International Journal of Population Data Science

Journal Website: www.ijpds.org

Health and household environment factors linked with early alcohol use in adolescence: a record-linked, data-driven, longitudinal cohort study

Amrita Bandyopadhyay[1,2], Sinead Brophy[1,2,3], Ashley Akbari[1,3], Joanne Demmler[3], Jonathan Kennedy[2], Shantini Paranjothy[4,5], Ronan A. Lyons[1,2,3], and Simon Moore[4,6,*]

Submission History

Submitted:	16/11/2021
Accepted:	06/05/2022
Published:	07/07/2022

[1]Administrative Data Research Wales, Swansea University Medical School, Wales SA2 8PP, UK
[2]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Wales, SA2 8PP, UK
[3]Health Data Research UK, Swansea University Medical School, Wales, SA2 8PP, UK
[4]School of Dentistry, Cardiff University, Cardiff, Wales, CF14 4XY, UK
[5]University of Aberdeen, Aberdeen Health Data Science Centre, Institute of Applied Health Sciences, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD
[6]Security, Crime and Intelligence University Innovation Institute, Social Science Research Park, Cardiff University, Maindy Road, Cardiff, CF24 4HQ, UK

Abstract

Introduction
Early alcohol use has significant association with poor health outcomes. Individual risk factors around early alcohol use have been identified, but a holistic, data-driven investigation into health and household environmental factors on early alcohol use is yet to be undertaken.

Objectives
This study aims to investigate the relationship between preceding health events, household exposures and early alcohol use during adolescence using a two-stage data-driven approach.

Methods
In stage one, a study population (N = 1,072) were derived from the Millennium Cohort Study (MCS) Wales (born between 2000–2002). MCS data were first linked with electronic-health records. Factors associated with early (<= eleven years old) alcohol use were identified using feature selection and stepwise logistic regression. In stage two, analogous risk factors from MCS were recreated for whole population (N = 59,231) of children (born between 1998-2002 in the Welsh Demographic Service Dataset) using routine data to predict the alcohol-related health events in hospital or GP records.

Results
Significant risk factors from stage two included poor maternal mental (adjusted odds ratio [aOR] = 1.31) and physical health (aOR = 1.25), living with someone with alcohol-related problem (aOR = 2.16), single-adult household (aOR = 1.45), ever in deprivation (aOR = 1.66), child's high hyperactivity (aOR = 3.57), and conduct disorder (aOR = 3.26). Children with health events, whose health needs are supported (e.g., are taken to the doctor), are at lower risk of early alcohol use.

Conclusion
Health events of the family members and the child can act as modifiable exposures and may therefore inform the development of prevention initiatives. Families with known alcohol problems, living in deprivation, experiencing child behavioural problems and those who are not taken to the doctor are at higher risk of early drinking behaviour and should be prioritised for early years support and interventions to target problem drinking in young people.

Keywords
alcohol; adolescent; data linkage; electronic health records (EHRS); cohort study

*Corresponding Author:
Email Address: MooreSC2@cardiff.ac.uk (Simon Moore)

https://doi.org/10.23889/ijpds.v7i1.1717

Introduction

Alcohol use in childhood is associated with the risk of later alcohol abuse, alcohol dependence [1] and several negative outcomes including poor educational achievement, death and disability [2–5]. Known factors that predict early alcohol use include a child's hyperactivity and conduct disorder [6, 7], lack of family support, household dysfunction, parental alcohol drinking pattern, parental indifference towards young persons' alcohol use [8–11] and adverse childhood experiences (ACEs) (e.g., child abuse and parental discord) [12]. Current research has largely focused on the family environment, individual level socio-demographic, neurocognitive, behavioural or emotional features, individually or in combination [13–15]. Although it is known that ACEs have a detrimental impact on a child's health in early life [16, 17], it is not known whether a child's own health status is associated with subsequent alcohol use and alcohol-related health outcomes.

Child health is a broad term that includes maintaining and protecting physical, mental and social health [18]. Broadly, there are two dominant methodological approaches in the investigation of child alcohol use that are increasingly regarded as complementary [19]. First, survey methodology allows researchers to focus on specific exposures and outcomes, such as volume of alcohol consumed, and to tailor validated [20] instruments to address preconceived study hypothesis [2]. Limitations include relatively small sample size, non-response, selection and volunteer bias [21]. Second, the analysis of routinely collected electronic health records (EHRs) facilitates the inclusion of a greater number of individuals, even entire populations, than is feasible using surveys. The analysis of whole population EHRs, however, imposes challenges relating to the processing and management of data, including addressing missing data on informative variables [22]. For example, EHRs are unlikely to capture occasional alcohol consumption but would be expected to capture health outcomes relating to hazardous alcohol use.

Existing literature on this topic has predominantly focused on preconceived study hypothesis [2], however this increases the chance of missing risk factors which have not already been identified. In contrast to this, a data-driven framework would avoid the limits of a pre-defined and hypothesis-bound investigation and significantly open up the exploration of the variable space. We anticipate that this will provide new insights and will ultimately help to develop a better understanding of the research problem under investigation. Hence, the current study does not focus on an explicit causal analysis, rather we aim to merge hypothesis-based knowledge with data-driven insights to investigate the risk factors associated with early alcohol use.

In this study we assess the relationship between childhood health factors, household environment and alcohol-related outcomes during adolescence using a two-stage data-driven approach. These broad categories of risk factors were based on hypothesis-based knowledge as discussed above. This method brings together a hypothesis-based study design followed by a data-driven approach which complements and minimises the limitation of both study designs.

Methods

A two-stage data-driven approach has been undertaken to investigate the association between the specific risk factors and the outcome in this study. In stage one, a machine learning feature selection algorithm and a classifier were used to identify the health conditions and socio-demographic factors associated with early alcohol use from linked EHRs and Millennium Cohort Study (MCS) survey data. In stage two, analogous risk factors identified from stage one were then sought in routine data and an analytic approach was used to determine the prediction model. The linked routinely collected EHRs and vast volume of administrative data from the whole population of Wales was analysed to determine the effect of the risk factors identified in the MCS data analysis as predictors to target alcohol-related health outcomes in the general adolescent population.

Stage one – Millennium Cohort Study (MCS)

Participants

The MCS is a longitudinal birth cohort of children born in the UK between the years 2000 and 2002 [23]. Parents of the original 18,819 singleton children were interviewed from all parts of UK when their child was nine months old, of those 1,951 were interviewed in Wales. Subsequent interviews took place at ages three, five, seven and eleven years of age. Written consent to link MCS children with their routine EHRs up to age fourteen years was obtained from their parents at the interview undertaken when children were seven years of age. Data of the 1,838 consented singleton children resident in Wales was subsequently linked with their EHRs. The study population included children who also participated in the interview at age eleven years, as the primary outcome data were collected at that point. The current study excluded participants who did not have a general practitioner (GP) record in the Welsh Longitudinal General Practice (WLGP) dataset before they were eleven years of age (Supplementary Figure 1).

Exposure

The study included parent reported socio-demographic and family-related variables for children from MCS interviews which took place between the age of nine months and seven years of the children. These include child's sex, mother's socio-economic classification (SEC), household poverty level (whether the household income was above/below 60% of national median using a modified Organisation for Economic Co-operation and Development scale), living area (based on 2005 Rural/Urban Area Classification), mother's alcohol use during and post pregnancy, lone parent carer, and number of children. Based on lone parent status, the total number of siblings at household and total number of household members, the study derived a binary variable to identify whether the child was residing with any other additional household members. Using both parents' responses on alcohol consumption, guardian alcohol use variables were derived. Children's emotional and behavioural difficulties were measured using the parent completed Strength and Difficulty Questionnaire (SDQ) [24]. Since most of these variables are time varying (and collected from MCS at ages nine months

until age eleven years) aggregated summary variables were derived based on average values. These variables include SDQ, mother's SEC, lone parent status, guardians' alcohol use, living area, poverty indicator, additional household member and mother's alcohol use after their child was born. The exposure variables from MCS have been described in Table 1.

