



Effect of timeframes to define long term conditions and sociodemographic factors on prevalence of multimorbidity using disease code frequency in primary care electronic health records: retrospective study

Thomas Beaney ,^{1,2} Jonathan Clarke,² Thomas Woodcock,¹ Azeem Majeed ,¹ Mauricio Barahona,² Paul Aylin¹

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjmed-2022-000474>).

For numbered affiliations see end of article.

Correspondence to: Dr Thomas Beaney, Department of Primary Care and Public Health, Imperial College London, London, UK; thomas.beaney@imperial.ac.uk

Cite this as: *BMJ MED* 2024;3:e000474. doi:10.1136/bmjmed-2022-000474

Received: 16 January 2023
Accepted: 12 December 2023

ABSTRACT

OBJECTIVE To determine the extent to which the choice of timeframe used to define a long term condition affects the prevalence of multimorbidity and whether this varies with sociodemographic factors.

DESIGN Retrospective study of disease code frequency in primary care electronic health records.

DATA SOURCES Routinely collected, general practice, electronic health record data from the Clinical Practice Research Datalink Aurum were used.

MAIN OUTCOME MEASURES Adults (≥18 years) in England who were registered in the database on 1 January 2020 were included. Multimorbidity was defined as the presence of two or more conditions from a set of 212 long term conditions.

Multimorbidity prevalence was compared using five definitions. Any disease code recorded in the electronic health records for 212 conditions was used as the reference definition. Additionally, alternative definitions for 41 conditions requiring multiple codes (where a single disease code could indicate an acute condition) or a single code for the remaining 171 conditions were as follows: two codes at least three months apart; two codes at least 12 months apart; three codes within any 12 month period; and any code in the past 12 months. Mixed effects regression was used to calculate the expected change in multimorbidity status and number of long term conditions according to each definition and associations with patient age, gender, ethnic group, and socioeconomic deprivation.

RESULTS 9718573 people were included in the study, of whom 7183662 (73.9%) met the definition of multimorbidity where a single code was sufficient to define a long term condition. Variation was substantial in the prevalence according to timeframe used, ranging from 41.4% (n=4023023) for three codes in any 12 month period, to 55.2% (n=5366285) for two codes at least three months apart. Younger people (eg, 50-75% probability for 18-29 years v 1-10% for ≥80 years), people of some minority ethnic groups (eg, people in the Other ethnic group had higher probability than the South Asian ethnic group), and people living in areas of lower socioeconomic deprivation were more likely to be re-classified as not multimorbid when using definitions requiring multiple codes.

CONCLUSIONS Choice of timeframe to define long term conditions has a substantial effect on the prevalence of multimorbidity in this nationally representative sample. Different timeframes affect prevalence for some people more than others, highlighting the need to consider the impact of bias in the choice of method when defining multimorbidity.

Introduction

Multimorbidity is defined as the co-existence of two or more long term conditions in the same person and is a priority for health systems globally.^{1,2} The number of people living with multimorbidity is growing, partly due to population ageing, with a

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Multimorbidity is defined as the presence of two or more long term conditions
- ⇒ Large differences in which conditions are included in this definition exist
- ⇒ How the time periods used to define a long term condition based on the frequency of disease codes in electronic health record data affect multimorbidity prevalence is unknown

WHAT THIS STUDY ADDS

- ⇒ Comparing five timeframes to define 41 of a set of 212 long term conditions, where presence of a single disease code in the electronic health record might represent an acute condition, substantial variation was shown in the prevalence of multimorbidity
- ⇒ Multimorbidity prevalence ranged from 41.4% (three codes required within any 12 month period) to 73.9% (if a single code was sufficient for diagnosis); using diseases recorded by clinicians in the electronic health record as active problems, multimorbidity prevalence was only 35.2%
- ⇒ Younger people, people in some minority ethnic groups, and people living in areas of lower socioeconomic deprivation were more likely to be re-classified as not multimorbid under timeframes requiring more than one code

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE, OR POLICY

- ⇒ Substantial variation was noted in multimorbidity prevalence according to the timeframe used to define a long term condition, highlighting the challenges in the direct comparison of estimates between studies
- ⇒ Recommendations are provided for researchers applying different timeframes for counting long term conditions using electronic health record data
- ⇒ Definitions of long term conditions requiring multiple codes may introduce bias, and researchers should provide a rationale for their choice of timeframe and consider sensitivity analyses to explore the impact on different patient groups

recent UK based study finding an increase in prevalence in adults from 31% in 2004 to 53% in 2019 and with higher incidence associated with increased socioeconomic deprivation.³ Although most studies define multimorbidity as the presence of two or more long term conditions,⁴ no consensus exists as to which conditions should be included. One systematic review identified a range from two to 285 separate conditions included in multimorbidity measures,⁴⁻⁶ making comparison between prevalence estimates challenging.

A further challenge to estimating multimorbidity prevalence is to distinguish acute from chronic presentations, particularly when using routinely collected electronic health record data. Most commonly, conditions are characterised as 'chronic' based on duration,⁶ ranging from at least three or 12 months, to lifelong.⁶⁻⁸ Although for many conditions, such as ischaemic heart disease, a single diagnostic code indicates a lifelong condition, for others, such as gastritis, a single code may not differentiate an acute episode from chronic disease, or active from inactive disease. Researchers using electronic health record data have used different approaches to distinguish these: some studies have required multiple entries of a diagnostic code spaced over time for particular diseases, while others have made use of prescriptions of related medications within a specific time period.^{3,9} A potential alternative to using occurrences of diagnostic codes is to use conditions that a clinician has classified as being a problem within the healthcare record. These conditions are recorded separately from clinical diagnoses and are used by clinicians as a summary of a patient's active clinical problems.¹⁰

Whether the prevalence of multimorbidity is substantially affected by the choices over the use of

diagnosis or problem codes is unclear, as is the timeframe for determining whether a person has a long term condition and whether definitions may impact some groups of people differently. The aim of this paper, therefore, is to determine the extent to which different strategies that are used to define multimorbidity in the primary care electronic health record affect the prevalence of multimorbidity overall and between different demographic groups. We used a range of timeframes from previous multimorbidity research to define long term conditions, and applied these to a large, nationally representative sample of patients in the primary care electronic health record in England.

Figure 1 shows the visual abstract.

Materials and methods

Data source

We used data provided from the Clinical Practice Research Datalink Aurum dataset, which includes routinely collected, patient level data from general practices using EMIS health web software in England.¹¹ We included all patients who were deemed research acceptable (a marker used by the Clinical Practice Research Datalink for patients meeting specific quality criteria¹²) registered to a general practice on 1 January 2020 and aged 18 years or older with at least one year of continuous practice registration. Observation data for each patient are recorded in the form of Medcodes, along with the date of observation and entry date. Any observation codes occurring after the 1 January 2020 were excluded and we included all historical observation codes. Data were linked to the 2019 Index of Multiple Deprivation of a patient's area of residence (categorised into tenths), which was used as a patient level marker of socioeconomic deprivation.¹³

Diagnostic codes and data cleaning

Multimorbidity was defined as the presence of two or more chronic conditions from a set of 212 conditions.⁴ We included 211 conditions selected by Head and colleagues, based on a set of 308 disease phenotypes developed by Kuan and colleagues.^{3,14} We reviewed these code lists and made changes for diabetes disease lists (online supplemental appendix) and also included "chronic primary pain" as a new condition. In the study by Head and colleagues, most conditions were defined by a diagnostic code ever being recorded, but for some conditions, the authors required three occurrences of a code within a 12 month period. Two clinicians (TB and JC) reviewed the codes and reassigned 41 diseases to require multiple codes where a single code might

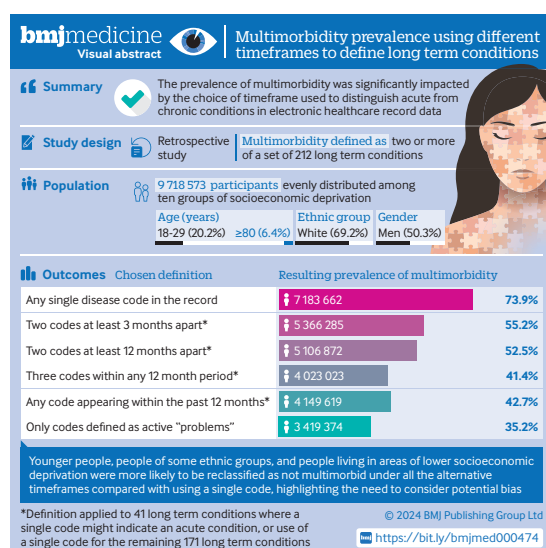


Figure 1 | Visual abstract

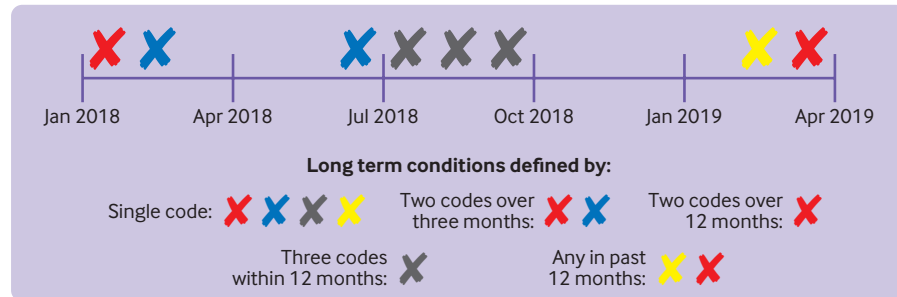


Figure 2 | Schematic diagram of assignment of a long term condition for a hypothetical patient under each alternative definition. Each cross represents a code for a condition in a single patient's electronic health record timeline from January 2018 to April 2019, with the colour representing the disease. The single code means a single code was sufficient for diagnosis results in all four long term conditions being counted. Two codes over three months and two codes over 12 months required at least two codes separated by at least three months and 12 months, respectively. Three within 12 months' requires three codes occurring within a 12 month time window. Any in past 12 months requires at least one code occurring in the past 12 months (measured in the 12 months from 1 January 2019 to 1 January 2020. Diseases defined by problems not included here

otherwise indicate an acute condition (see online supplemental appendix for further details and rationale). For diabetes, we assigned the most likely diabetes type (type 1, type 2, or other/unspecified) based on the frequency of codes. Further information on the long term condition categories and cleaning of ethnic group codes are given in the online supplemental appendix. Our code lists are available from: <https://tbeaney.github.io/MMclustering/>

Disease definitions

The number of chronic conditions for each person was calculated as of the 1 January 2020 under each of the following five definitions (figure 2). These definitions were selected to reflect the variety used in previous multimorbidity research and definitions of chronic conditions:

1. Any disease code in the record for all 212 conditions, which was used as the reference definition (single code);¹⁴
2. two codes at least three months apart for the 41 conditions requiring multiple codes,⁷ or a single code for the remaining 171 conditions;
3. two codes at least 12 months apart for the 41 conditions requiring multiple codes,^{7 8} or a single code for the remaining 171 conditions;
4. three codes within any 12 month period for the 41 conditions requiring multiple codes,³ or a single code for the remaining 171 conditions; and
5. any code appearing within the past 12 months for the 41 conditions requiring multiple codes,⁹ or a single code for the remaining 171 conditions.

