BMJ Open

Development of an algorithm to classify primary care electronic health records of alcohol consumption: experience using data linkage from UK Biobank and primary care electronic health data sources

David Fraile-Navarro [1,2] Amaya Azcoaga-Lorenzo [1] Utkarsh Agrawal [1] Bhautesh Jani [3] Adeniyi Fagbamigbe [1] Dorothy Currie [1] Alexander Baldacchino [1] Frank Sullivan [1]

For numbered affiliations see end of article.

Correspondence to
Dr Amaya Azcoaga-Lorenzo;
aal22@st-andrews.ac.uk

ABSTRACT

Objectives Develop a novel algorithm to categorise alcohol consumption using primary care electronic health records (EHRs) and asses its reliability by comparing this classification with self-reported alcohol consumption data obtained from the UK Biobank (UKB) cohort.

Design Cross-sectional study.

Setting The UKB, a population-based cohort with participants aged between 40 and 69 years recruited across the UK between 2006 and 2010.

Participants UKB participants from Scotland with linked primary care data.

Primary and secondary outcome measures Create a rule-based multiclass algorithm to classify alcohol consumption reported by Scottish UKB participants and compare it with their classification using data present in primary care EHRs based on Read Codes. We evaluated agreement metrics (simple agreement and kappa statistic).

Results Among the Scottish UKB participants, 18 838 (69%) had at least one Read Code related to alcohol consumption and were used in the classification. The agreement of alcohol consumption categories between UKB and primary care data, including assessments within 5 years was 59.6%, and kappa was 0.23 (95% CI 0.21 to 0.24). Differences in classification between the two sources were statistically significant (p<0.001); More individuals were classified as 'sensible drinkers' and in lower alcohol consumption levels in primary care records compared with the UKB. Agreement improved slightly when using only numerical values (k=0.29; 95% CI 0.27 to 0.31) and decreased when using qualitative descriptors only (k=0.18;95% CI 0.16 to 0.20).

Conclusion Our algorithm classifies alcohol consumption recorded in Primary Care EHRs into discrete meaningful categories. These results suggest that alcohol consumption may be underestimated in primary care EHRs. Using numerical values (alcohol units) may improve classification when compared with qualitative descriptors.

Strengths and limitations of this study

► This is the first study assessing the agreement between alcohol consumption in electronic health records (EHRs) from primary care and a different source at individual patient level.
► Our algorithm permits multiclass classification of alcohol consumption in primary care EHRs into five categories.
► Alcohol consumption patterns can vary in a short period of time and health records might not capture this change.
► The UK Biobank cohort is not representative of the whole population and therefore this data cannot be used to infer population levels of alcohol consumption.

BACKGROUND

Alcohol consumption is a major cause of morbidity and mortality globally.[1] According to WHO in 2016, harmful use of alcohol accounted for 3 million deaths worldwide and 132.6 million disability-adjusted life-years.[1] The Scottish Public Health Observatory, reported that in 2016, the proportion of adults who drink alcohol at levels beyond the recommended 14 units per week were around 30% of men and 16% of women.[2] Moreover, in 2020 the Scottish Health Survey (SHeS) reported that alcohol sales in 2019 were equivalent to 19.1 units per adult per week, exceeding 36% of the low-risk drinking guideline (14 units).[3] Evidence suggests that even low levels of regular alcohol consumption can cause harm.[4] Alcohol consumption plays a major role in precipitating and perpetuating mental health[5] and physical health

conditions including cancer[6] and heart disease[7] with subsequent increased overall mortality.[8]

Reliable estimates of levels of alcohol consumption in the population are required to guide and evaluate policies, to enable alcohol research and provide better individualised care. A patient's alcohol consumption amount can be a crucial factor as an individual risk factor. It is also extremely valuable when conducting epidemiological studies where it may be a confounder, a covariate or the primary exposure variable. Despite its importance, estimating how much people drink, is still a major problem when evaluating the effects of alcohol.[9]

Prospective studies are logistically complex and face difficulties with recruiting and retaining individuals.[10] Self-reported measures including surveys and standardised questionnaires are the most common methods for assessing alcohol consumption but are at risk of reporting and selection bias.[11] These studies can also be unreliable because of the inaccuracy of subjective recall[12] and because those who respond to a survey or enrol in a cohort typically differ from their non-responding counterparts.[13 14] It has been suggested that the downward trend in alcohol consumption in recent years is partially attributable to falling response rates with fewer heavy drinkers responding over time.[15 16] Population surveys like the SHeS use response probability weighting to make them nationally representative. These weights are based on limited sociodemographic information. A study published in 2014 found that survey participants in the SheS experienced lower rates of alcohol-related harm than the general population in Scotland[13] and worldwide.[17] Different approaches[18 19] have been used to improve the validity of data from surveys, as it is recognised they are not the perfect source to evaluate outcomes related to health behaviours such as alcohol consumption.[20] The current COVID-19 pandemic has probably introduced important changes in patterns of alcohol consumption.[21] It is still unknown if the pandemic will lead to an increase or a decrease in total alcohol consumption.[22] Lockdowns and other anti-COVID measures may affect the pattern of alcohol consumption.[3] Having reliable and regular estimates are more important than ever and primary care electronic health records (EHRs) have the potentiality to provide this data with reasonable investment and efforts.[23]

The use of routinely collected electronic data (RCD) and linkage from different sources are increasingly utilised in medicine offering an opportunity for developing observational research in biomedical sciences.[24] Linkage of data from the SheS, the Scottish Morbidity Records and National Records of Scotland has been used to improve the estimation of alcohol consumption in Scotland.[25] Nonetheless, the lack of high-quality RCD on alcohol consumption is a limiting factor for conducting population studies in this field. Developing valid and reliable instruments to categorise Primary Care EHRs will contribute to improving the assessment of alcohol consumption and target health interventions where appropriate.[26] The use of algorithmic approaches to analyse relational databases allows the classification of big datasets making them more usable for epidemiological research quality improvement and guiding patient care.