The health records of the children were also considered as the exposures for risk of early alcohol use. EHRs of the MCS children obtained from hospital admission record and primary care events within the Patient Episode Database for Wales (PEDW) and the WLGP dataset. A broad list of explanatory health codes was constructed using the three-digit ICD-10 codes and Read Code Version 2 recorded in PEDW and WLGP from birth until age ten (one year before the alcohol data were collected). Wales Electronic Cohort for Children (WECC) [25] containing further details on child health in Wales, were used to obtain age and maternal age at birth.

Outcome

Alcohol data for MCS children were obtained from a self-report questionnaire at age eleven (Supplementary Table 1). Based on the responses to the questionnaire the children were classified into two groups: those who had consumed alcohol (case) and those who had not (non-case). Those who did not answer or provided contradictory responses were removed from analyses (Supplementary Figure 1).

Statistical analysis

In the cohort exposure dataset, the participants with more than 10 missing variables (out of 13) were removed from analyses to ensure the accuracy of the data. An explanatory variable with less than 10% missing data had been imputed using a predictive mean matching (PMM) imputation method [26, 27].

To identify the health codes that were associated with early alcohol use from the large volume of linked EHRs spanning 10 years, a chi-square (χ^2) feature selection method was applied [28]. A critical threshold value $\chi^2 \geq 2.706$ (one degree of freedom, $p \leq 0.1$) was applied and health codes with a χ^2 above this threshold were retained in subsequent analyses. A multivariate stepwise logistic regression with bidirectional (forward and backward) search was then performed for the exposure variables to obtain the best-fit model [29]. In stepwise model the variables with least significance were removed at each iteration step and the final model was selected based on the minimum Akaike Information Criterion (AIC) value. From the final model, only the statistically significant ($p \leq 0.05$) variables were selected as significant predictors associated with the risk of early alcohol use leading to a further reduction in variable space. This is justified due to the following reasons.

- The variable selection process facilitates the choice of best model by incorporating the interdependence between the explanatory variables.

- The approach only considers the statistically significant variables for the stage two analysis which reduces the variable space and optimises the time to recreate analogous variables.

Stage two – whole population

Participants

All children born between 1st January 1998 and 31st December 2002 and were resident in Wales during the first fourteen years of their life were included in the whole population dataset. The study population was selected from the Welsh Demographic Service Dataset (WDSD), which is an administrative dataset of individuals living in Wales registered with a GP. The participants without continuous record in the WLGP from age six months to fourteen years were excluded to ensure a complete follow-up period.

Exposure

Analogous risk factors to those identified in the MCS analysis were created using the WDSD, WLGP and PEDW data. The study used an encrypted household identifier known as residential anonymised linking field (RALF) which enabled the participants to be linked with other household members and related records [30]. Each RALF is associated with the smallest geographical representation known as lower super output area (LSOA) which again is associated with a Welsh Index of Multiple Deprivation (WIMD) rank aggregated into a quintile or decile scale. Overall and employment WIMD scores were used as the measure of deprivation from routine data in the study. The main explanatory variables derived from routine data for the whole population analysis include child's sex, employment deprivation and overall deprivation, living with single adult, mother's alcohol-related condition during pregnancy, living with household member with alcohol-related condition, living area, maternal age, gestational age, and child mental and physical health. To be consistent with the MCS data, primary exposure data were collected for children up to age seven years. For time varying variables, the study used the same time points as MCS (birth to nine months, nine months to three years, three to five years, and five to seven years) and derived aggregated summary variables for the risk factors. Detailed descriptions of the variables are available in Supplementary Table 2.

Outcome

Alcohol-related health events across the whole population cohort were obtained from ICD-10 codes in PEDW (Supplementary Table 3) and Read codes in WLGP (Supplementary Table 4) between the age seven and fourteen years [31].

Statistical analysis

As the case (alcohol-related EHRs) to non-case (no alcohol-related EHRs) ratio was 1:99 in the whole population cohort and unbalanced, to improve the efficiency and the sensitivity of model performance case-control selection was undertaken by randomly selecting 20 non-cases for each sex matched case [32]. The dataset was randomly split into a training (70%) and test set (30%). Logistic regression was used to obtain the best-fit model on the training data. Model prediction on the test data provided a predictive probability of the expected outcome associated with each individual. Model prediction

Table 1: Socio-demographic characteristics of the MCS population (following imputation) and whole population sample with descriptive statistics

MCS			**Whole Population**		
Child Sex	n	%		n	%
Female	521	48.60	Female	28,770	48.57
Male	551	51.40	Male	30,461	51.43
Deprivation					
Mother Socio economic classification (SEC)			Overall deprivation		
Always managerial or intermediate	377	35.17	Low (WIMD quintile >=3)	29,102	49.13
Always semi-employed, self-employed, semi-routine or routine	280	26.12	High (WIMD quintile <3)	24,701	41.70
Unknown	415	38.71	Borderline (ever belong to high group but not always)	5,428	9.16
Poverty indicator			Employment deprivation		
Above poverty level	539	50.28	Low (WIMD quintile >= 3)	29,394	49.63
Below poverty level	270	25.19	High (WIMD quintile < 3)	24,774	41.83
Ever been below poverty level	263	24.53	Borderline (ever belong to high group but not always)	5,063	8.55
Household alcohol use					
Mother's alcohol use during pregnancy			Mother's alcohol-related health condition during pregnancy		
Never	752	70.15	No	55,251	93.28
Low (less than once a month or 1–2 times a month)	218	20.34	Yes	3,980	6.72
High (more than 1–2 times a month)	102	9.51			
Mother's alcohol use after child was born					
Never	82	7.65			
Low	500	46.64			
High	490	45.71			
Guardian alcohol use			Household member identified with alcohol-related hospital admission		
Low	247	23.04	No	57,799	97.58
Moderate	524	48.88	Yes	1,432	2.42
High	233	21.74			
Variable	68	6.34			
Living area					
Rural	238	22.20		14,760	24.92
Urban	779	72.67		41,907	70.75
Ever been urban	55	5.13		2,564	4.33
Maternal age at child's birth					
Less than 20 years	102	9.51		7,111	12.01
20 to 24 years	202	18.84		9,266	15.64
25 to 29 years	305	28.45		17,389	29.36
30 to 34 years	324	30.22		17,005	28.71
35 years and over	139	12.97		8,460	14.28
Gestational age					
Not term	52	4.85		1,317	2.22
Term	1,020	95.15		57,914	97.78
Household composition					
Siblings at home			Living with single adult		
No sibling	129	12.03	No	33,662	56.83
One sibling always or at some point	493	45.99	Yes	8,425	14.22
More than one sibling ever	450	41.98	Ever been	17,144	28.94

(Continued).

Table 1: Continued

MCS			Whole Population		
Lone parent					
No	754	70.34			
Yes	130	12.13			
Ever been	188	17.54			
Additional household member					
No	792	73.88			
Yes	118	11.01			
Ever had	162	15.11			
Mother's health					
Longstanding illness			Mother's any comorbidity		
No	589	54.94	No	46,170	77.95
Yes	170	15.86	Yes	13,061	22.05
Varies	313	29.20	Mother's psychosis disorder		
			No	58,924	99.48
			Yes	307	0.52
			Mother's common mental health condition		
			No	28,603	48.29
			Yes	30,628	51.71

Table 2: Health codes identified as risk factors for early alcohol use by chi-square feature selection method in the MCS cohort and the percent of sample with these codes present in whole population (WP) following selection

Health code	Description of the code	Type of code	chi-square	MCS (%)	WP (%)
Read code H05%	Upper respiratory infections	Diagnosis	.60	62.50	59.95
Read code K2%	Male genital organ diseases	Diagnosis	7.77	12.41	8.46
Read code 919%	Child health surveillance related administrative code	Administrative	6.07	25.56	30.70
Read code 64N%	Child physical health examination	Administrative	4.63	17.35	15.56
Read code 656%	Tetanus vaccination	Administrative	4.11	28.26	34.21
ICD-10 code Z%	Factors influencing health status and contact with health services	Diagnosis	3.90	27.99	20.68
Read code 654%	Diphtheria vaccination	Administrative	3.69	27.71	-
Read code 655%	Pertussis vaccination	Administrative	3.35	29.94	-
Read code F%	Nervous system and/or sense organ diseases	Diagnosis	3.04	70.24	-
Read code F4%	Disorders of eye and adnexa	Diagnosis	3.00	46.27	-
Read code K27%	Disorders of penis	Diagnosis	2.99	9.42	–
Read code etc.%	Trimethoprim, an antibiotic used mainly in the treatment of bladder infections	Medication	2.93	16.70	–
Read code 4%	Laboratory test and procedures (e.g. urine culture, blood test)	Administrative	2.89	60.73	–

codes were not selected by the logistic regression models, hence were not selected for WP analysis

was quantified by performance accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

MCS and routine EHRs were anonymously linked and accessed within the Secure Anonymised Information Linkage (SAIL) Databank. Linkage was completed using an encrypted person-based identifier known as the anonymised linkage field (ALF), generated by the Digital Health and Care Wales (DHCW) [33, 34]. Data preparation (extraction, cleaning, and linkage) was performed in Structured Query Language (SQL) on an IBM DB2 platform, with subsequent analyses performed in R v3.3.2 [35].