We also compared a definition of multimorbidity only including conditions defined as active problems (referred to here as problems).

Diagnostic codes entered into the record can be marked by clinicians as problems, which then appear on a patient's summary during a consultation.¹¹ These may be more visible to the clinician than historical diagnostic codes, which would require manual searching through the electronic notes and are also pulled across onto referral letters.¹⁰ Problems are recorded as either active or past problems and can be marked as ended if no longer active, with the end date recorded.¹¹ We included only those problems that were active, that is, those that did not have an associated end date before 1 January 2020 and were marked as active. The problem definition was applied to all 212 long term conditions and not only the 41 requiring multiple codes. Further details are given in the online supplemental appendix page 10.

Statistical analysis

We calculated the prevalence of patients with one, two, or three or more chronic conditions based on each definition, stratified by age group, gender, ethnic group, and socioeconomic deprivation. Age (years) was categorised into 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80 or more. For gender, we used the categories defined in Clinical Practice Research Datalink, as male, female, or indeterminate (where gender cannot be classified as male or female). No data for age or gender were missing due to our inclusion criteria and Clinical Practice Research Datalink's definition of research acceptable patients. In all analyses, missing data were assigned as categories to ensure no patients were excluded.

To calculate associations with the expected probability of a change in having multimorbidity under each definition, we used mixed effects logistic regression. We included only patients with multimorbidity defined using the single code definition and created a

Table 1 | Characteristics of study participants by mean number of conditions and prevalence of multimorbidity based on a single code definition of 212 long term conditions

Participant characteristics	Total, %	Mean number of long term conditions	Prevalence with one or more long term conditions, %	Prevalence with multimorbidity, %
Age (years):				
18-29	1 958 414 (20.2)	1.9	71.6	48.5
30-39	1 580 260 (16.3)	2.6	76.8	57.3
40-49	1 806 547 (18.6)	3.8	86.4	72.6
50-59	1 504 086 (15.5)	5.5	93.4	85.7
60-69	1 255 632 (12.9)	7.1	96.5	92.4
70-79	987 576 (10.2)	9.0	98.5	96.7
≥80	626 058 (6.4)	11.2	99.0	98.1
Gender:				
Female	4 831 012 (49.7)	5.3	89.4	78.5
Indeterminate	205 (<0.1)	4.2	88.3	74.1
Male	4 887 356 (50.3)	4.5	83.2	69.4
Ethnic group:				
White	6 721 471 (69.2)	5.5	90.2	79.6
South Asian	699 376 (7.2)	4.4	82.0	68.5
Black	390 621 (4.0)	4.0	83.8	68.8
Other	251 218 (2.6)	2.4	60.6	43.5
Mixed	132 957 (1.4)	3.4	79.7	62.8
Missing	1 522 930 (15.7)	3.3	76.4	58.6
Index of Multiple Deprivation group:				
1 (least deprived)	1 033 628 (10.6)	5.0	89.3	77.6
2	941 864 (9.7)	5.1	89.1	77.4
3	990 212 (10.2)	5.0	87.2	75.3
4	986 174 (10.1)	5.1	87.7	75.8
5	923 682 (9.5)	4.8	85.7	73.2
6	980 715 (10.1)	4.8	85.2	72.6
7	1 018 134 (10.5)	4.7	83.8	70.8
8	1 007 086 (10.4)	4.6	83.3	70.0
9	973 478 (10.0)	4.8	85.0	72.0
10 (most deprived)	855 807 (8.8)	5.2	86.8	74.7
Missing	7793 (0.1)	4.1	79.8	65.4
Total	9 718 573	4.9	86.3	73.9

binary outcome variable as 1 if multimorbidity status changed and 0 if multimorbidity status did not change, compared with the single code definition. We included age, gender, ethnic group, and Index of Multiple Deprivation as covariates in the models and random intercepts for general practice to account for clustering and potential coding differences between practices. We then calculated the marginal probability of a change in multimorbidity status overall, and separately for each covariate, holding other covariates at their mean values and assuming random effects of zero.

To calculate associations with the change in expected total count of conditions under each definition compared with the single code definition, we used mixed effects negative binomial regression. A negative binomial model was chosen because the conditional variance was found to be significantly larger than the conditional mean for each model ($P < 0.001$) making a Poisson model unsuitable.¹⁵ Models included the same covariates and random effects as in the mixed effects logistic regression models. We then calculated the marginal expected

change in the count of long term conditions overall and separately for each covariate, assuming other covariates in the model were held at their mean values and random effects of zero. All regression models assumed the single code definition as the reference group, given that this method will always result in the highest estimated prevalence. Further details, including equations for both models, are given in the online supplemental appendix.

Python version 3.10.6 and Pandas version 1.4.3 were used in data processing and plots and Stata version 17.0 was used for regression analyses.

Results

We included 9 718 573 adults aged 18 years or older and registered in the Clinical Practice Research Datalink Aurum on 1 January 2020 in the study. Table 1 shows the demographic characteristics of patients included. 20.2% of participants were aged 18-29 years and 6.4% were 80 years or older. Gender was almost equally split between women and men (49.7% and 50.3%, respectively) with less than 0.1% recorded in the Clinical

Practice Research Datalink as indeterminate gender. Most people (69.2%) were of white ethnicity, with 15.7% missing an ethnic group recording. Study participants were evenly distributed among the ten Index of Multiple Deprivation groups, but with relatively fewer (8.8%) in the most deprived tenth. Based on the single code definition (any code representing one of the 212 long term conditions present in the record), the number of long term conditions per person and prevalence of multimorbidity increased with age, and were higher in women and people of white ethnicity. Multimorbidity prevalence was highest in people living in the least deprived Index of Multiple Deprivation group (77.6%), with a declining prevalence up to the eighth group (70.0%) but with subsequently higher prevalence in the two most deprived groups (72.0% and 74.7%).

Multimorbidity prevalence

The prevalence of people with one or more, two or more, or three or more long term conditions varied substantially according to definition used (table 2). Under the single code definition, the prevalence of multimorbidity (two or more long term conditions) was 73.9%, reducing to 41.4% using the three codes within 12 months' definition and 35.2% using only conditions recorded as active in the problem list.

Of the 41 conditions included that were required to have multiple codes, elevated concentrations of total cholesterol, triglycerides, and low density lipoprotein; dermatitis; and enthesopathy and synovial disorders were the most common conditions, with a prevalence of at least 20% using the single code definition (online supplemental appendix figure A1). The prevalence of most of these conditions was substantially lower with each of the alternative definitions from two codes over three months to two codes over 12 months to three codes within 12 months, and any code in the past 12 months. However, reductions in the prevalence were not consistent between conditions, as evidenced by the prevalence of obesity, which was higher under the any in past 12 months definition compared with the three within 12 months definition.

Table 2 | Prevalence of people with one or more, two or more, or three or more long term conditions using different definitions. Data are number (percentage)

Long term condition definition	One or more long term conditions	Two or more long term conditions	Three or more long term conditions
Single code	8 386 988 (86.3)	7 183 662 (73.9)	6 077 925 (62.5)
Two over three months	7 027 508 (72.3)	5 366 285 (55.2)	4 209 213 (42.3)
Two over 12 months	6 797 097 (69.9)	5 106 872 (52.5)	3 980 370 (41.0)
Three within 12 months	5 937 863 (61.1)	4 023 023 (41.4)	2 864 536 (29.5)
Any in past 12 months	5 964 391 (61.4)	4 149 619 (42.7)	3 041 805 (31.3)
Problems only	5 546 217 (57.1)	3 419 374 (35.2)	2 172 454 (22.4)

For problem codes, the pattern was mixed according to condition. Hypertension, type 2 diabetes, thyroid disease, chronic kidney disease, atrial fibrillation, and chronic obstructive pulmonary disease had a similar prevalence applying the problems definition compared with using only the single code definition (online supplemental figure A2). However, those diagnoses based only on blood test results (raised concentrations of cholesterol or triglycerides) were represented rarely in the problem table (prevalence <0.1%; Online supplemental figure A2).

The prevalence of multimorbidity under each definition stratified by demographic group is shown in figure 3 (percentages are given in online supplemental table A2). Within all sociodemographic strata, multimorbidity prevalence was lowest applying the problems definition, from 12.1% in the 18-29 year age group to 87.8% in the 80 years and older age group, compared with 48.5% and 98.1%, respectively, applying a single code definition. Comparing the definitions using alternative timeframes, the absolute difference in prevalence between definitions was large for younger age groups but became smaller with increasing age. Within categories of gender, ethnic group, and Index of Multiple Deprivation group, absolute differences

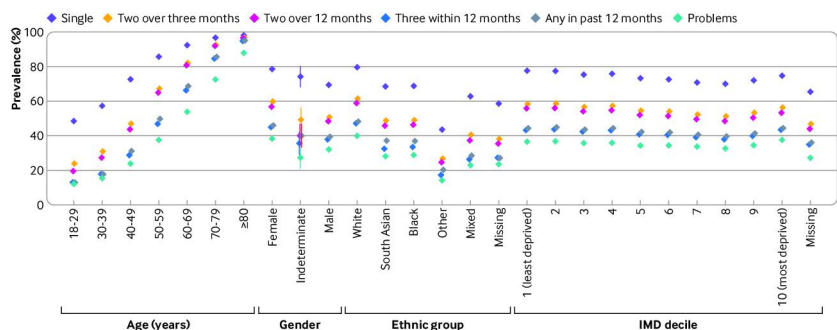


Figure 3 | Prevalence of multimorbidity using different definitions of long term condition, by demographic factors. Diamonds represent estimates and bars represent 95% confidence intervals from mixed effects logistic regression models. IMD=Index of Multiple Deprivation

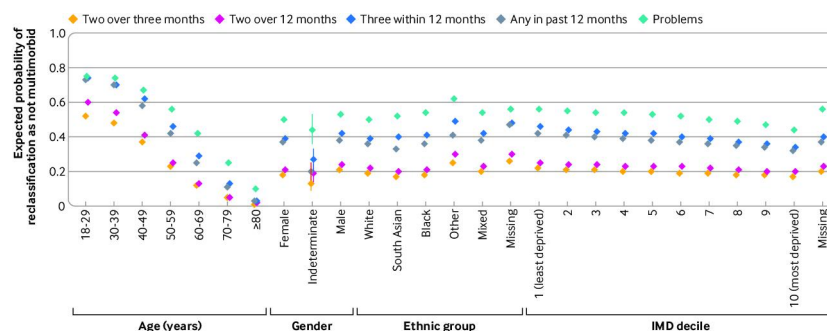


Figure 4 | Expected probabilities for being reclassified as not multimorbid with each definition compared with a single code definition. Diamonds represent estimates and bars represent 95% confidence intervals from mixed effects logistic regression models

in multimorbidity prevalence between definitions remained similar in size (figure 3).