We aim to develop a novel algorithm to categorise alcohol consumption using primary Care EHRs. We assess its reliability by comparing this classification with self-reported alcohol consumption data obtained from the UK Biobank (UKB) cohort from the same participants.

METHODS

Cross-sectional population data were obtained from the UKB cohort. We developed and evaluated our algorithm following a four-step process:

1. Classify the UKB participants in Scotland with available Primary Care EHR data into discrete alcohol consumption categories. These categories were established, based on previous research[27–30] and taking into account the characteristics of the data available as follows:
   – Non-drinker.
   – Sensible drinkers (1–14 units/week).
   – Moderate drinkers (15–21 units/week).
   – Hazardous drinkers (22–35 units/week).
   – Harmful drinkers (>35 units/week).
2. Combine Read Codes on alcohol consumption from Primary Care EHR recorded within 5 years of the UKB assessment to develop a 'primary care-based' classification on the same individual to match the abovementioned categories.
3. Calculate the agreement between the results of the algorithm and data from the UKB.
4. Evaluate if the agreement improves by restricting the algorithm to use different types of Read Codes and by limiting the period between assessments. To achieve this, we used deidentified participants information from the Scottish UKB cohort which also contains, when available, linked primary care data. We developed an algorithm using relevant Read Codes[31] from the primary care database to classify each participant into a drinking category to compare with their response to the UKB questionnaire

To achieve this, we used deidentified participants information from the Scottish UKB cohort which also contains, when available, linked primary care data. We developed an algorithm using relevant Read Codes[32] from the primary care database to classify each participant into a drinking category to compare with their response to the UKB questionnaire.

Data sources
UK Biobank
The UKB is a large and detailed population-based cohort with participants aged between 40 and 69 years at the time of recruitment. Participants were recruited across the UK between 2006 and 2010.[32] Of the 503 317 initially recruited 35 850 participants were from Scotland. One of the main advantages of using the UKB as a data source
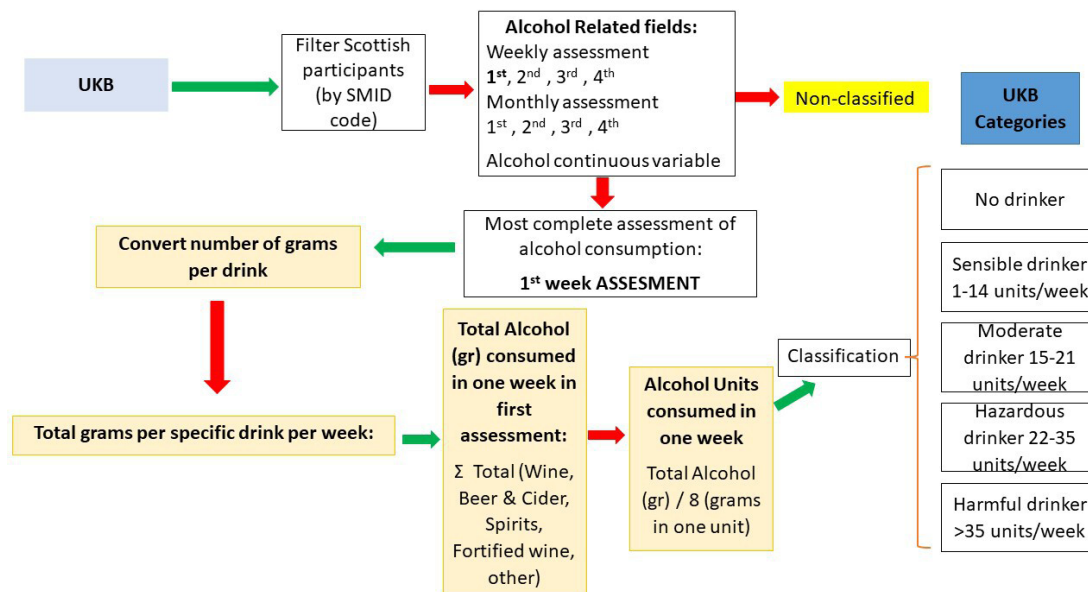
Figure 1 Classification of UKB participants based on reported alcohol consumption. SIMD, Scottish Index of Multiple Deprivation; UKB, UK Biobank.

is that alcohol consumption at enrolment and follow-up visits was assessed through a touchscreen questionnaire with quantity-frequency type questions and beverage specificity allowing accurate estimation of units of alcohol consumed. Evidence suggests that this approach may improve under-reporting.[33] We identified UKB participants from Scotland and classified them into alcohol consumption categories by calculating the number of units of alcohol consumed per week. This was based on the self-reported amount and type of beverage. Initially, we considered using the closest in time assessment to match with the primary care one. Preliminary analysis of the UKB database showed that second, third and fourth

assessments were incomplete and therefore were not used for classification. Figure 1 shows the classification of UKB participants based on reported alcohol consumption at enrolment.