Results

Stage one – MCS

Among the consented singleton children 1,838 were assigned an ALF, with 82% of the children having a GP registration record in SAIL before age eleven years (Supplementary Figure 1). Individual and household characteristics (following imputation) are described in Table 1. 7.6% of the MCS children were considered as 'case' based on their response. Health codes (256 ICD-10 and Read codes) were obtained

after merging the first ten years of EHRs from PEDW and WLGP. Feature selection method reduced this to 13 health features (Table 2).

After merging health and socio-demographic variables, 31 main explanatory variables (13 health codes and 18 socio-demographic variables) were available for the two-way logistic model. The final 19 features with significant p values were considered to be significantly associated with the risk profile of early alcohol use (Table 3).

Stage two – whole population

In Wales, 207,114 children were born in between 1st January 1998 and 31st December 2002, and their records were obtained from WDSD. After applying exclusion criteria there were 59,231 children as the study population (Supplementary Figure 2). Of the study population, 591 (0.99%) children had at least one alcohol-related event between seven and 14 years of age (Supplementary Figure 3) who were the cases from the whole population subset. After applying case control selection, the dataset had 591cases and 11,820 non-cases, which were further split into training and test set. There were 8,688 (417 cases and 8,271 non-cases) children in the training dataset. The variables identified as significantly associated with early alcohol use using MCS data were mapped into the whole population cohort (Supplementary Table 2). Table 1 presents descriptive statistics for this population. Mothers of 6.72% of the children had an alcohol-related event reported in PEDW or WLGP while pregnant. 2.42% children lived with a household member who had alcohol-related inpatient hospital admission. The adjusted odds ratio of the features with 95% confidence interval are presented in Table 4 (also see Supplementary Figure 4).

The model was run on the test dataset. The accuracy of the model was 61.32% with a sensitivity of 58.05% and specificity of 68.48% (additional details are provided in Supplementary Tables 5, 6). Out of 174 cases, the model was able to predict 101 (58%) children who had an alcohol-related health event recorded in the healthcare system between ages seven and fourteen.

Discussion

This study has developed a two-stage data-driven framework that can create a profile of the characteristics of children who end up with an alcohol problem in adolescence. The study undertook data linkage between a longitudinal survey data (MCS) and routine EHRs in stage one to select the significant risk factors associated with early alcohol use. Stage two built the analogous risk factors using only the linked routine data and based this, a prediction model was developed. Hybridisation of these two powerful data sources (routine and survey) enabled us to create a data-driven risk profile. The risk factors were significantly associated across both MCS and whole population analyses, but effect estimates varied. Children whose health needs are supported are at lower risk of early alcohol use, evidenced by protective effect of receiving vaccinations, attending routine health examinations with their GP, and contact with health services recorded in primary and secondary care were consistent across MCS and whole population analyses. Similarly, children with health codes relating to acute upper respiratory infections may have more protective guardians willing to consult medical professionals for mild conditions. Together, this suggests that the avoidance of regular healthcare contact is an indicator that increases the risk of early alcohol use. However, the trends relating to the two codes, the child surveillance administration code and the chapter heading linked to male genitals, differed between the whole population and the MCS analysis. The code linked to male genitals showed an association with higher risk of alcohol use in MCS but was statistically inconclusive for the whole population analysis. The child surveillance administration code was associated with higher risk for the MCS cohort in contrast to the whole population which can be attributed to the differential support received by two cohorts which was not captured by the data and hence this requires further investigation. Also, the proportion of cases obtained from MCS data (stage one) were higher than those obtained from the whole population data (stage two). This can be attributed to the fact that cases from stage one were based on the self-reported alcohol consumption data whereas the stage two routine data highlighted the most severe cases caused by alcohol among the adolescents and recorded on the healthcare system.

The overall risk profile obtained from MCS and whole population analyses were broadly consistent with each other and the research literature generally both in the UK and internationally. Similar risk factors include being male [13], ever living in an urban environment where there is a greater density of alcohol outlets [36], ever living in conditions of social deprivation, living in a household with higher level of alcohol use by household members [9]. Studies from USA highlighted that early onset of alcohol use was significantly associated with parental drinking pattern and living in a lone parent household [11], child's attention deficit hyperactivity disorder (ADHD) and conduct disorder [6, 7]. The stage one MCS analysis in this study revealed that emotional difficulty and a higher level of behavioural difficulty (as assessed by parents) were associated with a reduced risk of alcohol use. However, diagnosis of clinically relevant behavioural/emotional problems was protective in the population model. Poor maternal mental health was linked with adverse outcomes, consistent with family-level risk factors that promote children's alcohol use [12, 17]. A difference was observed in regards to the effect of maternal age at birth on the risk of a child's early alcohol use. The protective effect of higher maternal age was observed for the whole population but the finding on MCS data differed and requires further investigation. Further, employment deprivation in the whole population analysis was associated with lower risk of a child's early alcohol use after adjusting for overall deprivation. This finding is similar to the existing literature [15, 37], which found that early alcohol use is more common in higher income families. This suggests that reliance on employment indicators is not sufficient to understand the socio-economic factors influencing a child's early alcohol use, the overall deprivation (also measured by education, health, access to the service, physical environment of living) plays an important role as well.

The result of this study needs to be interpreted in conjunction with a number of limitations. Firstly, mapping the MCS survey to the routine data was challenging, not all

Table 3: The explanatory variables associated with higher and lower risk of early alcohol use for the MCS children (Stage one analysis) with the adjusted Odds Ratio (OR) and 95% confidence interval (CI)

Feature	Adjusted OR (95%CI)
Child's sex	
Female	1
Male	3.06 (2.35 to 3.99)***
Mother's Socio-economic classification (SEC)	
Always Managerial or intermediate	1
Always semi-employed, self-employed, semi-routine or routine	1.30 (0.93 to 1.81)
Unknown	1.94 (1.37 to 2.74)***
Lone parent	
Never lone parent	1
Lone parent	1.68 (1.07 to 2.65)*
Ever been	1.77 (1.27 to 2.49)**
Mother alcohol use during pregnancy	
Never	1
Low (less than once a month, 1–2 times a month)	2.48 (1.83 to 3.38)***
High	5.38 (3.58 to 8.15)***
Mother alcohol use after child was born	
Never	1
Low	1.15 (0.70–1.92)
High	0.70 (0.04 to 1.24)
Guardian alcohol use	
Low	1
Moderate	1.73 (1.22 to 2.25)**
High	1.07 (0.70 to 1.64)
Variable	0.91 (0.48 to 1.70)
Living area	
Rural	1
Urban	1.61 (1.17 to 2.23)**
Ever been urban	4.54 (2.69 to 7.75)***
Poverty indicator	
Above poverty level	1
Below poverty level	0.93 (0.60 to 1.45)
Ever been below poverty level	1.33 (0.95 to 1.86)
Maternal age at child's birth	
Less than 20 years	1
20 to 24 years	1.57 (0.97 to 2.58)
25 to 29 years	3.28 (2.03 to 5.36)***
30 to 34 years	2.68 (1.64 to 4.43)***
35 years or over	0.65 (0.35 to 1.21)
Gestational age	
Not term	1
Term	9.42 (4.22 to 23.03)***
Additional household member	
No	1
Yes	0.69 (0.45 to 1.06)
Ever had	0.57 (0.39 to 0.81)**
Hyperactivity	
Always normal	1
Any mention of higher level of hyperactivity	1.84 (1.37 to 2.47)***
Conduct disorder	
Always normal	1
Any mention of higher level of CP	2.10 (1.57 to 2.82)***
Emotional difficulty	
Always normal	1
Any mention of higher level of ED	0.68 (0.48–0.97)*

(Continued).