Demographic differences in multimorbidity status according to disease timeframe

In people with multimorbidity based on the single code definition, we compared the expected probability of a change in multimorbidity status under each different timeframe, holding age, gender, ethnic group and Index of Multiple Deprivation at their mean values. The expected probabilities by definition of a person being reclassified as not multimorbid (compared with the single code definition) were 18.7% for two codes over three months, 21.9% for two codes over 12 months, 40.4% for three codes within 12 months, 37.5% for any code in the past 12 months, and 51.6% for problems.

Older people were substantially less likely to be reclassified as not multimorbid compared with younger people after controlling for gender, ethnicity and deprivation, under each of the four alternative definitions compared with a single code definition (figure 4 and online supplemental table A3). People aged 80 years or older had only a 1-10% probability of being reclassified under the different definitions, compared with those aged 18-29 years, who had a 50-75% range of being reclassified. Within each definition, the expected probabilities of a change in multimorbidity status were similar for men and women. People of South Asian ethnicity had the lowest expected probability of a change in multimorbidity status, under all alternative definitions, except for the three codes within 12 months and problems definition, where the probability was lowest for people of white ethnicity. People of other, or unknown ethnic groups had the highest expected probabilities of no longer being multimorbid under all alternative definitions. The data showed a strong gradient towards a lower chance of being reclassified as not multimorbid with increasing deprivation across all alternative definitions.

Demographic differences in the number of chronic conditions according to disease timeframe

We used mixed effects multivariable negative binomial models to estimate the expected reduction in the number of long term conditions under each definition compared with the single code definition, holding age, gender, ethnic group, and Index of Multiple Deprivation at their mean values. The mean expected reduction in the number of long term conditions from the single code definition was 1.9 for two codes over three months, 2.1 for two codes over 12 months, 3.1 for three codes within 12 months, 3.0 for any code in the past 12 months, and 3.6 for problems.

Younger people had a smaller expected reduction in the number of long term conditions than older people, after controlling for gender, ethnicity and deprivation, relative to the single code definition. This finding was apparent across all alternative definitions but was more pronounced in the problems, three within 12 months and any in past 12 months definitions, with people of 18-29 years expected to have 2.0 conditions fewer by each definition, compared with people 80 years or older expected to have 6.4, 3.9 and 4.0 conditions fewer, respectively, when compared with the single code definition (online supplemental appendix figure A3 and table A4). Men had smaller expected reductions in the count of conditions compared with women across all definitions. People of South Asian ethnicity had the largest expected reduction in the number of long term conditions, while people of black, other, or unknown ethnic groups had smaller reductions in the expected count. Data were trending towards larger expected reduction in counts of long term conditions in people living in areas of higher deprivation.

Discussion

Our study found large variation in the prevalence of multimorbidity depending on the timeframe used to define long term conditions in a nationally representative cohort of patients in the general practice electronic health record in England, ranging from 41.4% where three codes were required within a 12 month period to 73.9% where a single code sufficed. Using

active problem codes resulted in the lowest prevalence of 35.2%. A large variation was found despite using the same set of conditions and applying the alternative definitions for only 41 (19%) of the 212 conditions included. Prevalence of multimorbidity by sociodemographic factors varied substantially, particularly for age. Our selection of alternative timeframes represents a range of approaches, rather than a comprehensive assessment of all possibilities. Although we were unable to assess from the electronic health record whether people are appropriately reclassified using alternative timeframes, our results highlight the large disparities in estimates of multimorbidity prevalence, and that choice of timeframe will affect estimates for some people more than others.

Prevalence of multimorbidity reported in the existing literature ranges widely, with differences explained in part by the number of long term conditions included.⁵ We used a similar number of conditions to those of Head and colleagues who reported a prevalence of 52.8% in England in 2019.³ Their reported prevalence of multimorbidity was much higher than those from two other UK based studies of 23.2% and 27.2%, which used a smaller set of long term conditions (40 and 36, respectively, compared with Head et al's 211).^{9,16} Our most comparable definition (three codes within 12 months) resulted in a lower prevalence of multimorbidity (41.4%) than that of Head and colleagues, which may relate to our recategorisation of codes from the CALIBER study, and our different set of conditions that required multiple codes. Using a single code definition, the detected prevalence in our study was substantially higher (74%), likely reflecting inclusion of acute or inactive conditions. Our study is novel in identifying that the timeframe for defining long term conditions also has a substantial effect on prevalence of multimorbidity, comparable to the differences in the number of long term conditions on the prevalence of multimorbidity.

Our results also show that differences in prevalence estimates according to timeframe do not have uniform effects across a population. Large differences were found between demographic factors and the prevalence of multimorbidity and the number of long term conditions. Older age, South Asian ethnicity, and greater deprivation were associated with a lower risk of being reclassified as not multimorbid (figure 4) across all alternative definitions, but with corresponding larger reductions in the total count of diseases (online supplemental figure A3). Various sociodemographic factors are known to affect incidence and prevalence of multimorbidity, with previous studies highlighting that risk is higher in people living in areas of greater socioeconomic deprivation and people of black ethnicity.^{3,17} Although we found that those living in less deprived areas had a higher prevalence of multimorbidity

when using a single code definition (table 1), this group were more likely to be reclassified as not multimorbid than people in more deprived groups when requiring multiple codes (figure 4). Our findings suggest differences in the composition of long term conditions contributing to multimorbidity in younger versus older age, and in people living in less versus more deprived areas. Groups that are more likely to be reclassified are more likely to have conditions where a single code could indicate an acute condition, for example, dermatitis and enthesopathy and synovial disorders, which were large contributors to overall multimorbidity burden (online supplemental figure A1). Other explanations for differences between groups might be explained by differential undercounting of disease burden or differences in health seeking behaviour and access to healthcare.^{18,19} Variation in how disease codes are entered between general practices and between clinicians within the same general practices which impact differently on patient groups.^{20,21}

We hypothesised that using only conditions designated by clinicians as active problems in the healthcare record could be an effective route to identifying active long term conditions. This method yielded the lowest prevalence of multimorbidity of 35.2% overall, and of 87.8% in those aged 80 years and older, but with substantial variation between conditions. In primary care in England, the management of some long term conditions is incentivised by the quality and outcomes framework which started in 2004.²² We found that conditions such as hypertension, type 2 diabetes, and chronic obstructive pulmonary disease, all included in the framework, were much more similar in prevalence comparing the single code and problems definition. This suggests there may be coding bias towards conditions present in the framework when using problem codes, and so we do not recommend this method for studies of multimorbidity including a broad range of conditions.

Strengths and limitations

A strength of our study is the use of large sample of adults registered to primary care in England, with previous studies of the Clinical Practice Research Datalink Aurum data finding it to be representative of the national population.¹¹ We adopted a broad set of disease codes, many of which have particular relevance to primary care, and applied different definitions of long term condition timeframe using the same set of conditions to the same patient cohort. To an extent, these definitions are arbitrary, and our analysis presented a range of approaches used in the literature before, rather than attempting to analyse all possible definitions. For example, a condition lasting more than six months is sometimes used to define a long term condition,⁶ but we did not include this definition because estimates would lie between our range of three and 12 months.

BOX 1 | RECOMMENDATIONS FOR RESEARCH IN APPLYING A TIMEFRAME FOR COUNTING LONG TERM CONDITIONS IN THE ELECTRONIC HEALTH RECORD

1. Provide a rationale for the choice of timeframe to define a long term condition
2. For studies of disease causes, use of a single code anywhere in the record may be preferred, as a condition even if historic and inactive may be relevant to subsequent disease development
3. Analyses of current interactions with healthcare services may be best suited to using codes appearing in a recent timeframe, for example, within the past 12 months
4. Timeframes requiring more than one code may be biased by factors related to repeated coding, therefore, we recommend considering sensitivity analyses using only a single code definition

A limitation of using routinely collected healthcare data is that some conditions may not be recorded, either because a person does not present with a condition or because when presenting, the condition is not coded by a clinician. The likelihood of missing codes is unlikely to be random with respect to either patients or diseases. Although a previous systematic review found good agreement in diseases recorded in Clinical Practice Research Datalink with other sources,²³ comparison against cancer diagnoses in the Clinical Practice Research Datalink with cancer registry data found a range of 9–26% for different cancers were missing in the Clinical Practice Research Datalink.²⁴ The financial incentives offered by quality and outcomes framework have improved data collection and coding in primary care²² and may therefore inflate the prevalence of long term conditions requiring multiple codes for conditions included in the framework relative to other conditions.

The code lists used for raised total cholesterol, raised low density lipoprotein cholesterol, low high density lipoprotein cholesterol and raised triglycerides available from the CALIBER study include test results (rather than diagnostic codes), which were rarely coded as problems, partially explaining the substantially lower prevalence of multimorbidity overall. Studies using problem codes alone would therefore need to include diagnostic codes, but was not feasible for comparison in our study as the granularity required in low density lipoprotein cholesterol and high density lipoprotein cholesterol measurements cannot be accounted for using diagnostic codes alone. A further issue of using problem codes is in defining active problems. In Clinical Practice Research Datalink Aurum no date is recorded at

which a problem is changed from active to inactive. Our data extraction occurred in May 2022, and so some codes marked as inactive were likely active at the study start date, and so our findings may represent an underestimate. However, given our focus on chronic conditions, the number of conditions to have resolved is unlikely to be large.