Primary care data

There is currently no UK national system for collecting or sharing primary care data. UKB has liaised with various data suppliers and other intermediaries to obtain primary care data for UKB participants, all of whom have provided written consent for linkage to their health-related records. Since September 2019, UKB has made available linked Primary Care data for 45% of the
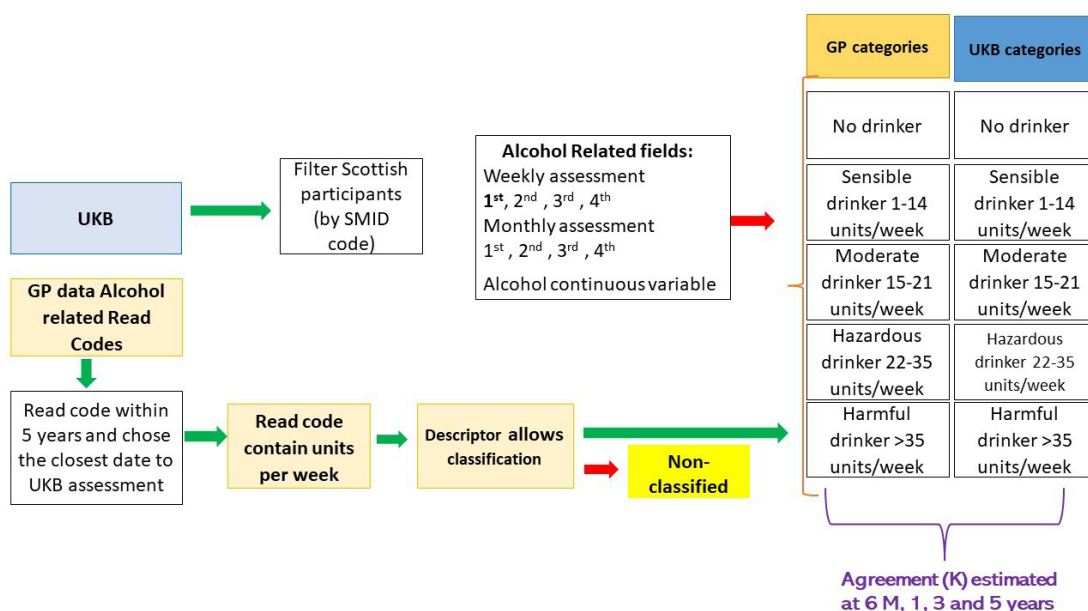

Figure 2 Algorithm to classify alcohol-related red codes from primary care. GP, general practitioner; SIMD, Scottish Index of Multiple Deprivation; UKB, UK Biobank.
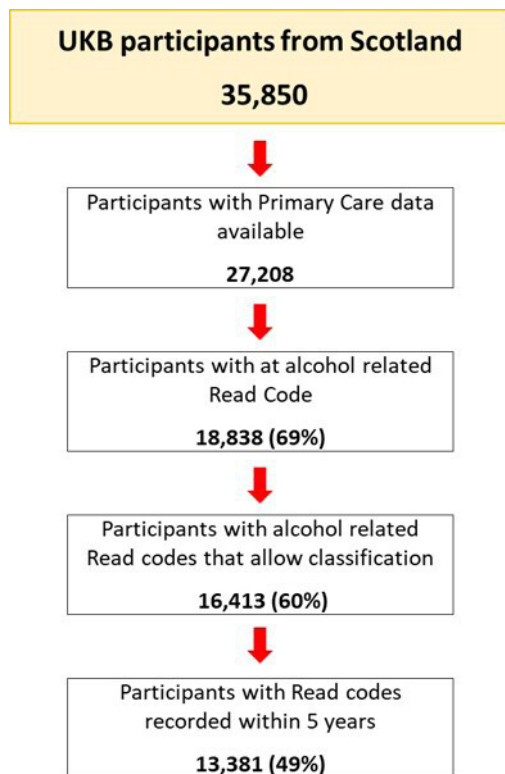
UKB participants from Scotland
35,850

↓

Participants with Primary Care data
available
27,208

↓

Participants with at alcohol related
Read Code
18,838 (69%)

↓

Participants with alcohol related
Read codes that allow classification
16,413 (60%)

↓

Participants with Read codes
recorded within 5 years
13,381 (49%)

Figure 3 Flow chart of participants. UKB, UK Biobank.

whole cohort. In Scotland, general practitioners (GPs) and other primary care professionals often ask patients about alcohol consumption and it has been one of the Quality and Outcome Framework indicators from 2004 to 2016. Alcohol consumption is recorded in the Primary Care database using Read Codes (Read V.2). As stated in National Health Service (NHS) Digital site:[34] 'Read Codes are a coded thesaurus of clinical terms. They have been used in the NHS since 1985. There are two versions: version 2 (V.2) and version 3 (CTV3 or V.3). Both versions provide a standard vocabulary for clinicians to record patient findings and procedures, in health and social care IT systems across primary and secondary care'. Over the years different nomenclatures to describe alcohol has been included in Read V.2 and V.3. There is no clear guidance on which is the preferred method to code alcohol consumption using this system.

Algorithm development
First, we created a comprehensive list of all Read V.2 alcohol-related codes by exploring all the categories and subcategories of the thesaurus with an explicit mention of alcohol or alcohol-related terms. Second, the primary care database was queried to select all participants who had a Read Code indicating alcohol intake. Qualitative descriptors of alcohol consumption (eg, codes for light/moderate/heavy drinker) or a quantitative record containing the number of units of alcohol consumed per week were extracted. Two clinicians (DF-N and AA-L) independently assessed the different descriptors and assigned these to one of the previously described five categories.

Disagreement was minor and resolved by discussion or if needed with the help of a third clinician (FS) to make the final decision. The full list of Read Codes contributing to this algorithm is given in online supplemental file 1. Only participants who had a relevant Read Code recorded within 5 years of the UKB assessment were considered for the final analysis. If both a qualitative descriptor and a quantitative one were available, the numerical value was used in preference to categorise them. The algorithm was used to classify each participant in the same alcohol consumption categories (figure 2).