Table 3: Continued

Feature	Adjusted OR (95%CI)
Total Difficulty Score	
Always normal	1
Any mention of higher level of TDS	0.45 (0.31 to 0.66)***
Mother longstanding illness	
No	1
Yes	1.53 (1.09 to 2.16)*
Varies	1.25 (0.96 to 1.65)
Other acute upper respiratory infections (Read code H05%)	
No	1
Yes	0.43 (0.34–0.55)***
Male genital organ diseases (Read code K2%)	
No	
Yes	2.77 (1.58–4.94)***
Child surveillance administration (Read code 919%)	
No	
Yes	1.38 (1.06 to 1.81)*
Child exam (Read code 64N%)	
No	
Yes	0.51 (0.35 to 0.75)**
Tetanus vaccination (Read code 656%)	
No	
Yes	0.60 (0.45 to 0.79)***
General examination (ICD10 code Z%)	
No	
Yes	0.73 (0.55 to 0.99)*
Disorders of penis (Read code K27%)	
No	
Yes	0.63 (0.33 to 1.19)

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

MCS variables were available in the routine data. In some instances, multiple variables had to be merged to derive summary variables. This may result in a degree of uncertainty about the information captured in the summary variables. Secondly, it was necessary to aggregate some time-varying variables into a single point estimate and, as such, the analyses are unable to capture how the recency of some events might influence results. Thirdly, due to unavailability of continuous GP records of some participants between six months and fourteen years (if the participants changed their GP and the their registered GP was not contributing to SAIL), they were removed from the whole population analysis. Similarly, the follow-up of children was not possible where they who moved out of the study area (Wales, UK), or died under age fourteen, because of which their exposure (sociodemographic and health related data) and outcome (alcohol data) data were not available. This resulted in a large reduction of the number of children in the study population. However, this did not contribute to selection bias as this happened randomly and the losses had no direct relationship with alcohol-related outcome. Fourthly, the EHRs did not include Emergency Department (ED) attendance data (but does include admissions into hospital via the ED) as there are no uniformly applicable codes for alcohol-related attendances in ED, and even when available, these are sparsely populated [38]. Lastly, in this study the model performance, measured by sensitivity and specificity, was moderate. However, even if we had a sensitivity and specificity of 90% the maximum positive predictive value, we can get is 31%, given the low prevalence of alcohol-related medical contact, as the prevalence influences the positive and negative predictive value of a model performance [39]. Machine learning approaches generally aim to achieve the best predictive models from the available data. The low positive predictive value, obtained here, suggests that the variables needed to improve model performance are not available in the data (e.g., genetic information, peer alcohol-related data).

Routine EHRs and administrative data are available to healthcare professionals and are used by policy makers and commissioners to determine how resources are best utilised to manage preventive interventions. However, the bulk of research considering early alcohol use and related outcomes has relied on self-report surveys. It has been shown that linking survey and routine data can offer new insights [40]. The results presented here are novel in that our approach generalised results from an established survey to a whole population analysis using predictive analytic techniques. This provides in-depth knowledge about the profile of the children susceptible to early alcohol use and can feasibly be used to inform population health strategies designed to reduce the

Table 4: The explanatory variables associated with higher and lower risk of early alcohol-related health outcomes for the whole population (Stage two analysis) with the adjusted Odds Ratio (OR) and 95% confidence interval (CI)

Feature	**Adjusted OR (95% CI)**
Child's Sex	
Female	1
Male	1.09 (1.02 to 1.17)**
Overall deprivation:	
Low	1
High	1.11 (0.98 to 1.25)
Borderline	1.66 (1.41 to 1.95)***
Employment deprivation:	
Low	1
High	0.84 (0.75 to 0.95)**
Borderline	0.82 (0.69 to 0.97)*
Living with single adult:	
No	1
Yes	1.45 (1.32 to 1.59)***
Ever been	1.17 (1.08 to 1.26)***
Mother's alcohol-related condition during pregnancy	
No	1
Yes	0.88 (0.77 to 1.00)*
Household member with alcohol-related condition	
No	1
Yes	2.16 (1.80 to 2.60)***
Living area	
Rural	1
Urban	0.99 (0.92 to 1.08)
Ever in urban	2.42 (2.08 to 2.81)***
Maternal age at birth	
Less than 20 years	1
20 to 24 years	0.88 (0.79 to 0.99)*
25 to 29 years	0.79 (0.71 to 0.87)***
30 to 34 years	0.68 (0.61 to 0.76)***
35 years or over	0.53 (0.46 to 0.60)***
Gestational age	
Not-term	1
Term	1.11 (0.89 to 1.40)
Child – Attention deficit hyperactive disorder (ADHD)	
No	1
Yes	3.57 (2.52 to 5.15)***
Child - Conduct disorder	
No	1
Yes	3.26 (2.14 to 5.07)***
Child – Depression/Anxiety	
No	1
Yes	0.75 (0.34 to 1.69)
Mother's any comorbidity	
No	1
Yes	1.25 (1.16 to 1.34)***
Mother's common mental health condition	
No	1
Yes	1.31 (1.23 to 1.40)***
Mother's psychosis disorder	
No	1
Yes	3.12 (2.04 to 4.90)***

(Continued).

Table 4: Continued

Feature	Adjusted OR (95% CI)
Other acute upper respiratory infections (Read code H05%)	
No	1
Yes	0.97 (0.91 to 1.04)
Male genital organ diseases (Read code K27%)	
No	1
Yes	0.90 (0.79 to 1.02)
Child surveillance administration (Read code 919%)	
No	1
Yes	0.80 (0.75 to 0.86)***
Tetanus vaccination (Read code 656%)	
No	1
Yes	0.47 (0.44 to 0.51)***
Child exam (Read code 64N%)	
No	1
Yes	0.59 (0.53 to 0.65)***
General examination (ICD10 code Z%)	
No	1
Yes	0.84 (0.78 to 0.92)***

Note: *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001.

prevalence of early alcohol use in children and related health outcomes.

Conclusions

The hybridisation of data of different nature, as carried out in this study, is a novel approach that combines the complementary advantages of EHRs with more personal insights from questionnaire-based cohort data. This provides a robust resource on which findings can be based and generalised to the wider population. The identified risk factors such as living with a single parent, alcohol problem in the household, social deprivation and children receiving poor support from the healthcare system indicate that involvement and support for the family is important in breaking cycles and improving children's outcomes.

Acknowledgements

This research has been carried out as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1). This work was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR). The research was supported by DECIPHer, a UKCRC Public Health Research Centre of Excellence, which receives funding from the British Heart Foundation, Cancer Research UK, Medical Research Council, the Welsh Government and the Wellcome Trust (WT087640MA), under the auspices of the UK Clinical Research Collaboration. This work was supported by Health Data Research UK which receives its funding from HDR UK Ltd (NIWA1) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Welcome Trust.

The authors are grateful to the Centre for Longitudinal Studies, UCL Institute of Education and the UK Data Service. The co-operation of the participating Cohort families is also gratefully acknowledged. This work uses data provided by patients and collected by the NHS as part of their care and support. This study used anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research. Compliance with ethical standards.

Funding

This work was supported by funds from the Economic and Social Research Council, the Medical Research Council and Alcohol Research UK to the ELAStiC Project (ES/L015471/1).