We used a large set of chronic conditions, which will have unequal burden on clinical outcomes, healthcare use, and quality of life. Many may not be included in other multimorbidity measures, with a recent Delphi study identifying a core set of 24 conditions to always include and 35 to usually include and with no conditions identified for exclusion.²⁵ A smaller and standardised set of conditions can aid comparability and reproducibility but has drawbacks. Firstly, a smaller set limits the scope to detect novel associations between less common conditions. Secondly, choices over conditions are subjective and those that are less frequent can still have a large burden on individuals.

The overlap between some of the long term conditions included here may lead to double counting and over-estimation of multimorbidity, for example, combining myocardial infarction and angina as ischaemic heart disease. However, when investigating the cause of disease, more granular categories may be a benefit because a person with a myocardial infarction who subsequently develops angina may have a different trajectory and opportunities for intervention and prevention to a person with angina who subsequently develops a myocardial infarction, despite a common pathophysiology.

Combining diagnostic codes with relevant medications or treatments may help to distinguish active versus inactive conditions; for example, use of a proton pump inhibitor in gastritis. Similarly, the quality and outcomes framework excludes patients from the asthma register if no asthma related drugs were prescribed in the past 12 months.²⁶ However, a difficulty with this approach for some diseases is that drugs can have multiple indications; for example, proton pump inhibitors being co-prescribed with non-steroidal anti-inflammatory medications, which may be more common in those with a history of gastro-intestinal problems irrespective of active symptoms.²⁷

Implications

Despite the consensus in the medical literature on the definition of multimorbidity as the co-occurrence of two or more chronic conditions,⁴ which conditions should be included or how to determine chronicity has not been agreed.^{5 25} Our findings highlight that even when using the same set of medical codes, decisions on how to define a long term condition can change the prevalence of multimorbidity almost twofold, which has important implications for the direct comparison of estimates between studies.

Results also suggest that a universally agreed metric of multimorbidity prevalence may be an unrealistic target and that estimates are highly context dependent. Rather than seeking one rigid definition of multimorbidity, we believe that researchers should instead embrace the variety of approaches, better reflecting the variety of lived experiences of people having multimorbidity. As such, we do not advocate only one of our approaches as the best, with choice dependent in part on the aims of the research. Nevertheless, some approaches may be more suitable in particular contexts, while the impact of bias, and exclusion of specific groups from multimorbidity measures, highlighted in our work, should be assessed (Box 1).

Researchers should consider which conditions to include, and may opt for a narrower set of diseases by inclusion of only conditions for which a single code would indicate chronic risk, rather than those that may also represent acute conditions. Where inclusion of a greater breadth of diseases is preferred, researchers should decide whether some of these require presence of multiple codes over time. For studies focused on cause or accumulation of diseases over time, use of any diagnostic code in the record may be the preferred approach, as a disease, even if not currently active and only recorded once, may be relevant to subsequent disease development. However, this approach may lead to inclusion of acute or inactive conditions, inflating the prevalence of multimorbidity. For studies focused on associations between current disease burden and healthcare service use and treatment, use of more contemporaneous active diseases may be preferable. For example, use of codes from the past 12 months, or by incorporating prescription data where possible for some conditions. However, any approach that uses multiple codes may be at greater risk of bias in the frequency of coding of conditions, which may depend on factors related to the patient, clinician, general practice, and coding incentives.²¹ Therefore, we suggest that use of a single code definition should always be considered as a sensitivity analysis to understand the effect on prevalence and differential impact between patients.

Conclusion

The choice of timeframe to define the chronicity of a long term condition based on code frequency in the electronic health record has a substantial effect on the reported prevalence of multimorbidity. Younger people, people of some ethnic groups, and people living in areas of lower socioeconomic deprivation were more likely to be reclassified as not multimorbid under alternative timeframes compared with using a single code, highlighting the need to consider the potential bias arising from the choice of definition.

AUTHOR AFFILIATIONS

¹Department of Primary Care and Public Health, Imperial College London, London, UK

²Department of Mathematics, Imperial College London, London, UK

Twitter Thomas Beaney @drtombeaney

Acknowledgements Data management was provided by the Big Data and Analytical Unit at the Institute of Global Health Innovation. We thank Mark Cunningham for assistance with data extraction and management.

Contributors TB conceptualised the study, conducted the data management and formal analysis, and wrote the first draft of the manuscript. All authors contributed to the study design, method, interpretation of findings, and reviewing and editing the manuscript. TB is the guarantor and accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Transparency: The lead author (the guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Funding This research is funded through a clinical PhD fellowship awarded to TB from the Wellcome Trust 4i programme at Imperial College London (grant number N/A). JC acknowledges support from the Wellcome Trust (215938/Z/19/Z). MB acknowledges support from EPSRC grant EP/N014529/1 supporting the EPSRC Centre for Mathematics of Precision Healthcare. TW, AM and PA acknowledge support from the National Institute for Health and Care Research (NIHR) under the Applied Research Collaboration (ARC) Northwest London (grant number N/A). Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre (BRC). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, the Wellcome Trust or the Department of Health and Social Care.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Patient and public involvement This work is supported by a patient and public involvement group. This group met prior to the study start and inputted into discussions on which conditions were included in the multimorbidity measure. The group also met after the analysis and contributed to the interpretation of the findings.

Ethics approval Ethical approval and data access to the Clinical Practice Research Datalink (CPRD) was granted by CPRD's Research Data Governance Process on 28 April 2022 (Protocol reference: 22_001818).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data used in this study are not publicly available as access is subject to approval processes. More information is available from Clinical Practice Research Datalink: <https://cprd.com/research-applications>.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original

work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Thomas Beaney <http://orcid.org/0000-0001-9709-7264>

Azeem Majeed <http://orcid.org/0000-0002-2357-9858>

REFERENCES

- World Health Organization. Multimorbidity: technical series on safer primary care; 2016.
- Pearson-Stuttard J, Ezzati M, Gregg EW. Multimorbidity—a defining challenge for health systems. *Lancet Public Health* 2019;4:e599–600. 10.1016/S2468-2667(19)30222-1
- Head A, Fleming K, Kypridemos C, *et al*. Inequalities in incident and prevalent multimorbidity in England, 2004–19: a population-based, descriptive study. *Lancet Healthy Longev* 2021;2:e489–97. 10.1016/S2666-7568(21)00146-X
- Johnston MC, Crilly M, Black C, *et al*. Defining and measuring multimorbidity: a systematic review of systematic reviews. *Eur J Public Health* 2019;29:182–9. 10.1093/eurpub/cky098
- Ho I-S, Azcoaga-Lorenzo A, Akbari A, *et al*. Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies. *Lancet Public Health* 2021;6:e587–97. 10.1016/S2468-2667(21)00107-9
- O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICD-2. *Fam Pract* 2004;21:381–6. 10.1093/fampra/cmh407
- Bernell S, Howard SW. Use your words carefully: what is a chronic disease. *Front Public Health* 2016;4:159. 10.3389/fpubh.2016.00159
- Centers for Disease Control and Prevention. About chronic diseases. 2022 Available: <https://www.cdc.gov/chronicdisease/about/index.htm>
- Barnett K, Mercer SW, Norbury M, *et al*. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet* 2012;380:37–43. 10.1016/S0140-6736(12)60240-2
- South NHS. Central and West commissioning support unit. Emis web clinician quick reference guide. 2022. Available: <https://gdoc.org.uk/wp-content/uploads/2019/10/20180214-Emis-QRG-Clinicians.pdf>
- Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: clinical practice research datalink (CPRD) aurum. *Int J Epidemiol* 2019;48:1740–1740g. 10.1093/ije/dyzo34
- Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol* 2015;44:827–36. 10.1093/ije/dyv098
- Ministry of Housing Communities & Local Government. English indices of deprivation. 2019. Available: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019>
- Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al*. A chronological map of 308 physical and mental health conditions from 4 million individuals in the english national health service. *Lancet Digit Health* 2019;1:e63–77. 10.1016/S2589-7500(19)30012-3
- Dean CB. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association* 1992;87:451–7. 10.1080/01621459.1992.10475225
- Cassell A, Edwards D, Harshfield A, *et al*. The epidemiology of multimorbidity in primary care: a retrospective cohort study. *Br J Gen Pract* 2018;68:e245–51. 10.3399/bjgp18X695465
- Bisquera A, Turner EB, Ledwaba-Chapman L, *et al*. Inequalities in developing multimorbidity over time: a population-based cohort study from an urban, multi-ethnic borough in the United Kingdom. *Lancet Reg Health Eur* 2022;12. 10.1016/j.lanepe.2021.100247
- Kontopantelis E, Roland M, Reeves D. Patient experience of access to primary care: identification of predictors in a national patient survey. *BMC Fam Pract* 2010;11:61. 10.1186/1471-2296-11-61
- Parsons J, Bryce C, Atherton H. Which patients miss appointments with general practice and the reasons why: a systematic review. *Br J Gen Pract* 2021;71:e406–12. 10.3399/bjgp.2020.1017
- de JJ, Groenewegen P, Westert G. Medical practice variation: does it cluster within general practitioners practices? In: *Morbidity, Performance and Quality in Primary Care*. CRC Press, 2006.
- Beaney T, Clarke J, Salman D, *et al*. Identifying potential biases in code sequences in primary care electronic healthcare records: a retrospective cohort study of the determinants of code frequency. *BMJ Open* 2023;13. 10.1136/bmjopen-2023-072884
- Forbes LJ, Marchand C, Doran T, *et al*. The role of the quality and outcomes framework in the care of long-term conditions: a systematic review. *Br J Gen Pract* 2017;67:e775–84. 10.3399/bjgp17X693077
- Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the general practice research database: a systematic review. *Br J Gen Pract* 2010;60:e128–36. 10.3399/bjgp10X483562
- Arhi CS, Bottle A, Burns EM, *et al*. Comparison of cancer diagnosis recording between the clinical practice research datalink, cancer registry and hospital episodes statistics. *Cancer Epidemiol* 2018;57:148–57. 10.1016/j.canep.2018.08.009
- Ho ISS, Azcoaga-Lorenzo A, Akbari A, *et al*. Measuring multimorbidity in research: delphi consensus study. *BMJ Med* 2022;1:e000247. 10.1136/bmjmed-2022-000247
- England NHS. Quality and outcomes framework guidance for 2021/22. 2022 Available: <https://www.england.nhs.uk/wp-content/uploads/2021/03/Bo456-update-on-quality-outcomes-framework-changes-for-21-22-.pdf>
- Scheiman JM. The use of proton pump inhibitors in treating and preventing NSAID-induced mucosal damage. *Arthritis Res Ther* 2013;15:S5. 10.1186/ar4177

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjmed-2022-000474>).