Statistical analysis
Cohen's kappa statistic[35] and McNemar-Bowker test[36 37] were used to evaluate the agreement between the classifications from both sources (UKB and primary care EHR data). Alcohol consumption was classified into five groups. The kappa statistic was estimated between assessments recorded within 5 years. When more than one assessment was available in primary Care EHR data, the record nearest in time to the UKB assessment was used to calculate the level of agreement. We performed further subgroup analysis stratifying data by age and sex. Additional agreement measures using Kappa statistics were calculated after restricting the Read Codes (only Read Codes containing numerical values and Read Codes with a qualitative descriptor) and limiting the periods between assessments. As suggested by Landis and Koch, we interpreted the kappa values as follows: ≤0.20 indicates poor agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good agreement and ≥0.81 indicates excellent agreement.[38] The McNemar-Bowker test is a modification of the McNemar test for a 2×2 paired table for symmetry. The test describes whether the marginal distributions of two different measures or ratings are similar, as would be expected when measures agree.[36 37 39] Additionally, weighted kappa was also calculated. Cohen's (unweighted) kappa accounts for the disagreement between the two rating methods, but not the extent of disagreement. This is especially relevant when the ratings are ordered. The weighted kappa coefficient takes into consideration the different levels of disagreement between categories. For example, if one rater 'strongly disagrees' and another 'strongly agrees' this must be considered a greater level of disagreement than when one rater 'agrees' and another 'strongly agrees'.[40]

Data were processed and analysed using Python[41] V.3.8 and R, V.3.4 statistical software.[42] Statistical significance was set at p≤0.05, and CI to 95%.[43]

Patient and public involvement
Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS
Algorithm results
Of 502 493 UKB participants with available data at the time of extraction, 35 850 were from Scotland and therefore, eligible for inclusion. Of those, 27 208 had linked

Table 1 Per cent of participants in each alcohol consumption category by age and sex according to UKB and primary care—EHR data

	Non-alcohol (%)		Sensible drinker (%)		Moderate (%)		Hazardous (%)		Harmful (%)	
	GP	UKB	GP	UKB	GP	UKB	GP	UKB	GP	UKB
All participants	6.7	5.8	80.8	56.7	2.6	14.5	8.5	12.9	1.4	10.1
Age bands										
<50	5.6	4.3	81.6	57.5	3.3	15.3	7.8	13.1	1.7	9.8
50–60	6.3	5	80.1	55.5	2.7	15.1	9.3	13.5	1.7	10.5
>60	6.7	5.8	80.8	56.7	2.6	14.5	8.5	12.9	1.4	10.1
Sex										
Female	7.6	6.3	86.3	71.5	1.6	12.7	3.9	7.1	0.6	2.4
Males	5.5	5.2	74.4	39.3	3.8	16.5	13.9	19.8	2.4	19.1

EHR, electronic health record; GP, general practitioner; UKB, UK Biobank.

primary care records and 18838 (69%) had at least one Read Code related to alcohol consumption and were included in the analysis (figure 3). The mean age of this subgroup of participants was 57 years (SD=8). More than half (53%) were women, which was expected considering the characteristics of the underlying UKB cohort.

The primary care data from the 18838 participants with at least one alcohol-related Read Code, contained 86 different Read Codes and 102 descriptors related to alcohol consumption. The median number of records per individual was 4 (range 1–59). Certain Read Codes did not permit meaningful classification of alcohol consumption as they did not provide enough information (eg, 136F. Spirit drinker) or essential information was missing (eg, 136 Alcohol consumption should contain a numeric value, but this was not available). This reduced the number of individuals with records in both data sources to 16413. Subsequently, only Read Codes recorded within 5 years of the assessment date from the UKB were considered for classification and 13381 individuals (54% women) were finally included in our algorithm.

Prevalence of alcohol consumption categories
The most common alcohol consumption category in all participants using both sources was 'sensible drinkers (1–14 units per week)'. However, individuals were classified more often in this category using primary care HER (80.8%) compared with the UKB (56.7%) p<0.001. On the contrary, more individuals were classified in higher consumption categories and less in the non-alcohol consumption group based on UKB assessment compared with primary care p<0.001. Findings were similar when analysed by age bands and sex. UKB data reported higher alcohol consumption than primary care EHRs. Males were assigned to higher drinking categories than females from both sources (table 1, figure 4 and online supplemental file 2).

Agreement between the UKB and primary care classifications
The level of agreement between UKB and Primary care classification including assessments within 5 years was 59.6% and the kappa analysis showed only a fair agreement (κ=0.23, 95% CI 0.21 to 0.24). The difference between the two sources of classifications was statistically significant (McNemar's χ^2=3550, df=10, p<0.001) (table 2).

Agreement between the UKB and primary care classifications by age, sex and at different times between assessments
The overall agreement between the classifications from the two sources ranged from 59.6% to 60.2% when considering different periods between both assessments (taken
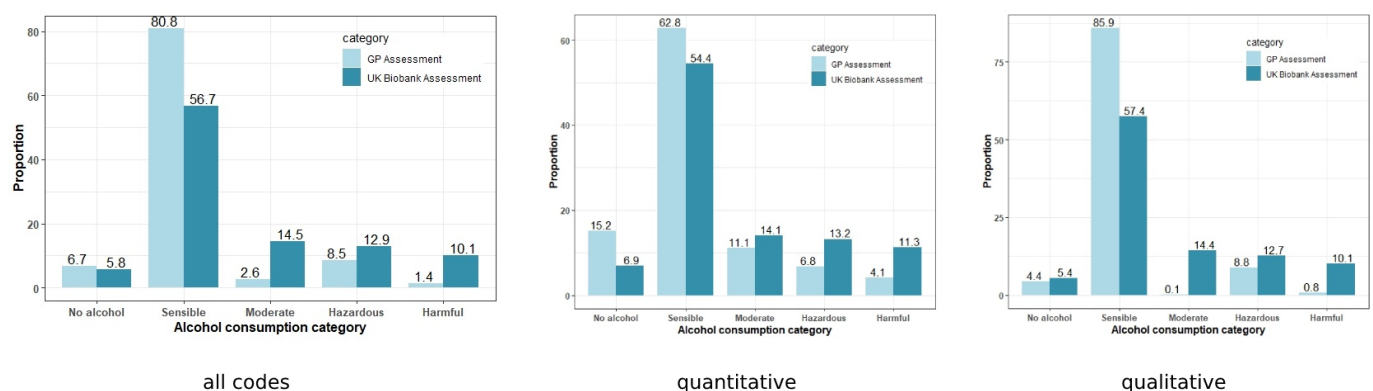