The study funders had no involvement in the study design; the collection, analysis, and interpretation of data; the writing

of the report; and the decision to submit the paper for publication.

Dedication

This work was designed with Professor Damon Berridge. Damon passed away April 12th, 2019, and is greatly missed by us all.

Contributorship statement

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Amrita Bandyopadhyay and Sinead Brophy. The first draft of the manuscript was written by Amrita Bandyopadhyay, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conceptualization: Sinead Brophy and Amrita Bandyopadhyay; Methodology: Amrita Bandyopadhyay, Damon Berridge, and Sinead Brophy; Formal analysis and investigation: Amrita Bandyopadhyay Writing - original draft preparation: Amrita Bandyopadhyay; Writing - review and editing: Simon Moore, Sinead Brophy, Ashley Akbari, Joanne Demmler, Shantini Paranjothy, Jonathan Kennedy and Ronan A Lyons; Funding acquisition: Simon Moore, Shantini Paranjothy and Ronan A Lyons; Resources: Ashley Akbari; Supervision: Sinead Brophy and Simon Moore.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics statement

Ethics approval for the fourth survey of the Millennium Cohort Study was received from the Northern and Yorkshire Research Ethics Committee (07/MRE03/32). This study was approved by the SAIL Databank independent Information Governance Review Panel (IGRP) (project number 0336).

References

1. Hingson RW, Heeren T, Winter MR. Age at Drinking Onset and Alcohol Dependence: Age at Onset, Duration, and Severity. *Arch Pediatr Adolesc Med*. 2006; 160(7):739–746. https://doi.org/10.1001/archpedi.160.7.739

2. Bi J, Sun J, Wu Y, Tennen H, Armeli S. A Machine Learning Approach to College Drinking Prediction and Risk Factor Identification. *ACM Trans Intell Syst Technol*. 2013;4(4):72:1–72:24. https://doi.org/10.1145/2508037.2508053

3. Hingson RW, Zha W, Weitzman ER. Magnitude of and Trends in Alcohol-Related Mortality and Morbidity Among U.S. College Students Ages 18–24, 1998–2005. *J Stud Alcohol Drugs Suppl*. 2009;(16):12–20. Accessed August 2, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701090/

4. National Institute on Alcohol Abuse and Alcoholism. Underage Drinking: A Major Public Health Challenge – Alcohol Alert No. 59. Published April 2003. Accessed August 2, 2018. https://pubs.niaaa.nih.gov/publications/aa59.htm

5. Office of the Surgeon General (US), National Institute on Alcohol Abuse and Alcoholism (US), Substance Abuse and Mental Health Services Administration (US). *The Surgeon General's Call to Action To Prevent and Reduce Underage Drinking*. Office of the Surgeon General (US); 2007. Accessed April 3, 2019. http://www.ncbi.nlm.nih.gov/books/NBK44360/

6. Molina BS, Pelham WE. Childhood predictors of adolescent substance use in a longitudinal study of children with ADHD. *J Abnorm Psychol*. 2003;112(3): 497–507.

7. Sibley MH, Pelham WE, Molina BSG, et al. The role of early childhood ADHD and subsequent CD in the initiation and escalation of adolescent cigarette, alcohol, and marijuana use. *J Abnorm Psychol*. 2014;123(2):362–374. https://doi.org/10.1037/a0036585

8. Kelly Y, Goisis A, Sacker A, Cable N, Watt RG, Britton A. What influences 11-year-olds to drink? Findings from the Millennium Cohort Study. *BMC Public Health*. 2016;16(1):169. https://doi.org/10.1186/s12889-016-2847-x

9. Mahedy L, MacArthur GJ, Hammerton G, et al. The effect of parental drinking on alcohol use in young adults: the mediating role of parental monitoring and peer deviance. *Addiction*. 2018;113(11):2041–2050. https://doi.org/10.1111/add.14280

10. Simantov E, Schoen C, Klein JD. Health-Compromising Behaviors: Why Do Adolescents Smoke or Drink?: Identifying Underlying Risk and Protective Factors. *Arch Pediatr Adolesc Med*. 2000;154(10):1025–1033. https://doi.org/10.1001/archpedi.154.10.1025

11. Donovan JE, Molina BSG. Childhood Risk Factors for Early-Onset Drinking*. *J Stud Alcohol Drugs*. 2011;72(5):741–751. https://doi.org/10.15288/jsad.2011.72.741

12. Dube SR, Miller JW, Brown DW, et al. Adverse childhood experiences and the association with ever using alcohol and initiating alcohol use during adolescence. *J Adolesc Health*. 2006;38(4):444.e1–444.e10. https://doi.org/10.1016/j.jadohealth.2005.06.006

13. Kelly Y, Britton A, Cable N, Sacker A, Watt RG. Drunkenness and heavy drinking among 11year olds - Findings from the UK Millennium Cohort Study. *Prev Med*. 2016;90:139–142. https://doi.org/10.1016/j.ypmed.2016.07.010

14. Marshall EJ. Adolescent Alcohol Use: Risks and Consequences. *Alcohol Alcohol*. 2014;49(2):160–164. https://doi.org/10.1093/alcalc/agt180

15. Melotti R, Heron J, Hickman M, Macleod J, Araya R, Lewis G. Adolescent Alcohol and Tobacco Use and Early Socioeconomic Position: The ALSPAC Birth Cohort. *Pediatrics*. 2011;127(4):e948–e955. https://doi.org/10.1542/peds.2009-3450

16. Mersky JP, Topitzes J, Reynolds AJ. Impacts of adverse childhood experiences on health, mental health, and substance use in early adulthood: A cohort study of an urban, minority sample in the U.S. *Child Abuse Negl*. 2013;37(11):917–925. https://doi.org/10.1016/j.chiabu.2013.07.011

17. Paranjothy S, Evans A, Bandyopadhyay A, et al. Risk of emergency hospital admission in children associated with mental disorders and alcohol misuse in the household: an electronic birth cohort study. *Lancet Public Health*. 2018;3(6):e279–e288. https://doi.org/10.1016/S2468-2667(18)30069-0

18. Huber M, Knottnerus JA, Green L, et al. How should we define health? *BMJ*. 2011;343:d4163. https://doi.org/10.1136/bmj.d4163

19. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*. 2004;26(1):99–105. https://doi.org/10.1002/bies.10385

20. Saracci R. Epidemiology in wonderland: Big Data and precision medicine. *Eur J Epidemiol*. 2018;33(3):245–257. https://doi.org/10.1007/s10654-018-0385-9

21. Sedgwick P. Questionnaire surveys: sources of bias. *BMJ*. 2013;347:f5265. https://doi.org/10.1136/bmj.f5265

22. Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of Big Data challenges and analytical methods. *J Bus Res*. 2017;70:263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

23. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). *Int J Epidemiol*. 2014;43(6):1719–1725. https://doi.org/10.1093/ije/dyu001

24. Goodman A, Goodman R. Strengths and Difficulties Questionnaire as a Dimensional Measure of Child Mental Health. *J Am Acad Child Adolesc Psychiatry*. 2009; 48(4):400–403. https://doi.org/10.1097/CHI.0b013e3181985068

25. Hyatt M, Rodgers SE, Paranjothy S, Fone D, Lyons RA. The wales electronic cohort for children (WECC) study. *Arch Dis Child - Fetal Neonatal Ed*. 2011;96(Suppl 1):Fa18–Fa18. https://doi.org/10.1136/archdischild.2011.300164.6

26. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw*. Published online 2010:1–68.

27. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377–399. https://doi.org/10.1002/sim.4067

28. Cantú-Paz E, Newsam S, Kamath C. Feature Selection in Scientific Applications. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. ACM; 2004:788–793. https://doi.org/10.1145/1014052.1016915

29. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media; 2003.