Appendix

Impact and inequalities in the prevalence of multimorbidity using different timeframes to define long term conditions: retrospective study of disease code frequency in primary care electronic healthcare records

Beane T^{1,2} (0000-0001-9709-7264), Clarke J², Woodcock T¹, Majeed A¹, Barahona M², Aylin P¹

1. Department of Primary Care and Public Health, Imperial College London, London, W6 8RP, United Kingdom
2. Centre for Mathematics of Precision Healthcare, Department of Mathematics, Imperial College London, London, SW7 2AZ, United Kingdom

Corresponding Author:

Dr Thomas Beane

Department of Primary Care and Public Health, Imperial College London, London, W6 8RP,
United Kingdom

Email: thomas.beane@imperial.ac.uk

This section summarises the data cleaning approach. Code lists for ethnicity and disease code lists are available from our GitHub repository:

<https://tbeaney.github.io/MMclustering/>

Observation data is included in the form of Medcodes. We included observations only where the date of observation and the date of data entry occurred on or before 1st January 2020.

Duplicate observations were excluded.

Ethnicity

We used ethnicity codes developed by Davidson and colleagues.¹ We removed four codes from these list representing examination findings, as suggesting clinician observed findings rather than reported by the patient. We followed the algorithm from Mathur *et al* to identify ethnicity:²

1. Remove duplicate ethnicity codes recorded on the same date
2. Assign each code to the 5 and 16 level ethnicity categories
3. If only one ethnicity code, use single code to define ethnicity
4. If more than one code, use most frequently occurring ethnicity code
5. If more than one code with the same count, use the latest ethnicity code instead.

Chronic conditions

These were mapped using code lists developed for the CALIBER study, and adapted for use in multimorbidity in CPRD Aurum.^{3,4} We reviewed the codes in these lists, and made amendments to the code lists for diabetes, to remove Type 1 and Type 2 codes from the other/unspecified code list. We added chronic primary pain to the set of included conditions and created a new code list. Previous studies of multimorbidity in primary care settings have found a high prevalence and burden of chronic pain.^{5,6} However, in order to avoid double counting of pain related to another chronic condition included, we excluded secondary causes, and included only primary pain conditions. Codes were reviewed by two clinicians and co-authors, TB and JC, and conditions requiring multiple codes were selected based on the following criteria:

1. A disease category indicates symptoms, e.g. urinary incontinence/dysmenorrhoea.
2. A disease category indicates a disease which may be acute or present as a single short-lived episode, e.g. gastritis, pancreatitis, sinusitis, depression, anxiety.
3. A disease category may be easily confused with another disease and be diagnosed in primary care clinically without investigations, e.g. skin disorders, asthma.
4. A disease category may indicate the result of a single blood test, e.g. anaemias (except for aplastic anaemia). The disease code of Chronic Kidney Disease (CKD) is not included here as the definition implies two tests at least 3 months apart.

Where these criteria were not met, then a single code was deemed sufficient for diagnosis:

- | | |
|-------------------------------------|--|
| 1. Abdominal Aortic Aneurysm | 24. Chronic viral hepatitis |
| 2. Abdominal Hernia | 25. Cirrhosis |
| 3. Alcoholic liver disease | 26. Coeliac disease |
| 4. Angiodysplasia of colon | 27. Collapsed vertebra |
| 5. Ankylosing spondylitis | 28. Congenital Septal Defect |
| 6. Aplastic anaemias | 29. Coronary Heart Disease (not otherwise specified) |
| 7. Asbestosis | 30. Crohn's disease |
| 8. Atrial Fibrillation | 31. Cystic Fibrosis |
| 9. Atrioventricular blocks | 32. Hearing loss |
| 10. Autism and Asperger's syndrome | 33. Dementia |
| 11. Autoimmune liver disease | 34. Diabetes Mellitus: other or not specified |
| 12. Autonomic Neuropathy | 35. Diabetic Eye Disease |
| 13. Barrett's oesophagus | 36. Diabetic Neuropathy |
| 14. Benign Prostatic Hyperplasia | 37. Diaphragmatic hernia |
| 15. Visual impairment and blindness | 38. Dilated cardiomyopathy |
| 16. Bronchiectasis | 39. Diverticular Disease |
| 17. Cardiomyopathy: other | 40. Down's syndrome |
| 18. Cataract | 41. Eating Disorders |
| 19. Cerebral Palsy | 42. Endometrial hyperplasia and hypertrophy |
| 20. Chronic Fatigue Syndrome | 43. Endometriosis |
| 21. Chronic Kidney Disease | |
| 22. COPD | |
| 23. Chronic primary pain | |

44. Enteropathic arthropathy
45. Epilepsy
46. Benign essential tremor
47. Fatty Liver
48. Fibromatosis
49. Giant Cell arteritis
50. Glaucoma
51. Glomerulonephritis
52. Gout
53. Heart failure
54. HIV
55. Hodgkin Lymphoma
56. Hyperkinetic disorders
57. Hyperparathyroidism
58. Hypersplenism
59. Hypertension
60. Hypertrophic Cardiomyopathy
61. Hyposplenism
62. Idiopathic Intracranial Hypertension
63. Immunodeficiencies
64. Intellectual disability
65. Intervertebral disc disorders
66. Intracerebral haemorrhage
67. Ischaemic stroke
68. Left bundle branch block
69. Leukaemia
70. Hepatic failure
71. Lupus Erythematosus
72. Macular degeneration
73. Meniere's Disease
74. Motor neurone disease
75. Multiple sclerosis
76. Multiple valve disorder
77. Myasthenia gravis
78. Myelodysplastic Syndrome
79. Myocardial Infarction
80. Neuropathic Bladder
81. Non-Hodgkin Lymphoma
82. Nonrheumatic aortic valve disorders
83. Nonrheumatic mitral valve disorders
84. Obsessive-compulsive disorder
85. Obstructive and reflux uropathy
86. Oesophageal varices
87. Osteoarthritis (excl spine)
88. Osteoporosis
89. Other haemolytic anaemias
90. Parkinson's disease
91. Peripheral Arterial Disease
92. Peripheral Neuropathy
93. Personality disorders
94. Plasma Cell Malignancy
95. Pleural plaque
96. Polycystic ovarian syndrome
97. Polycythaemia vera
98. Polymyalgia Rheumatica
99. Portal hypertension
100. Primary Malignancy: Biliary Tract
101. Primary Malignancy: Bladder
102. Primary Malignancy: Bone
103. Primary Malignancy: Bowel
104. Primary Malignancy: Brain

105. Primary Malignancy:
Breast

106. Primary Malignancy:
Cervix

107. Primary Malignancy:
Kidney

108. Primary Malignancy: Liver

109. Primary Malignancy: Lung

110. Primary Malignancy:
Melanoma

111. Primary Malignancy:
Mesothelioma

112. Primary Malignancy:
Multiple Sites

113. Primary Malignancy:
Oesophageal

114. Primary Malignancy:
Oropharyngeal

115. Primary Malignancy: other

116. Primary Malignancy: Ovary

117. Primary Malignancy:
Pancreas

118. Primary Malignancy:
Prostate

119. Primary Malignancy: Skin

120. Primary Malignancy:
Stomach

121. Primary Malignancy: Testis

122. Primary Malignancy:
Thyroid

123. Primary Malignancy:
Uterus

124. Primary pulmonary
hypertension
125. Primary
Thrombocytopaenia

126. Psoriatic Arthritis

127. Pulmonary Fibrosis

128. Raynaud's syndrome

129. Retinal vascular occlusions

130. Rheumatic Valve Disorder

131. Rheumatoid Arthritis

132. Right bundle branch block
combinations

133. Sarcoidosis

134. Schizophrenia

135. Scleritis and episcleritis

136. Scleroderma

137. Secondary Malignancy:
Adrenal Gland

138. Secondary Malignancy:
Bone

139. Secondary Malignancy:
Bowel

140. Secondary Malignancy:
Brain

141. Secondary Malignancy:
Liver

142. Secondary Malignancy:
Lung

143. Secondary Malignancy:
Lymph Nodes

144. Secondary Malignancy:
other

145. Secondary Malignancy:
Peritoneum

146. Secondary Malignancy:
Pleura

147. Secondary polycythaemia

148. Secondary pulmonary hypertension

149. Secondary Thrombocytopaenia

150. Sick sinus syndrome

151. Sickle Cell Disease

152. Sjogren's Syndrome

153. Spina bifida

154. Spinal stenosis

155. Spondylolisthesis

156. Spondylosis

157. Stable Angina

158. Stroke: not otherwise specified
159. Subarachnoid haemorrhage

160. Supraventricular tachycardia

161. Thalassaemia

162. Thrombophilia

163. Thyroid Disease

164. Transient ischaemic attack

165. Tubulo-interstitial nephritis

166. Type 1 Diabetes Mellitus

167. Type 2 Diabetes Mellitus

168. Ulcerative colitis

169. Unstable Angina

170. Venous thromboembolic disease (Excl PE)

171. Ventricular tachycardia

The following 41 diseases required multiple codes:

1. Acne

2. Alcohol Misuse

3. Allergic and chronic rhinitis

4. Alopecia areata

5. Anaemia: other

6. Anterior and Intermediate Uveitis

7. Anxiety disorders

8. Asthma

9. Bipolar affective disorder and mania

10. Chronic Cystitis

11. Chronic sinusitis

12. Depression

13. Dermatitis

14. Dysmenorrhoea

15. Enthesopathy and synovial disorder
16. Folate deficiency anaemia

17. Gastritis and duodenitis

18. Gastro-oesophageal reflux disease

19. Hidradenitis suppurativa

20. Hypertrophic Nasal Turbinates

21. Irritable bowel syndrome

22. Low HDL-C

23. Migraine

24. Obesity

25. Pancreatitis

26. Pericardial Effusion

27. Pleural effusion

28. Posterior Uveitis

29. Psoriasis

30. Raised LDL-C

31. Raised Total Cholesterol

32. Raised Triglycerides

33. Rosacea

34. Seborrheic dermatitis

35. Sleep apnoea

36. Substance Misuse

37. Tinnitus

38. Trigeminal neuralgia

39. Urinary Incontinence

40. Vitamin B12 deficiency anaemia

41. Vitiligo

Data were de-duplicated based on disease categories on the same date, i.e., where two different codes indicating the same disease were recorded on the same day, only one was counted.