Figure 4 Alcohol consumption categories from both sources. GP, general practitioner.

Table 2 Intersource agreement within 5 years between UKB and primary care EHR assessments

	UKB					
	No alcohol	Sensible	Moderate	Harmful	Hazardous	Total
Primary care EHR						
No alcohol	409	415	30	19	17	890
Sensible	317	7019	1643	622	1215	10816
Moderate	4	53	98	68	126	349
Harmful	27	10	13	110	32	192
Hazardous	21	96	150	526	341	1134
Total	778	7593	1934	1345	1731	13381

EHR, electronic health record; UKB, UK Biobank.

at the end of the first 6 months, first, third and fifth year). Having a nearer value in time to the UKB assessment did not improve that. Differences between the UKB classification and primary care EHRs were statistically significant in both males and females (McNemar's χ^2 for male=2494.7, p<0.0001 and McNemar's χ^2 for female=1125.0, p<0.0001) although the agreement and kappa were significantly higher for females (71.64% k=0.24) than men (45.33% k=0.18) p<0.0001. The level of agreement and the Kappa values are summarised in table 3.

Algorithm including only quantitative Read Codes

When restricting the algorithm to include exclusively Read Codes containing numerical values, 6629 participants contributed to this classification. The most common category was 'sensible drinker'. There was less difference between both sources (62.8% from primary care data and 54.4% from UKB) (figure 4) compared with the comparison with qualitative codes. Simple agreement was 56,1%. However, the kappa analysis was slightly better (κ=0.29, 95 % CI 0.27 to 0.31) although the difference between the two sources of classifications was still statistically significant (McNemar's χ^2=1449.6, df=10, p<0.001). Simple agreement and Kappa analysis did not improve either when using Primary Care EHR values closer to the UKB assessment date (tables 4 and 5).

Classification including only qualitative Read Codes

A total of 10065 individuals were available to include in the algorithm considering only qualitative Read Codes. Using these codes only, simple agreement was similar to those from the quantitative Read Codes (59.9% vs 56.1%). However, the kappa analysis showed poorer agreement (k=0.18; 95% CI 0.16 to 0.20). The most common category was 'sensible drinkers' (85.9% from primary Care EHR and 57.4% from UKB). Only 0.1% were classified as 'moderate drinker' and 0.8% 'harmful' using primary care EHR data. The difference between the two sources of classifications was statistically significant (McNemar's χ^2=3176.4, df=10, p<0.001) (figure 4). As in the previous classifications, agreement and kappa did not improve by using assessments which were closer in time (table 5).

DISCUSSION
Key results

We developed an algorithm that allows the reduction of 86 Read Codes containing 102 different descriptors from Primary Care into a meaningful classification of five categories of alcohol consumption. In our sample, 69% of UKB participants with linked Primary Care EHR had at least one Read Code related to alcohol consumption and 60% had a code that allowed classification into a drinking category.

Classification into the different alcohol consumption categories from the UKB data and the primary care Read Codes showed significant differences. In both cases the most prevalent category was 'sensible drinker', but our algorithm assigned more people to this category and consistently less to higher consumption groups and more to 'non-alcohol'. Although it is not possible to consider either source as the 'gold standard', the UKB assessment used a detailed and comprehensive self-administered questionnaire. We consider that the UKB data are more likely to be accurate than the data which is routinely recorded in primary care. Our results suggest that alcohol consumption could be systematically underestimated in primary care records and that individuals might be classified into more socially desirable categories, either by biases introduced by professionals or by patients themselves, when reporting their own consumption.

The overall agreement between the UKB data and the algorithm using primary care Read Codes was 59.6% and this proportion only varied by −2.1%+1.7% regardless of the algorithm rules and periods in between assessments and age. Interestingly, although the kappa statistic was not substantially different, female participants showed a much higher agreement than their male counterparts (up to +13.5%). Kappa analysis also showed slightly better agreement for the algorithm of Read Codes only containing numerical values of units of alcohol. This finding suggests that when a more objective measurement is used classification improves. The great variety of Read Codes related to alcohol consumption containing a qualitative description of the drinking patterns introduces

Table 3 Participants, agreement and kappa values at the four different periods of assessment using all values