30. Rodgers SE, Demmler JC, Dsilva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. *Health Place*. 2012;18(2):209–217. https://doi.org/10.1016/j.healthplace.2011.09.006

31. Trefan L, Akbari A, Paranjothy S, et al. Electronic Longitudinal Alcohol Study in Communities (ELAStiC) Wales – protocol for platform development. *Int J Popul Data Sci*. 2019;4(1). https://doi.org/10.23889/ijpds.v4i1.581

32. Rose S, van der Laan MJ. Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation. *Int J Biostat*. 2009;5(1). https://doi.org/10.2202/1557-4679.1127

33. Ford DV, Jones KH, Verplancke JP, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009;9:157. https://doi.org/10.1186/1472-6963-9-157

34. Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak*. 2009;9(1):3. https://doi.org/10.1186/1472-6947-9-3

35. R Core Team. R: A Language and Environment for Statistical Computing. Published 2018. Accessed November 22, 2018. https://doi.org/10.1186/1472-6947-9-3

36. Gartner A, Farewell DM, Morgan J, et al. Association between alcohol outlet density and alcohol-related mortality in Wales: an e-cohort study. *The Lancet*. 2017;390:S14. https://doi.org/10.1016/S0140-6736(17)32949-5

37. Moore SC, Orpen B, Smith J, et al. Alcohol affordability: implications for alcohol price policies. A cross-sectional analysis in middle and older adults from UK Biobank. *J Public Health*. 2021;(fdab095). https://doi.org/10.1093/pubmed/fdab095

38. Fone D, Dunstan F, White J, et al. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). *BMC Public Health*. 2012;12(1):428. https://doi.org/10.1186/1471-2458-12-428

39. Brenner H, Gefeller O. Variation of Sensitivity, Specificity, Likelihood Ratios and Predictive Values with Disease Prevalence. *Stat Med*. 1997;16(9):981–991. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N

40. Gray L, Batty GD, Craig P, et al. Cohort Profile: The Scottish Health Surveys Cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *Int J Epidemiol*. 2010;39(2):345–350. https://doi.org/10.1093/ije/dyp155

Abbreviations

ALF: Anonymised linkage field #
ED: Emergency Department
EHR: Electronic health record
LSOA: Lower super output area
MCS: Millennium Cohort Study
NWIS: National Health Service Wales Informatics Service
PEDW: Patient Episode Database for Wales
PMM: predictive mean matching
RALF: Residential anonymised linking field
SAIL: Secure Anonymised Information Linkage
SDQ: Strength and Difficulty Questionnaire
SEC: socio-economic classification
SQL: Structured Query Language
WDSD: Welsh Demographic Service Dataset
WECC: Wales Electronic Cohort for Children
WLGP: Welsh Longitudinal General Practice
WIMD: Welsh Index of Multiple Deprivation

Supplementary Appendices

Supplementary table 1: MCS alcohol-related questions and criteria for inclusion in the case group

Questions	Criteria
How many times have you had an alcoholic drink in the last 12 months?	3–5 times or more
How many times have you had an alcoholic drink in the last four weeks?	1–2 times or more
Have you ever drunk enough to feel drunk?	Yes
Have you ever had five or more alcoholic drinks at a time? A drink is half a pint of lager, beer or cider, one alcopop, a small glass of wine, or a measure of spirits.	Yes
How many times have you had five or more alcoholic drinks at a time?	Once or more

Supplementary table 2: MCS to Whole Population explanatory variables mapping

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Gender	Gender	WDSD	No	1 = male 0 = female	
Lone parent	Living with single adult	WDSD	Yes		
Additional household member	Living with single adult	WDSD	Yes	0 = Never with a single adult 1 = Always with a single adult 2 = Ever been with single adult	1. Using RALF, number of people sharing same house with child at the above mentioned 4 time points were derived 2. Based on household members' age at the 4 time points, the number of adults staying with child was determined 3. A binary variable was created based on the number adults at the household at 4 time points 4. A categorical summary variable was created to identify the overall status of the concept variable
Mother's SEC	Employment deprivation	WIMD reference data from Welsh Government	Yes	0 = Always in least deprived group 1 = Always in most deprived group 2 = Ever belong to most deprived group	1. Welsh Index of Multiple Deprivation (WIMD) quintile scale on employment and overall deprivation at each time point for each RALF was achieved. 2. WIMD quintile scale between 1 and 5 (from most to least deprivation). 3. The study combined the scale 1 and 2 to indicate the most deprived group and the rest 3 scales were classified as least deprived group 4. A categorical summary variable was created to identify the overall status of the concept variable

(Continued).

Supplementary table 2: Continued

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Mother alcohol use during pregnancy	Mother's alcohol-related condition during pregnancy	WECC, WLGP, PEDW	No	1 = Yes 0 = No	1. From WECC the maternal ALF was obtained 2. Based on gestational age and the week of birth the pregnancy period was calculated 3. If the mother had an alcohol-related code recorded in WLGP or PEDW during the pregnancy period then a binary flag variable was created
Guardian alcohol use	Household member with alcohol-related hospital admission record	WDSD, PEDW	Yes	0 = Never lived with someone who had an alcohol hospital admission 1 = Ever lived with someone who had an alcohol hospital admission	1. Using RALF, any household member had an alcohol-related event recorded in WLGP or PEDW between birth to < nine months, nine months to < three years, three years to < five years and five years to < seven years -was identified 2. A categorical summary variable was created
Living area	Living area	WDSD and Rural Urban indicator reference data from Welsh Government	Yes	0 = Always lived in rural area 1 = Always lived in urban area 2 = Ever lived in urban area	1. Each RALF is always within a Lower super Output Area (LSOA) code. 2. Each LSOA code is further categorised using the rural urban indicators into urban, village and town. 3. In this study village and town are grouped together and classified as rural. 4. A categorical summary variable was created
Maternal age at birth	Maternal age at birth	WECC	No	Less than 20 years 20 to 24 years 25 to 29 years 30 to 34 years 35 years or over	
Gestational age	Gestational age	WECC	No	1 = not term 0 = term	

(Continued).

Supplementary table 2: Continued

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Mother longstanding illness	Mother's any comorbidity Mother's psychosis disorder	WLGP, PEDW	No	1 = yes	Any longstanding health condition, common mental health condition and psychosis disorder between their birth and the seven years of their child's age
	Mother's common mental health condition			0 = no	
Conduct disorder	Conduct disorder (CD)	WLGP	No	1 = yes	CD diagnosis/treatment by GP between birth and age seven
				0 = no	
Hyperactivity	Attention Deficit Hyperactivity disorder (ADHD)	WLGP	No	1 = yes	ADHD diagnosis/treatment by GP between birth and age seven
				0 = no	
Emotional difficulty Total difficulty score	Other mental health condition	WLGP, PEDW	No	1 = yes	Any mental health condition (apart from ADHD and CD codes) reported in GP
				0 = no	Any mental health condition related hospital admission between birth and age seven
Health codes: 5 Read codes and 1 ICD10 codes	Health codes: 5 Read codes and 1 ICD10 codes	Read codes from WLGP and ICD10 codes from PEDW	No		Individual code recorded in WLGP and PEDW between birth and age 7