For diabetes, many codes are used for both Type 1 and Type 2 diabetes, e.g., ‘Diabetes mellitus’ and ‘O/E – diabetic maculopathy present both eyes’. Furthermore, the initial diagnosis may be mis-classified, with the correct diabetes type later recorded. To avoid double or triple counting multimorbidity burden in people with diabetes, we selected the most likely diabetes type using the following algorithm:

1. Count all occurrences of Type 1, Type 2 and unspecified/other codes for each patient
2. If only one type, then assign this as disease type
3. If Type 1 and no Type 2 codes, then assign as Type 1
4. If Type 2 and no Type 1 codes, then assign as Type 2
5. If a mix of Type 1 and Type 2 codes, then look within last 3 years of observations:
 - a. If Type 1 and no Type 2 codes, then assign as Type 1
 - b. If Type 2 and no Type 1 codes, then assign as Type 2
 - c. If a mix of both Type 1 and Type 2 codes, then assign as unspecified

Four diseases in the CALIBER code lists were defined by blood tests alone (Raised Total Cholesterol, Raised LDL-C, Low HDL-C and Raised Triglycerides). We retained only values with abnormal readings defined in the CALIBER study and made distributional assumptions regarding the units (mmol/L or mg/dL) where the defined units were implausible (Table A1). In the original CALIBER study, CKD also included blood test results, based on the eGFR, but we did not include those here, and included CKD only where a diagnostic code was present.

Table A1: mapping of cholesterol and triglyceride measurements

disease	gender	measurement	Operator1	Value1	Operator2	Value2	units
Raised Total Cholesterol		Serum	>	5	<	38	mmol/L
Raised Total Cholesterol		Serum	>	193.35			mg/dL
Raised Total Cholesterol		Plasma	>	4.8544	<	36	mmol/L
Raised Total Cholesterol		Plasma	>	187.7184			mg/dL
Raised LDL-C		Serum	>	3	<	38	mmol/L
Raised LDL-C		Serum	>	116.01			mg/dL
Raised LDL-C		Plasma	>	2.9126	<	36	mmol/L
Raised LDL-C		Plasma	>	112.6311			mg/dL
Low HDL-C	F	Serum	<	1.2	>	12	mmol/L
Low HDL-C	F	Serum	<	46.404			mg/dL
Low HDL-C	F	Plasma	<	1.165	>	12	mmol/L
Low HDL-C	F	Plasma	<	45.0524			mg/dL
Low HDL-C	M	Serum	<	1	>	10	mmol/L
Low HDL-C	M	Serum	<	38.67			mg/dL
Low HDL-C	M	Plasma	<	0.9709	>	10	mmol/L
Low HDL-C	M	Plasma	<	37.5437			mg/dL
Low HDL-C	I	Serum	<	1	>	10	mmol/L
Low HDL-C	I	Serum	<	38.67			mg/dL
Low HDL-C	I	Plasma	<	0.9709	>	10	mmol/L
Low HDL-C	I	Plasma	<	37.5437			mg/dL
Raised Triglycerides		Serum	>	1.7		38	mmol/L
Raised Triglycerides		Serum	>	150.569			mg/dL
Raised Triglycerides		Plasma	>	1.6521		36	mmol/L
Raised Triglycerides		Plasma	>	146.3256			mg/dL

Adapted from Kuan *et al* (2019)⁴; F = female, M = male, I = indeterminate

Body Mass Index (BMI) incorporated both codes and BMI measurements. We included only BMI measurements taken from age 16 years and adapted the approach used by Bhaskaran and colleagues with CPRD data:⁷

1. Weight records outside the range 20kg – 450 kg dropped.
2. Height records outside the range 121 – 214 cm dropped.
3. Duplicated records dropped.
4. BMI calculated as weight/height² for all those with height and weight on the same day

5. Where a weight was recorded on a given date, but no height recording, the most recent older height record was used.
6. Remaining weight records without a matching height, the first available future height measurement was used.
7. Where BMI could not be calculated on a given date with a height or weight measurement, but a BMI was recorded in CPRD, this was used in preference.
8. BMIs outside the range 5 – 200 kg were dropped.

Problem definition

In CPRD, a separate ‘Problem’ table records observation codes that have been marked as problems. All entries in the Problem table link to one (and only one) corresponding observation in the Observation table. We extracted all active problems in the Problem table that had a corresponding link to a code in the Observation table, using the *obsid* variable. 12.1% of our extracted observation codes had a link to a problem code.

Statistical analysis

1. Linear equation for the mixed effects logistic regression model:

$$\text{logit}(y_{i,j}) = \beta_0 + \beta_1 \text{age}_{i,j} + \beta_2 \text{gender}_{i,j} + \beta_3 \text{ethnicity}_{i,j} + \beta_4 \text{IMD}_{i,j} + u_j$$

Where j represents the GP practice and i represents each patient in practice j and:

$$y_{i,j} = \begin{cases} 0, & x_{\text{single},i,j} = x_{\text{alternative},i,j} \\ 1, & x_{\text{single},i,j} \neq x_{\text{alternative},i,j} \end{cases}$$

And where $x_{\text{single},i,j}$ and $x_{\text{alternative},i,j}$ are binary indicators for having multimorbidity under the single code definition and each of the alternative definitions, respectively, in patient i in practice j .

2. Linear equation for the mixed effects negative binomial regression model:

$$\log(y_{i,j}) = \beta_0 + \beta_1 \text{age}_{i,j} + \beta_2 \text{gender}_{i,j} + \beta_3 \text{ethnicity}_{i,j} + \beta_4 \text{IMD}_{i,j} + u_j$$

Where j represents the GP practice and i represents each patient in practice j and $y_{i,j}$ is the change in the number of LTCs for patient i in GP practice j for each definition, compared to a single code definition.

Table A2: prevalence of multimorbidity by demographic group according to each definition of chronic disease

Variable	Single	2 over 3	2 over 12	3 within 12	Any in last 12	Problems
Age (years)						
18-29	48.5%	23.9%	19.6%	13.1%	13.0%	12.1%
30-39	57.3%	30.9%	27.3%	17.8%	17.8%	15.4%
40-49	72.6%	46.9%	43.7%	28.7%	31.2%	23.9%
50-59	85.7%	67.2%	64.9%	46.9%	49.8%	37.5%
60-69	92.4%	82.1%	80.7%	66.3%	68.7%	53.8%
70-79	96.7%	92.5%	91.9%	84.6%	85.6%	72.5%
80+	98.1%	96.7%	96.6%	94.8%	95.0%	87.8%
Gender						
Female	78.5%	59.8%	56.8%	45.0%	46.0%	38.3%
Indeterminate	74.1%	49.3%	40.0%	35.6%	40.0%	27.3%
Male	69.4%	50.7%	48.4%	37.9%	39.4%	32.1%
Ethnicity						
White	79.6%	61.5%	58.8%	47.2%	48.2%	39.9%
South Asian	68.5%	48.8%	45.8%	32.5%	37.1%	28.2%
Black	68.8%	49.0%	46.3%	33.5%	36.9%	28.8%
Other	43.5%	26.8%	24.7%	17.3%	20.3%	14.2%
Mixed	62.8%	40.6%	37.3%	26.4%	28.5%	22.9%
Missing	58.6%	38.2%	35.5%	27.3%	27.2%	23.6%
IMD decile						
1 (least deprived)	77.6%	58.3%	55.8%	43.3%	44.3%	36.5%
2	77.4%	58.5%	56.0%	43.7%	44.9%	36.8%
3	75.3%	56.7%	54.1%	42.4%	43.6%	35.8%
4	75.8%	57.3%	54.7%	43.2%	44.4%	35.9%
5	73.2%	54.6%	52.0%	40.8%	42.2%	34.4%
6	72.6%	54.1%	51.4%	40.5%	41.9%	34.4%
7	70.8%	52.2%	49.6%	39.1%	40.6%	33.7%
8	70.0%	51.2%	48.5%	38.0%	39.7%	32.6%
9	72.0%	53.3%	50.4%	39.9%	41.3%	34.5%
10 (most deprived)	74.7%	56.4%	53.3%	43.5%	44.4%	37.6%
Missing	65.4%	46.8%	44.0%	34.9%	36.1%	27.2%

Figure A1: Prevalence of conditions requiring multiple codes under alternative timeframes for diagnosis (for conditions with at least 2% prevalence)

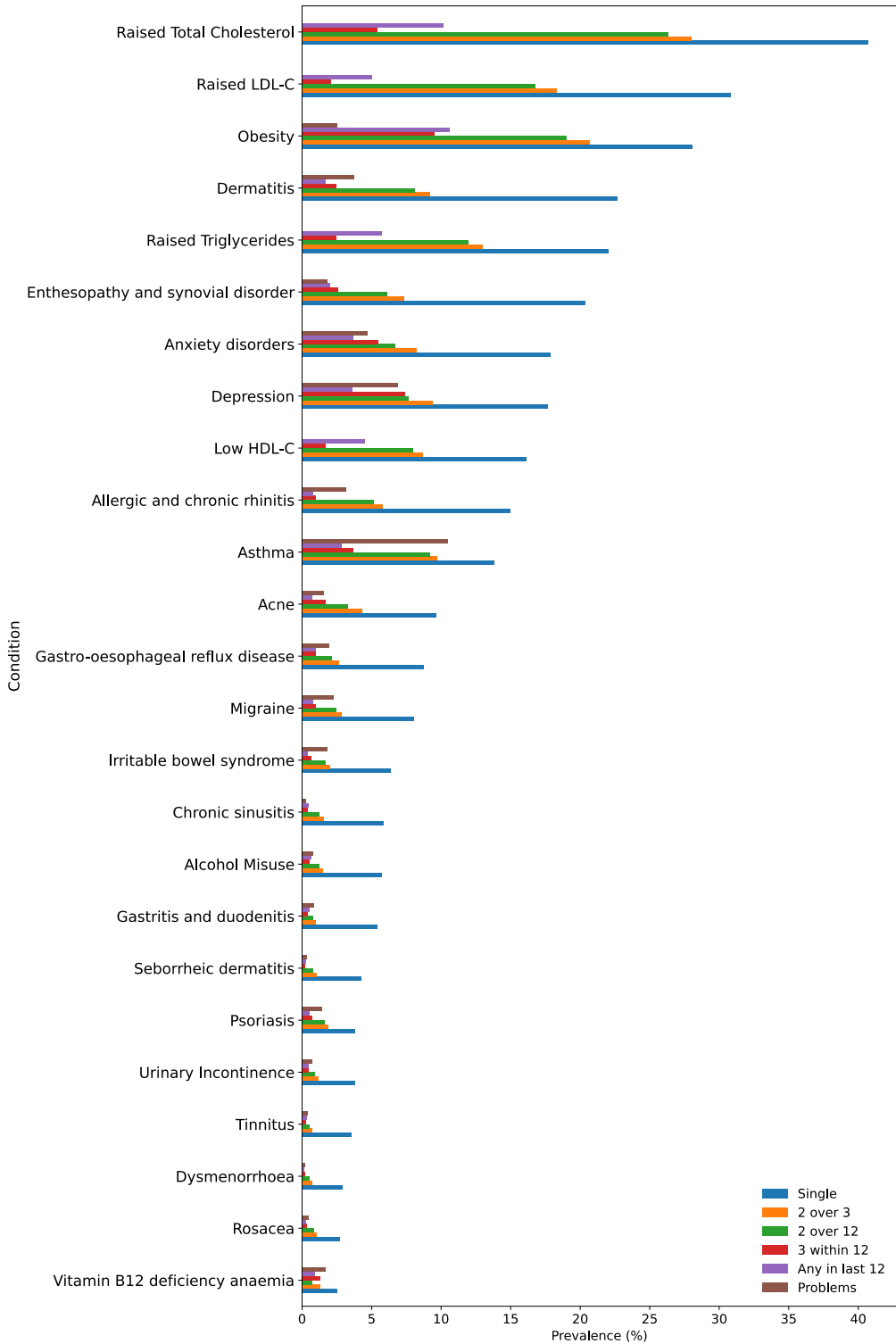


Figure A2: Prevalence of conditions using codes appearing in the problem table compared to a single code definition (for conditions with at least 1% prevalence)