Time frame between UKB and primary care assessments	No of participants	Agreement N (%)	Unweighted kappa	Weighted kappa
All participants				
6 months	3720	2240 (60.2)	0.20	0.32
1 year	6000	3592 (59.9)	0.20	0.32
3 years	10764	6460 (60.0)	0.21	0.34
5 years	13381	7977 (59.6)	0.23	0.35
Females				
6 months	1969	1447 (73.5)*	0.24	0.31
1 year	3200	2349 (73.4)*	0.24	0.30
3 years	5787	4227 (73.0)*	0.23	0.30
5 years	7264	5204 (71.64)*	0.24	0.30
Males				
6 months	1751	793 (45.2)*	0.15	0.28
1 year	2800	1243 (44.4)*	0.15	0.27
3 years	4977	2233 (44.9)*	0.16	0.29
5 years	6117	2773 (45.3)*	0.18	0.31
Age bands				
<50 years				
6 months	629	371 (58.9)	0.19	0.34
1 year	1132	681 (60.2)	0.20	0.35
3 years	2306	1412 (61.2)*	0.20	0.34
5 years	2999	1802 (60.1)*	0.22	0.35
50–60 years				
6 months	1226	738 (60.2)	0.19	0.31
1 year	2059	1271 (61.7)	0.19	0.30
3 years	3827	2209 (57.7)*	0.19	0.33
5 years	4763	2740 (57.5)*	0.20	0.34
>60 years				
6 months	1865	1131 (60.6)	0.20	0.32
1 year	2809	1694 (60.3)	0.20	0.32
3 years	4631	2839 (61.3)*	0.21	0.34
5 years	5619	3435 (61.1)*	0.23	0.35

*Significant at p<0.0001, equality of proportions test.
UKB, UK Biobank.

obvious subjectivity in the assessment process as the professional will have to make a judgement and decide which category the individual falls in.

Out of the 102 descriptors, 36 did not provide sufficient information to be used for classification into one of the five categories (see online supplemental file 1 for full details). Some of these codes could be relevant to make individual clinical decisions (eg, 136E. ex-very heavy drinker-(>9u/day)) but others are useless if units of alcohol are not added (eg, 136F Spirit drinker). It is difficult to justify making all these codes available as good clinical practice cannot be based on an unreliable coding system. What is certain is that the coding process

for clinicians is more laborious than it needs to be to find the most appropriate code. Inaccurate recording means that these data are potentially less useful for epidemiological purposes and we recommend that the consumption of alcohol is recorded in grams or units of alcohol.

Strengths and limitations

To our knowledge, this is the first time that an algorithm for classifying alcohol consumption has been developed using the UKB and primary care EHR data together, allowing us to compare alcohol consumption between these two linked sources. Our algorithm permits meaningful classification into five categories, which

Table 4 Participants, agreement and kappa values at the four different periods of assessment using only numerical values and qualitative descriptors

Time frame between UKB and primary care assessments	No of participants	Agreement n (%)	Unweighted kappa	Weighted kappa
Using numerical values only				
6 months	983	597 (60.7)	0.35	0.55
1 year	1438	870 (60.5)	0.35	0.56
3 years	2836	1672 (58.9)	0.31	0.53
5 years	6623	3722 (56.2)	0.29	0.50
Using qualitative descriptors only				
6 months	3296	1988 (60.3)	0.18	0.29
1 year	5253	3145 (59.9)	0.18	0.29
3 years	9029	5441 (60.3)	0.19	0.30
5 years	10065	6027 (59.9)	0.19	0.30

UKB, UK Biobank.

considerably reduces the number of descriptors currently utilised in the primary care records. We have also shown that there may be potentially a systematic misclassification of patients in GP records. This hypothesis merits more in-depth study to confirm our preliminary findings and its repercussions for health data research.

This study has several limitations. First, we made the arbitrary decision to consider only Read Codes recorded within 5 years of the UKB assessment. It is well known that assessing drinking patterns is especially challenging compared with other health behaviours. Underreporting has been proven when compared with objective measures even at very sensitive periods like pregnancy.[44] Alcohol consumption is not necessarily stable[45] and might change considerably over time depending on different factors that cannot be assessed using RCD. However, we have not seen differences in terms of the agreement based on the period between assessments when covering a maximum span of 5 years. We cannot confirm whether or not this persistent disagreement is due to genuine changes in lifestyle over time or to inaccuracies of the primary care EHR.

Table 5 Intersource agreement of values within 5 years between both assessments using only numerical values and qualitative descriptors

	UKB					
	No alcohol	**Sensible**	**Moderate**	**Hazardous**	**Harmful**	**Total**
Numerical values only						
Primary care						
No alcohol	317	561	51	40	38	1007
Sensible	125	2874	609	394	163	4161
Moderate	7	121	201	248	160	737
Hazardous	4	32	57	149	209	451
Harmful	4	20	19	45	181	269
Total	457	3608	937	876	751	6629
Qualitative descriptors only						
Primary care						
No alcohol	237	183	9	6	5	440
Sensible	268	5508	1311	1013	549	8649
Moderate	0	2	1	3	2	8
Hazardous	18	80	121	245	421	885
Harmful	24	4	5	14	36	83
Total	547	5777	1447	1281	1013	10065

UKB, UK Biobank.

Although our objective was not epidemiological in nature, another limitation to consider is that the UKB cohort is not representative of the general population and there is evidence of a 'healthy volunteer' selection bias.[46] Researchers need to be cautious when extrapolating selected cohort results to the overall population, and in this case, this would limit our ability to estimate the alcohol consumption in the population. Instead, to evaluate the primary care EHR, our priority was to have the most accurate assessment of alcohol consumption as a comparator, so we do not consider that the lack of representativeness would affect our findings. As the availability of linked primary care data was dependent on the UKB provision we could not analyse the complete database and/or associations with different characteristics of participants.

Another potential limitation is that the decision of in which categories the qualitative Read Codes were allocated was done by applying the researcher's clinical criteria and this could inevitably have introduced new assumptions. Although most of them were very straightforward (1364-Moderate drinker—3–6 u/day), others were less obvious (E250 Inebriety NOS, allocated to Hazardous) (see online supplemental file 1X). For this reason, when planning the algorithm, we decided to prioritise, when available, those Read Codes containing units of alcohol in an attempt to minimise this bias. The finding that agreement improves by using numerical values supports this decision.