Supplementary table 3: Alcohol-related ICD10 codes

ICD10 Code	Description
E244	Alcohol-induced pseudo-Cushing's syndrome
E512	Wernicke's encephalopathy
F100	Mental and behavioural disorders due to use of alcohol
F101	Mental and behavioural disorders due to use of alcohol
F102	Mental and behavioural disorders due to use of alcohol
F103	Mental and behavioural disorders due to use of alcohol
F104	Mental and behavioural disorders due to use of alcohol
F105	Mental and behavioural disorders due to use of alcohol
F106	Mental and behavioural disorders due to use of alcohol
F107	Mental and behavioural disorders due to use of alcohol
F108	Mental and behavioural disorders due to use of alcohol
F109	Mental and behavioural disorders due to use of alcohol
G312	Degeneration of nervous system due to alcohol
G405	Special epileptic syndromes
G621	Alcoholic polyneuropathy
G721	Alcoholic myopathy
I426	Alcoholic cardiomyopathy
K292	Alcoholic gastritis
K700	Alcoholic fatty liver
K701	Alcoholic hepatitis
K702	Alcoholic fibrosis and sclerosis of liver
K703	Alcoholic cirrhosis of liver
K704	Alcoholic hepatic failure
K709	Alcoholic liver disease, unspecified
K852	Alcohol-induced acute pancreatitis
K860	Alcohol-induced chronic pancreatitis
O354	Maternal care for (suspected) damage to fetus from alcohol
Q860	Fetal alcohol syndrome (dysmorphic)
R780	Finding of alcohol in blood
T510	Toxic effect: Ethanol
X450–X459	Accidental poisoning by and exposure to alcohol
X650–X659	Intentional self-poisoning by and exposure to alcohol
Y150	Poisoning by and exposure to alcohol, undetermined intent
Y152	Poisoning by and exposure to alcohol, undetermined intent
Y154	Poisoning by and exposure to alcohol, undetermined intent
Y158	Poisoning by and exposure to alcohol, undetermined intent
Y159	Poisoning by and exposure to alcohol, undetermined intent
Y900	Blood alcohol level of less than 20 mg/100 ml
Y901	Blood alcohol level of 20–39 mg/100 ml
Y902	Blood alcohol level of 40–59 mg/100 ml
Y903	Blood alcohol level of 60–79 mg/100 ml
Y904	Blood alcohol level of 80–99 mg/100 ml
Y905	Blood alcohol level of 100–119 mg/100 ml
Y906	Blood alcohol level of 120–199 mg/100 ml
Y907	Blood alcohol level of 200–239 mg/100 ml
Y908	Blood alcohol level of 240 mg/100 ml or more
Y909	Presence of alcohol in blood, level not specified
Y910	Mild alcohol intoxication
Y911	Moderate alcohol intoxication
Y912	Severe alcohol intoxication
Y913	Very severe alcohol intoxication
Y919	Alcohol involvement, not otherwise specified
Z502	Alcohol rehabilitation
Z714	Alcohol abuse counselling and surveillance
Z721	Alcohol use

Supplementary table 4: Alcohol-related read codes

Read Code	Description
136..	Alcohol consumption
1362	Trivial drinker – <1 u/day
1363	Light drinker – 1–2 u/day
1364	Moderate drinker – 3–6 u/day
1365	Heavy drinker – 7–9 u/day
1366	Very heavy drinker – >9 u/day
1368	Alcohol consumption unknown
1369	Suspect alcohol abuse – denied
136F.	Spirit drinker
136G.	Beer drinker
136H.	Drinks beer and spirits
136I.	Drinks wine
136J.	Social drinker
136K.	Alcohol intake above recommended sensible limits
136L.	Alcohol intake within recommended sensible limits
136N.	Light drinker
136O.	Moderate drinker
136P.	Heavy drinker
136Q.	Very heavy drinker
136R.	Binge drinker
136S.	Hazardous alcohol use
136T.	Harmful alcohol use
136V.	Alcohol units per week
136W.	Alcohol misuse
136X.	Alcohol units consumed on heaviest drinking day
136Y.	Drinks in morning to get rid of hangover
136Z.	Alcohol consumption NOS
136a.	Increasing risk drinking
136b.	Feels should cut down drinking
136c.	Higher risk drinking
136d.	Lower risk drinking
136e.	Declines to state current alcohol consumption
13Y8.	Alcoholics anonymous
13ZY.	Disqualified from driving due to excess alcohol
1462	H/O: alcoholism
1B1c.	Alcohol induced hallucinations
1F9D.	Replaces meals with drinks
2126C	Alcohol dependence resolved
2577	O/E – breath – alcohol smell
388u.	Fast alcohol screening test
38D2.	Single alcohol screening questionnaire
38D3.	Alcohol use disorders identification test
38D4.	Alcohol use disorder identification test consumption questionnaire
38D5.	Alcohol use disorder identification test Piccinelli consumption questionnaire
38Df.	Five-shot questionnaire on heavy drinking
38Dz.	Severity of alcohol dependence questionnaire
38P03	Health of the Nation Outcome Scale for Children and Adolescents item 4 – alcohol, substance/solvent misuse
38QA.	CIWA-Ar - Clinical Institute Withdrawal Assessment for Alcohol scale, revised
38QE.	Addiction Research Foundation Clinical Institute Withdrawal Assessment for Alcohol
44X3.	Blood ethanol level
66e..	Alcohol disorder monitoring
66e0.	Alcohol abuse monitoring
6792	Health ed. – alcohol
67A5.	Pregnancy alcohol advice
67H0.	Lifestyle advice regarding alcohol

(Continued).

Supplementary table 4: Continued

Read Code	Description
67K6.	Cycle of change stage, alcohol
6892	Alcohol consumption screen
68S..	Alcohol consumption screen
7P221	Delivery of rehabilitation for alcohol addiction
8BA8.	Alcohol detoxification
8BAs.	Alcohol relapse prevention
8BAu.	Alcohol harm reduction programme
8CAM.	Patient advised about alcohol
8CAM0	Advised to abstain from alcohol consumption
8CAv.	Advised to contact primary care alcohol worker
8CE1.	Alcohol leaflet given
8CdK.	Specialist alcohol treatment service signposted
8G32.	Aversion therapy – alcoholism
8H35.	Admitted to alcohol detoxification centre
8H7p.	Referral to community alcohol team
8HHe.	Referral to community drug and alcohol team
8HkG.	Referral to specialist alcohol treatment service
8HkJ.	Referral to alcohol brief intervention service
8IA7.	Alcohol consumption screening test declined
8IAF.	Brief intervention for excessive alcohol consumption declined
8IAJ.	Declined referral to specialist alcohol treatment service
8IAt.	Extended intervention for excessive alcohol consumption declined
8IEA.	Referral to community alcohol team declined
8IH4.	Alcohol Use Disorders Identification Test declined
8W2..	Referral to mental health services deferred until alcohol misuse resolved
9EQ..	HO/RTS-police:venesect alc
9EVD.	Hospital alcohol liaison team report received
9NJz.	In-house alcohol detoxification
9NN2.	Under care of community alcohol team
9NgzH	Withdrawn from alcohol detoxification programme
9NzA.	Hospital attendance related to personal alcohol consumption
9k1..	Alcohol misuse – enhanced services administration
9k11.	Alcohol consumption counselling
9k12.	Alcohol misuse – enhanced service completed
9k13.	Alcohol questionnaire completed
9k14.	Alcohol counselling by other agencies
9k15.	Alcohol screen – alcohol use disorder identification test completed
9k16.	Alcohol screen – fast alcohol screening test completed
9k17.	Alcohol screen – alcohol use disorder identification test consumption questions completed
9k18.	Alcohol screen – alcohol use disorder identification test Piccinelli consumption questions completed
9k19.	Alcohol assessment declined – enhanced services administration
9k1A.	Brief intervention for excessive alcohol consumption completed
9k1B.	Extended intervention for excessive alcohol consumption completed
C1505	Alcohol-induced pseudo-Cushing's syndrome
E01..	Alcoholic psychoses
E010.	Alcohol withdrawal delirium
E011.	Alcohol amnestic syndrome
E0110	Korsakov's alcoholic psychosis
E0111	Korsakov's alcoholic psychosis with peripheral neuritis
E011z	Alcohol amnestic syndrome NOS
E012.	Other alcoholic dementia
E0120	Chronic alcoholic brain syndrome
E013.	Alcohol withdrawal hallucinosis
E014.	Pathological alcohol intoxication
E015.	Alcoholic paranoia

(Continued).