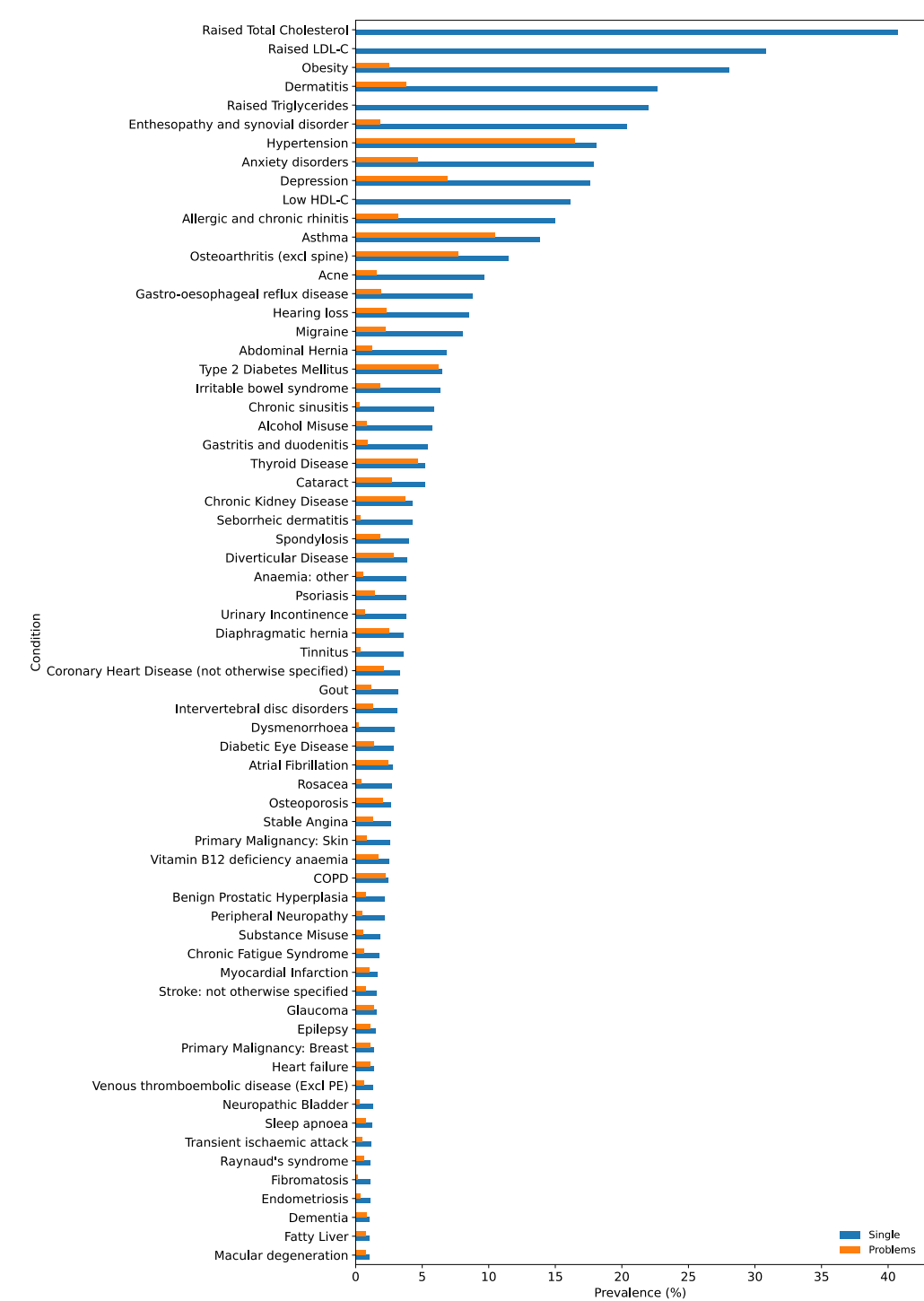
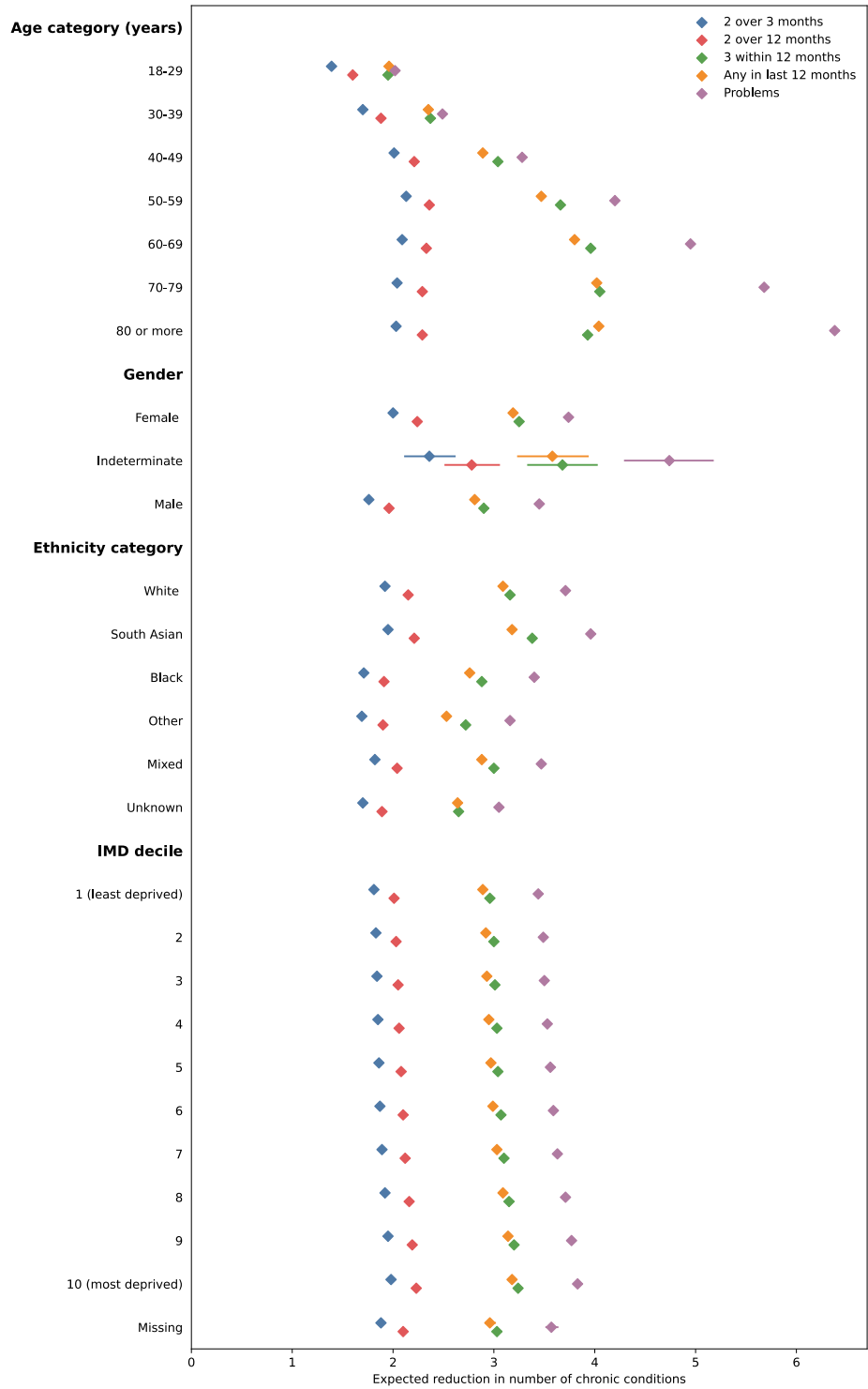


Table A3: Expected probabilities and 95% confidence intervals for being reclassified as not multimorbid with each definition compared to a single code definition