Interpretation, generalisability and future directions

Atkinson *et al* published an algorithm to categorise the EHR on smoking status with a very high agreement that demonstrates the validity of smoking status in primary care records.[47] Although our study did not confirm this for alcohol records, it is not surprising to find a lower level of agreement regarding alcohol consumption. Smoking tends to be more stable across time than drinking[45] and it is more commonly reported in a numerical form (cigarettes per day or packets per year). In addition, we have established five categories as in terms of alcohol a binary classification consumption vs non-consumption would not be very useful from a clinical point of view. In our population, 60% of the participants had at least one record regarding alcohol consumption that could be used for classification. This contrasts with previous reports that found a poorer (51.9%) recording of alcohol consumption in the UK.[48]

The poor agreement and allocation into lower categories of consumption by primary care EHR is especially significant in men. Previous research has consistently found higher alcohol consumption level among them compared with women, but this finding would merit special attention. High-risk drinkers should be targeted in preventive and risk reduction interventions. It is difficult to do this effectively if, as our results suggest, there is an underestimation of the prevalence of hazardous and harmful drinkers among half of the patients.

Primary care EHR data are potentially an accessible and valuable source of information on alcohol consumption that may be useful for a range of purposes. Our findings suggest, however, that validation with additional sources would be required before they can be used routinely to estimate alcohol consumption in the population. The relatively low level of agreement at an individual level also suggests the need for data quality improvement. Read Codes are due to be replaced by SNOMED-CT[49] in the near future in many health systems.[50] Given the numerous Read Codes available with many qualitative and unclear descriptors, we consider it would be useful to standardise alcohol recording and prioritise those containing grams or units of alcohol. This recommendation we believe is valid, both when using the current system as well as when planning the implementation of future ones. Making the process of calculating alcohol consumption for clinicians easier at the point of care, for instance, integrating calculators of units of alcohol based on the type of beverage, could improve data quality significantly. The fact that qualitative descriptors are used more often than quantitative probably reflects the fact that is easier to use these Read Codes than calculate units of alcohol consumed manually.

CONCLUSION

Considering the logistical difficulties and cost that health surveys at a population level imply and the clinical importance of having good estimates of alcohol consumption, it seems sensible to make efforts to improve the quality and accessibility of primary care EHR records. Rule-based algorithmic approaches as we have developed, are easy to adjust to local contexts to capture singularities and can easily be implemented periodically to monitor trends. This will improve the ability to plan and allocate resources based on more recent data. As the NHS and more broadly health and social care systems worldwide are starting to grasp the potentials of machine learning, the first step to build reliable prediction tools is to assure the quality and the robustness of the underlying data.[51] Improving and standardising the recording system of alcohol consumption should be a priority that would be relatively easy to implement. The analysis of this data in clinical records serves as a good example of how progress in Health Data Science can contribute to improvement in individual and societal health.

Author affiliations
[1]Population and Behavioural Science Division, School of Medicine Medical & Biological Sciences, University of St Andrews, St Andrews, UK
[2]Faculty of Medicine, Health and Human Sciences, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, New South Wales, Australia
[3]General Practice and Primary Care, Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK

Twitter David Fraile-Navarro @dafraile, Bhautesh Jani @bhauteshjani and Adeniyi Fagbamigbe @franstel74

ORCID iDs
David Fraile-Navarro http://orcid.org/0000-0002-1108-7071
Amaya Azcoaga-Lorenzo http://orcid.org/0000-0003-3307-878X
Utkarsh Agrawal http://orcid.org/0000-0001-5181-6120
Bhautesh Jani http://orcid.org/0000-0001-7348-514X
Adeniyi Fagbamigbe http://orcid.org/0000-0001-9184-8258
Dorothy Currie http://orcid.org/0000-0001-7321-9394
Alexander Baldacchino http://orcid.org/0000-0002-5388-7376
Frank Sullivan http://orcid.org/0000-0002-6623-4964

