RELEVANT PERMISSIONS YOU HAVE OBTAINED OR THAT ARE BEING SOUGHT

Not required

PLEASE STATE THE NAME OF THE COMMITTEE THAT IS BEING APPLIED TO/HAS GIVEN APPROVAL, AS APPLICABLE

Not Required

THE PROJECT USES WILL USE ONLY ANONYMISED DATA, AND THEREFORE RESEARCH ETHICS REVIEW IS NOT REQUIRED

No

PLEASE SPECIFY WHY RESEARCH ETHICS PERMISSION IS NOT REQUIRED:

This project will only used anonymously linked data at the SAIL databank and therefore is exempt from needing formal ethical approval as covered by the SAIL Databank's Terms and Conditions.

Funding

TOTAL AMOUNT OF FUNDING AWARDED

750

PLEASE PROVIDE THE SOURCE OF THE FUNDING

Swansea University

3.Description

AREA OF RESEARCH

Child Health

PROVIDE A LAY SUMMARY OF THE PROJECT

Obesity is a leading public health concern in Wales, with one in eight reception age children being obese (https://phw.nhs.wales/topics/obesity/obesity-in-wales2019/). This increases their risk for comorbid conditions like type 2 diabetes, constipation, fatty liver, and mental health issues, the latter of which is another condition which has increased over the last few years (https://digital.nhs.uk/dataand-information/publications/statistical/mental-health-of-children-and-youngpeople-in-england/2020-wave-1-follow-up). We plan on conducting research into the social demographics of children diagnosed with obesity in Wales, and its correlation with secondary comorbidities such as type 2 diabetes, constipation, fatty liver disease, and mental health issues. Along with opensource government postcode statistics (i.e. LSOAs only), we would like to use the SAIL databank to look for any relationships between these factors and obesity using statistical and machine learning techniques.

PROVIDE THE AIM OF THE PROJECT, INCLUDING ANTICIPATED OUTCOMES

The aim of this project is to fulfil the requirements for a 3rd year dissertation at Swansea University Medical School. The scope of the project will entail exploring and quantifying any difference in comorbidities between children who are not classes as obese and those who are. Using machine learning we may also try to predict which children do go onto develop comorbidities and look for latent factors that contribute to this. if there is an opportunity to publish from this work we would aim to submit to Archives of Diseases in Childhood in the first instance.

PROVIDE AN OUTLINE OF YOUR ANALYSIS PLAN INCLUDING THE ANTICIPATED OUTPUTS

1. We will obtain BMI (and height/weight) data from multiple datasets (GP, National Community Child Health) and note that the Child Measurement Programme for Wales mention the ability to link these data to SAIL so we will explore this too. http://www.wales.nhs.uk/sitesplus/documents/888/Child%20Measurement%20report%20%28Eng%29.pdf 2. We will obtain comorbidity data for type 2 diabetes, constipation, fatty liver, neurology and mental health issues from primary and secondary care datasets in SAIL and create a case-control cohort of obese vs nonobese children matched on age, sex, birth weight, gestational age, maternal age, Welsh Index of Multiple Deprivation and year of birth. Birth data will be obtained from the National Community Child Health dataset. If we aren't able to achieve at least a 1:1 match on obese vs not obese then we will use all children and matching variables as model covariates instead. 3. We will use logistic regression to model the prevalence of comorbidities between the two groups, and a cox-proportional hazards model to determine

time to event of incident comorbidity status. We will use machine learning techniques (random forest, logistic regression, naive bases) to predict future risk of developing comorbidities where we would define the risk window for obese children to be up to 1 year before a healthcare record to accommodate potential time for known obesity and up to 5 years after the healthcare record. The latter approach may help us identify latent factors and groups within both obese and non-obese children to may contribute to developing comorbidites.

4.Dissemination and Impact

PROVIDE AN OUTLINE OF THE PUBLIC ENGAGEMENT STRATEGY FOR THE STUDY, OR A BRIEF EXPLANATION WHY THERE IS NOT PUBLIC ENGAGEMENT

Given that this will be a 4 month 3rd year student dissertation project, there won't be a public engagement strategy. However should potential for a publication arise towards the end of the project we will apply for a new IGRP or extension and propose a strategy.

PLEASE INDICATE YOUR PLANS FOR PUBLISHING THE RESULTS OF YOUR PROJECT E.G. TARGET JOURNAL OR INTENDED RECIPIENTS OF REPORT

Should potential for publication arise from this project we would aim to submit to Archives in Diseases in Childhood.

PLANS TO GENERATE IMPACT

This project aims to provide evidence of the risk of obesity and associated outcomes in children in Wales as outlined as a major priority in the 10 year Healthy Weight: Healthy Wales initiative https://gov.wales/sites/default/files/publications/2019-10/healthy-weight-healthywales-youth-community.pdf Should this project provide insights that are aligned with the Health Weight: Healthy Wales initiative we would seek to present these results to them so that this work can add to any evidence required for policy change in Wales.

ARE THE RESULTS/METHODS DEVELOPED LIKELY TO HAVE OTHER POTENTIAL APPLICATIONS?

None

WHAT ARE THE POTENTIALLY SENSITIVE ISSUES THAT NEED TO BE TAKEN INTO ACCOUNT WHEN PUBLICISING THE FINDINGS OF THE PROJECT?

We do not foresee any potential sensitive issue beyond working within SAIL guidelines to ensure that patients anonymity is preserved.

5.Data Requested

Requested Study Data Scope

TIME PERIOD FOR WHICH THE DATA IS REQUESTED FROM

Earliest available date

TO 30/06/2022 00:00:00

GEOGRAPHIC AREA FOR WHICH DATA IS REQUESTED

All-Wales

DEMOGRAPHIC REQUIREMENTS

GENDER:

Any

AGE

All available ages

OTHER RESTRICTIONS

N/A

SAIL Data Sets

DATASET NAME

NCCH - National Community Child Health Database

PLEASE LIST THE INFORMATION NEEDED FROM THIS DATASET

Maternal ALF_E, birth weight, maternal age, gestational age, birth weight, BMI/Height/Weight measurements

DATA SCOPE

Study data scope

DATASET NAME

WDSD - Welsh Demographic Service

PLEASE LIST THE INFORMATION NEEDED FROM THIS DATASET

Address

DATA SCOPE

Study data scope

DATASET NAME

PEDW - Patient Episode Database for Wales

PLEASE LIST THE INFORMATION NEEDED FROM THIS DATASET

Healthcare records pertaining to comorbidities listed in analysis plan

DATA SCOPE

Study data scope

DATASET NAME

WLGP - GP Primary Care – Audit

PLEASE LIST THE INFORMATION NEEDED FROM THIS DATASET

Healthcare records pertaining to comorbidities listed in analysis plan

DATA SCOPE

Study data scope

External Datasets

Sensitive Data Items

PLEASE LIST ANY SENSITIVE DATA ITEMS REQUIRED WITH JUSTIFICATION FOR THEIR REQUIREMENT.

Data provision schedule

Once at the beginning of the project