VARIABLE	DEFINITION									
	2 over 3		2 over 12		3 within 12		Any in last 12		Problems	
	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI
AGE CATEGORY (YEARS)										
18-29	0.52	0.51 - 0.52	0.60	0.60 - 0.61	0.74	0.74 - 0.75	0.73	0.73 - 0.74	0.75	0.75 - 0.75
30-39	0.48	0.48 - 0.48	0.54	0.54 - 0.54	0.70	0.70 - 0.71	0.70	0.70 - 0.70	0.74	0.74 - 0.74
40-49	0.37	0.37 - 0.37	0.41	0.41 - 0.41	0.62	0.61 - 0.62	0.58	0.57 - 0.58	0.67	0.67 - 0.67
50-59	0.23	0.22 - 0.23	0.25	0.25 - 0.25	0.46	0.46 - 0.47	0.42	0.42 - 0.42	0.56	0.56 - 0.56
60-69	0.12	0.12 - 0.12	0.13	0.13 - 0.13	0.29	0.28 - 0.29	0.25	0.25 - 0.26	0.42	0.41 - 0.42
70-79	0.05	0.04 - 0.05	0.05	0.05 - 0.05	0.13	0.12 - 0.13	0.11	0.11 - 0.11	0.25	0.25 - 0.25
80 or more	0.01	0.01 - 0.01	0.02	0.02 - 0.02	0.03	0.03 - 0.03	0.03	0.03 - 0.03	0.10	0.10 - 0.10
GENDER										
Female	0.18	0.18 - 0.18	0.21	0.21 - 0.22	0.39	0.39 - 0.39	0.37	0.36 - 0.37	0.50	0.50 - 0.50
Indeterminate	0.13	0.09 - 0.17	0.19	0.14 - 0.24	0.27	0.20 - 0.33	0.20	0.15 - 0.25	0.44	0.36 - 0.53
Male	0.21	0.21 - 0.21	0.24	0.24 - 0.24	0.42	0.42 - 0.42	0.38	0.38 - 0.39	0.53	0.53 - 0.53
ETHNICITY CATEGORY										
White	0.19	0.19 - 0.19	0.22	0.21 - 0.22	0.39	0.39 - 0.39	0.36	0.36 - 0.37	0.50	0.50 - 0.50
South Asian	0.17	0.17 - 0.18	0.20	0.20 - 0.21	0.40	0.40 - 0.40	0.33	0.33 - 0.33	0.52	0.52 - 0.52
Black	0.18	0.18 - 0.19	0.21	0.21 - 0.21	0.41	0.40 - 0.41	0.36	0.35 - 0.36	0.54	0.54 - 0.54
Other	0.25	0.25 - 0.26	0.30	0.29 - 0.30	0.49	0.49 - 0.50	0.41	0.41 - 0.42	0.62	0.62 - 0.62
Mixed	0.20	0.20 - 0.20	0.23	0.23 - 0.24	0.42	0.42 - 0.43	0.38	0.38 - 0.38	0.54	0.54 - 0.54
Unknown	0.26	0.25 - 0.26	0.30	0.29 - 0.30	0.48	0.47 - 0.48	0.47	0.47 - 0.47	0.56	0.56 - 0.56
IMD DECILE										
1 (least deprived)	0.22	0.21 - 0.22	0.25	0.25 - 0.25	0.46	0.45 - 0.46	0.42	0.42 - 0.43	0.56	0.56 - 0.56
2	0.21	0.21 - 0.21	0.24	0.24 - 0.24	0.44	0.44 - 0.44	0.41	0.40 - 0.41	0.55	0.55 - 0.55
3	0.21	0.21 - 0.21	0.24	0.24 - 0.24	0.43	0.43 - 0.44	0.40	0.40 - 0.40	0.54	0.54 - 0.54
4	0.20	0.20 - 0.21	0.23	0.23 - 0.24	0.42	0.42 - 0.43	0.39	0.39 - 0.40	0.54	0.54 - 0.54
5	0.20	0.20 - 0.20	0.23	0.23 - 0.23	0.42	0.41 - 0.42	0.38	0.38 - 0.39	0.53	0.53 - 0.53
6	0.19	0.19 - 0.20	0.23	0.22 - 0.23	0.40	0.40 - 0.41	0.37	0.37 - 0.38	0.52	0.52 - 0.52
7	0.19	0.19 - 0.19	0.22	0.22 - 0.22	0.39	0.39 - 0.39	0.36	0.36 - 0.37	0.50	0.50 - 0.50
8	0.18	0.18 - 0.18	0.21	0.21 - 0.21	0.37	0.37 - 0.38	0.35	0.35 - 0.35	0.49	0.49 - 0.49
9	0.18	0.17 - 0.18	0.20	0.20 - 0.21	0.36	0.35 - 0.36	0.34	0.33 - 0.34	0.47	0.47 - 0.47
10 (most deprived)	0.17	0.17 - 0.17	0.20	0.19 - 0.20	0.34	0.33 - 0.34	0.32	0.31 - 0.32	0.44	0.43 - 0.44
Missing	0.20	0.19 - 0.21	0.23	0.22 - 0.24	0.40	0.38 - 0.42	0.37	0.35 - 0.38	0.56	0.54 - 0.57

Figure A3: Expected reduction in number of long-term conditions with each definition, compared to a single code definition



From mixed effects negative binomial regression models. Points represent estimates and bars represent 95% confidence intervals of the estimates.

Table A4: Expected reduction in number of long-term conditions and 95% confidence interval with each definition compared to a single code definition

VARIABLE	DEFINITION									
	2 over 3		2 over 12		3 within 12		Any in last 12		Problems	
	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI
AGE CATEGORY (YEARS)										
18-29	1.39	1.38 - 1.40	1.60	1.59 - 1.61	1.95	1.94 - 1.96	1.96	1.94 - 1.97	2.02	2.01 - 2.04
30-39	1.70	1.69 - 1.70	1.88	1.87 - 1.89	2.37	2.35 - 2.38	2.35	2.33 - 2.37	2.49	2.47 - 2.51
40-49	2.01	1.99 - 2.02	2.21	2.20 - 2.22	3.04	3.02 - 3.05	2.89	2.87 - 2.92	3.28	3.26 - 3.31
50-59	2.13	2.12 - 2.14	2.36	2.35 - 2.37	3.66	3.64 - 3.68	3.47	3.44 - 3.50	4.20	4.16 - 4.23
60-69	2.09	2.08 - 2.10	2.33	2.32 - 2.35	3.96	3.94 - 3.99	3.80	3.77 - 3.84	4.95	4.91 - 4.98
70-79	2.04	2.03 - 2.05	2.29	2.28 - 2.31	4.05	4.03 - 4.08	4.02	3.99 - 4.06	5.68	5.64 - 5.73
80 or more	2.03	2.02 - 2.04	2.29	2.27 - 2.30	3.93	3.91 - 3.96	4.04	4.00 - 4.07	6.38	6.33 - 6.43
GENDER										
Female	2.00	1.99 - 2.01	2.24	2.23 - 2.25	3.25	3.23 - 3.27	3.19	3.17 - 3.22	3.74	3.71 - 3.77
Indeterminate	2.36	2.11 - 2.62	2.78	2.51 - 3.06	3.68	3.33 - 4.03	3.58	3.23 - 3.94	4.74	4.29 - 5.18
Male	1.76	1.75 - 1.77	1.96	1.95 - 1.97	2.90	2.88 - 2.91	2.81	2.79 - 2.83	3.45	3.42 - 3.48
ETHNICITY CATEGORY										
White	1.92	1.91 - 1.93	2.15	2.14 - 2.16	3.16	3.14 - 3.18	3.09	3.07 - 3.11	3.71	3.68 - 3.73
South Asian	1.95	1.94 - 1.96	2.21	2.20 - 2.22	3.38	3.36 - 3.40	3.18	3.16 - 3.21	3.96	3.93 - 3.99
Black	1.71	1.70 - 1.72	1.91	1.90 - 1.92	2.88	2.87 - 2.90	2.76	2.74 - 2.78	3.40	3.37 - 3.43
Other	1.69	1.68 - 1.70	1.90	1.89 - 1.91	2.72	2.70 - 2.74	2.53	2.51 - 2.55	3.16	3.13 - 3.18
Mixed	1.82	1.81 - 1.84	2.04	2.03 - 2.06	3.00	2.98 - 3.02	2.88	2.85 - 2.90	3.47	3.44 - 3.50
Unknown	1.70	1.69 - 1.70	1.89	1.88 - 1.90	2.65	2.63 - 2.66	2.64	2.62 - 2.66	3.05	3.03 - 3.07
IMD DECILE										
1 (least deprived)	1.81	1.80 - 1.82	2.01	2.00 - 2.02	2.96	2.95 - 2.98	2.89	2.86 - 2.91	3.44	3.41 - 3.47
2	1.83	1.82 - 1.84	2.03	2.02 - 2.04	3.00	2.98 - 3.01	2.92	2.90 - 2.94	3.49	3.46 - 3.51
3	1.84	1.83 - 1.85	2.05	2.04 - 2.06	3.01	2.99 - 3.02	2.93	2.91 - 2.96	3.50	3.47 - 3.53
4	1.85	1.84 - 1.86	2.06	2.05 - 2.07	3.03	3.01 - 3.04	2.95	2.92 - 2.97	3.53	3.50 - 3.55
5	1.86	1.85 - 1.87	2.08	2.06 - 2.09	3.04	3.03 - 3.06	2.97	2.94 - 2.99	3.56	3.53 - 3.58
6	1.87	1.86 - 1.88	2.10	2.09 - 2.11	3.07	3.05 - 3.09	2.99	2.97 - 3.02	3.59	3.57 - 3.62
7	1.89	1.88 - 1.90	2.12	2.11 - 2.13	3.10	3.08 - 3.12	3.03	3.01 - 3.05	3.63	3.61 - 3.66
8	1.92	1.91 - 1.93	2.16	2.15 - 2.17	3.15	3.13 - 3.17	3.09	3.06 - 3.11	3.71	3.68 - 3.74
9	1.95	1.94 - 1.96	2.19	2.18 - 2.20	3.20	3.18 - 3.21	3.14	3.11 - 3.16	3.77	3.74 - 3.80
10 (most deprived)	1.98	1.97 - 1.99	2.23	2.22 - 2.25	3.24	3.22 - 3.26	3.18	3.16 - 3.21	3.83	3.80 - 3.86
Missing	1.88	1.84 - 1.91	2.10	2.06 - 2.14	3.03	2.98 - 3.08	2.96	2.91 - 3.02	3.57	3.51 - 3.64

References

1. Davidson, J. *et al.* Codelists for: ‘Ethnic differences in the incidence of clinically diagnosed influenza: an England population-based cohort study 2008-2018’. <https://datacompass.lshtm.ac.uk/id/eprint/2102/> (2021) doi:10.17037/DATA.00002102.
2. Mathur, R., Palla, L., Farmer, R. E., Chaturvedi, N. & Smeeth, L. Ethnic differences in the severity and clinical management of type 2 diabetes at time of diagnosis: A cohort study in the UK Clinical Practice Research Datalink. *Diabetes Research and Clinical Practice* **160**, (2020).
3. Head, A. *et al.* Inequalities in incident and prevalent multimorbidity in England, 2004–2013;19: a population-based, descriptive study. *The Lancet Healthy Longevity* **2**, e489–e497 (2021).
4. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health* **1**, e63–e77 (2019).
5. Bisquera, A. *et al.* Inequalities in developing multimorbidity over time: A population-based cohort study from an urban, multi-ethnic borough in the United Kingdom. *Lancet Reg Health Eur* **12**, 100247 (2021).
6. Ashworth, M. *et al.* Journey to multimorbidity: longitudinal analysis exploring cardiovascular risk factors and sociodemographic determinants in an urban setting. *BMJ Open* **9**, (2019).
7. Bhaskaran, K., Forbes, H. J., Douglas, I., Leon, D. A. & Smeeth, L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open* **3**, e003389 (2013).