REFERENCES

1 World Health Organization (WHO). *Global status report on alcohol and health 2018*. World Health Organization, 2019.
2 Scottish public health Observatory (ScotPHO). Available: https://www.scotpho.org.uk/ pho.org.uk/behaviour/alcohol/introduction/
3 Cabinet Secretary for Health and Social Care. Scottish Health Survey – telephone survey – August/September 2020: main report [Internet], 2021. Available: https://www.gov.scot/publications/scottish-health-survey-telephone-survey-august-september-2020-main-report/documents/ [Accessed 16 Nov 2021].
4 World Health Organization (WHO). *Global health risks. mortality and burden of disease attributable to selected major risks*, 2009.
5 Puddephatt J-A, Jones A, Gage SH, *et al*. Associations of alcohol use, mental health and socioeconomic status in England: findings from a representative population survey. *Drug Alcohol Depend* 2021;219:108463.
6 Boffetta P, Hashibe M. Alcohol and cancer. *Lancet Oncol* 2006;7:149–56.
7 Piano MR. Alcohol's effects on the cardiovascular system. *Alcohol Res Curr Rev* 2017;38:219.
8 Department of Health. *UK Chief Medical Officers' Alcohol Guidelines Review: Summary of the proposed new guidelines 2015*, 2016.
9 Catto S. How much are people in Scotland really drinking? *Scottish Public Heal Obs* 2008:2–60.
10 Walters SJ, Bonacho Dos Anjos Henriques-Cadby I, Bortolami O, *et al*. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom health technology assessment programme. *BMJ Open* 2017;7:e015276.
11 Boniface S, Scholes S, Shelton N, *et al*. Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in England. *PLoS One* 2017;12:e0170892.
12 Babor TF, Steinberg K, Anton R, *et al*. Talk is cheap: measuring drinking outcomes in clinical trials. *J Stud Alcohol* 2000;61:55–63.
13 Gorman E, Leyland AH, McCartney G, *et al*. Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. *Am J Epidemiol* 2014;180:941–8.
14 Keyes KM, Rutherford C, Popham F, *et al*. How healthy are survey Respondents compared with the general population?: using Survey-linked death records to compare mortality outcomes. *Epidemiology* 2018;29:299–307.
15 Gray L, McCartney G, White IR, *et al*. Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol. *BMJ Open* 2013;3:e002647.
16 Boniface S, Kneale J, Shelton N. Drinking pattern is more strongly associated with under-reporting of alcohol consumption than socio-demographic factors: evidence from a mixed-methods study. *BMC Public Health* 2014;14:1297.
17 Livingston M, Callinan S. Underreporting in alcohol surveys: whose drinking is underestimated? *J Stud Alcohol Drugs* 2015;76:158–64.
18 Beck F, Guignard R, Legleye S. Does computer survey technology improve reports on alcohol and illicit drug use in the general population? A comparison between two surveys with different data collection modes in France. *PLoS One* 2014;9:e85810.
19 Greenfield TK, Bond J, Kerr WC. Biomonitoring for improving alcohol consumption surveys: the new gold standard? *Alcohol Res* 2014;36:39–45.
20 Nugawela MD, Langley T, Szatkowski L, *et al*. Measuring alcohol consumption in population surveys: a review of international guidelines and comparison with surveys in England. *Alcohol Alcohol* 2016;51:84–92.
21 Pollard MS, Tucker JS, Green HD. Changes in adult alcohol use and consequences during the COVID-19 pandemic in the US. *JAMA Netw Open* 2020;3:e2022942.
22 Rehm J, Kilian C, Ferreira-Borges C, *et al*. Alcohol use in times of the COVID 19: implications for monitoring and policy. *Drug Alcohol Rev* 2020;39:301–4.
23 Lee S, Xu Y, D Apos Souza AG, *et al*. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci* 2020;5:1123.
24 McDonnell L, Delaney BC, Sullivan F. Finding and using routine clinical datasets for observational research and quality improvement. *Br J Gen Pract* 2018;68:147–8.
25 Gorman E, Leyland AH, McCartney G, *et al*. Adjustment for survey non-representativeness using record-linkage: refined estimates of alcohol consumption by deprivation in Scotland. *Addiction* 2017;112:1270–80.
26 Bell S, Daskalopoulou M, Rapsomaniki E, *et al*. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. *BMJ* 2017;356:j909.
27 Holmes J, Meng Y, Meier PS, *et al*. Effects of minimum unit pricing for alcohol on different income and socioeconomic groups: a modelling study. *Lancet* 2014;383:1655–64.
28 Davies S. *UK Chief Medical Officers' alcohol guidelines review. Summary of the proposed new guidelines*. Department of Health, 2016.
29 Cheong CK, Dean L, Dougall I, *et al. The Scottish health survey 2018 edition; amended in February*, 2020.

30 Jani BD, McQueenie R, Nicholl BI, *et al*. Association between patterns of alcohol consumption (beverage type, frequency and consumption with food) and risk of adverse health outcomes: a prospective cohort study. *BMC Med* 2021;19:1–14.

31 Booth N. What are the read codes? *Health Libr Rev* 1994;11:177–82.

32 Sudlow C, Gallacher J, Allen N, *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779.

33 Feunekes GI, van 't Veer P, van Staveren WA, *et al*. Alcohol intake assessment: the sober facts. *Am J Epidemiol* 1999;150:105–12.

34 Read Codes - NHS Digital [Internet]. Available: https://digital.nhs.uk/services/terminology-and-classifications/read-codes [Accessed 14 Apr 2021].

35 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.

36 Bowker AH. A test for symmetry in contingency tables. *J Am Stat Assoc* 1948;43:572–4.

37 McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12:153–7.

38 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.

39 Rahardja D, Yang Y, Zhang Z. A comprehensive review of the two-sample independent or paired binary data, with or without stratum effects. *J Mod Appl Stat Methods* 2016;15:215–23.

40 Tang W, Hu J, Zhang H, Wan T, Jun HU, Hui ZHANG PWU, *et al*. Kappa coefficient: a popular measure of rater agreement. *Shanghai Arch Psychiatry* 2015;27:62.

41 Van Rossum G, Drake Jr FL. *Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam*, 1995.

42 Team RC. *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria, 2016.

43 UK Biobank. Available: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us

44 Abernethy C, McCall KE, Cooper G, *et al*. Determining the pattern and prevalence of alcohol consumption in pregnancy by measuring biomarkers in meconium. *Arch Dis Child Fetal Neonatal Ed* 2018;103:F216–20.

45 Room R. Smoking and drinking as complementary behaviours. *Biomed Pharmacother* 2004;58:111–5.

46 Fry A, Littlejohns TJ, Sudlow C, *et al*. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017;186:1026–34.

47 Atkinson MD, Kennedy JI, John A, *et al*. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. *BMC Med Inform Decis Mak* 2017;17:1–12.

48 Mansfield K, Crellin E, Denholm R, *et al*. Completeness and validity of alcohol recording in general practice within the UK: a cross-sectional study. *BMJ Open* 2019;9:e031537.

49 Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 2006;121:279.

50 ISD Services | Terminology Services and Clinical Coding | Coding & Terminology Systems | ISD Scotland [Internet]. Available: https://www.isdscotland.org/products-and-services/terminology-services/coding-and-terminology-systems/#SNOMED-CT [Accessed 05 May 2021].

51 Verheij RA, Curcin V, Delaney BC, *et al*. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res* 2018;20:e9134.