Supplementary table 4: Continued

Read Code	Description
E01y.	Other alcoholic psychosis
E01y0	Alcohol withdrawal syndrome
E01yz	Other alcoholic psychosis NOS
E01z.	Alcoholic psychosis NOS
E23..	Alcohol dependence syndrome
E230.	Acute alcoholic intoxication in alcoholism
E2300	Acute alcoholic intoxication, unspecified, in alcoholism
E2301	Continuous acute alcoholic intoxication in alcoholism
E2302	Episodic acute alcoholic intoxication in alcoholism
E2303	Acute alcoholic intoxication in remission, in alcoholism
E230z	Acute alcoholic intoxication in alcoholism NOS
E231.	Chronic alcoholism
E2310	Unspecified chronic alcoholism
E2311	Continuous chronic alcoholism
E2312	Episodic chronic alcoholism
E2313	Chronic alcoholism in remission
E231z	Chronic alcoholism NOS
E23z.	Alcohol dependence syndrome NOS
E250.	Nondependent alcohol abuse
E2500	Nondependent alcohol abuse, unspecified
E2501	Nondependent alcohol abuse, continuous
E2502	Nondependent alcohol abuse, episodic
E2503	Nondependent alcohol abuse in remission
E250z	Nondependent alcohol abuse NOS
Eu10.	[X]Mental and behavioural disorders due to use of alcohol
Eu100	[X]Mental and behavioural disorders due to use of alcohol: acute intoxication
Eu101	[X]Mental and behavioural disorders due to use of alcohol: harmful use
Eu102	[X]Mental and behavioural disorders due to use of alcohol: dependence syndrome
Eu103	[X]Mental and behavioural disorders due to use of alcohol: withdrawal state
Eu104	[X]Mental and behavioural disorders due to use of alcohol: withdrawal state with delirium
Eu105	[X]Mental and behavioural disorders due to use of alcohol: psychotic disorder
Eu106	[X]Mental and behavioural disorders due to use of alcohol: amnesic syndrome
Eu107	[X]Mental and behavioural disorders due to use of alcohol: residual and late-onset psychotic disorder
Eu108	[X]Alcohol withdrawal-induced seizure
Eu10y	[X]Mental and behavioural disorders due to use of alcohol: other mental and behavioural disorders
Eu10z	[X]Mental and behavioural disorders due to use of alcohol: unspecified mental and behavioural disorder
F11x0	Cerebral degeneration due to alcoholism
F1440	Cerebellar ataxia due to alcoholism
F25B.	Alcohol-induced epilepsy
F375.	Alcoholic polyneuropathy
F3941	Alcoholic myopathy
G555.	Alcoholic cardiomyopathy
G8523	Oesophageal varices in alcoholic cirrhosis of the liver
J153.	Alcoholic gastritis
J610.	Alcoholic fatty liver
J611.	Acute alcoholic hepatitis
J612.	Alcoholic cirrhosis of liver
J6120	Alcoholic fibrosis and sclerosis of liver
J613.	Alcoholic liver damage unspecified
J6130	Alcoholic hepatic failure
J617.	Alcoholic hepatitis
J6170	Chronic alcoholic hepatitis
J6708	Alcohol-induced acute pancreatitis
J6710	Alcohol-induced chronic pancreatitis

(Continued).

Supplementary table 4: Continued

Read Code	Description
L2553	Maternal care for (suspected) damage to fetus from alcohol
PK80.	Fetal alcohol syndrome
PK83.	Fetus and newborn affected by maternal use of alcohol
Q0071	Fetus or neonate affected by placental or breast transfer of alcohol
R103.	[D]Alcohol blood level excessive
SLH3.	Alcohol deterrent poisoning
SM0..	Alcohol causing toxic effect
SM00.	Ethyl alcohol causing toxic effect
SM000	Ethanol causing toxic effect
SM002	Grain alcohol causing toxic effect
SM00z	Ethyl alcohol causing toxic effect NOS
SM0z.	Alcohol causing toxic effect NOS
T90..	Accidental poisoning by alcohol, NEC
T900.	Accidental poisoning by alcoholic beverages
T901.	Accidental poisoning by other ethyl alcohol and its products
T9012	Accidental poisoning by grain alcohol NOS
T901z	Accidental poisoning by ethyl alcohol NOS
T90z.	Accidental poisoning by alcohol NOS
TJH3.	Adverse reaction to alcohol deterrents
U1A9.	[X]Accidental poisoning by and exposure to alcohol
U1A90	[X]Accidental poisoning by and exposure to alcohol, occurrence at home
U1A91	[X]Accidental poisoning by and exposure to alcohol, occurrence in residential institution
U1A92	[X]Accidental poisoning by and exposure to alcohol, occurrence at school, other institution and public administrative area
U1A93	[X]Accidental poisoning by and exposure to alcohol, occurrence at sports and athletics area
U1A94	[X]Accidental poisoning by and exposure to alcohol, occurrence on street and highway
U1A95	[X]Accidental poisoning by and exposure to alcohol, occurrence at trade and service area
U1A96	[X]Accidental poisoning by and exposure to alcohol, occurrence at industrial and construction area
U1A97	[X]Accidental poisoning by and exposure to alcohol, occurrence on farm
U1A9y	[X]Accidental poisoning by and exposure to alcohol, occurrence at other specified place
U1A9z	[X]Accidental poisoning by and exposure to alcohol, occurrence at unspecified place
U209.	[X]Intentional self poisoning by and exposure to alcohol
U2090	[X]Intentional self poisoning by and exposure to alcohol, occurrence at home
U2091	[X]Intentional self poisoning by and exposure to alcohol, occurrence in residential institution
U2092	[X]Intentional self poisoning by and exposure to alcohol, occurrence at school, other institution and public administrative area
U2093	[X]Intentional self poisoning by and exposure to alcohol, occurrence at sports and athletics area
U2094	[X]Intentional self poisoning by and exposure to alcohol, occurrence on street and highway
U2095	[X]Intentional self poisoning by and exposure to alcohol, occurrence at trade and service area
U2096	[X]Intentional self poisoning by and exposure to alcohol, occurrence at industrial and construction area
U2097	[X]Intentional self poisoning by and exposure to alcohol, occurrence on farm
U209y	[X]Intentional self poisoning by and exposure to alcohol, occurrence at other specified place
U209z	[X]Intentional self poisoning by and exposure to alcohol, occurrence at unspecified place
U4097	[X]Poisoning by and exposure to alcohol, occurrence on farm, undetermined intent
U60H3	[X]Alcohol deterrents causing adverse effects in therapeutic use
U8...	[X]Supplementary factors related to causes of morbidity and mortality classified elsewhere
U80..	[X]Evidence of alcohol involvement determined by blood alcohol level
U800.	[X]Evidence of alcohol involvement determined by blood alcohol level of less than 20 mg/100 ml
U801.	[X]Evidence of alcohol involvement determined by blood alcohol level of 20–39 mg/100 ml
U802.	[X]Evidence of alcohol involvement determined by blood alcohol level of 40–59 mg/100 ml
U803.	[X]Evidence of alcohol involvement determined by blood alcohol level of 60–79 mg/100 ml
U804.	[X]Evidence of alcohol involvement determined by blood alcohol level of 80–99vmg/100 ml
U805.	[X]Evidence of alcohol involvement determined by blood alcohol level of 100–119 mg/100 ml
U806.	[X]Evidence of alcohol involvement determined by blood alcohol level of 120–199 mg/100 ml

(Continued).

Supplementary table 4: Continued

Read Code	Description
U807.	[X]Evidence of alcohol involvement determined by blood alcohol level of 200–239 mg/100 ml
U808.	[X]Evidence of alcohol involvement determined by blood alcohol level of 240 mg/100 ml or more
U80z.	[X]Evidence of alcohol involvement determined by presence of alcohol in blood, level not specified
U81..	[X]Evidence of alcohol involvement determined by level of intoxication
U810.	[X]Evidence of alcohol involvement determined by level of intoxication, mild alcohol intoxication
U811.	[X]Evidence of alcohol involvement determined by level of intoxication, moderate alcohol intoxication
U812.	[X]Evidence of alcohol involvement determined by level of intoxication, severe alcohol intoxication
U813.	[X]Evidence of alcohol involvement determined by level of intoxication, very severe alcohol intoxication
U814.	[X]Evidence of alcohol involvement determined by level of intoxication, alcohol involvement, not otherwise specified
ZV113	[V]Personal history of alcoholism
ZV4KC	[V] Alcohol use
ZV57A	[V]Alcohol rehabilitation
ZV6D6	[V]Alcohol abuse counselling and surveillance
ZV704	[V]Medicolegal examination
ZV70L	[V]Blood-alcohol and blood-drug test
ZV791	[V]Screening for alcoholism
du11.	DISULFIRAM 200 mg tablets
du12.	ANTABUSE 200 mg tablets

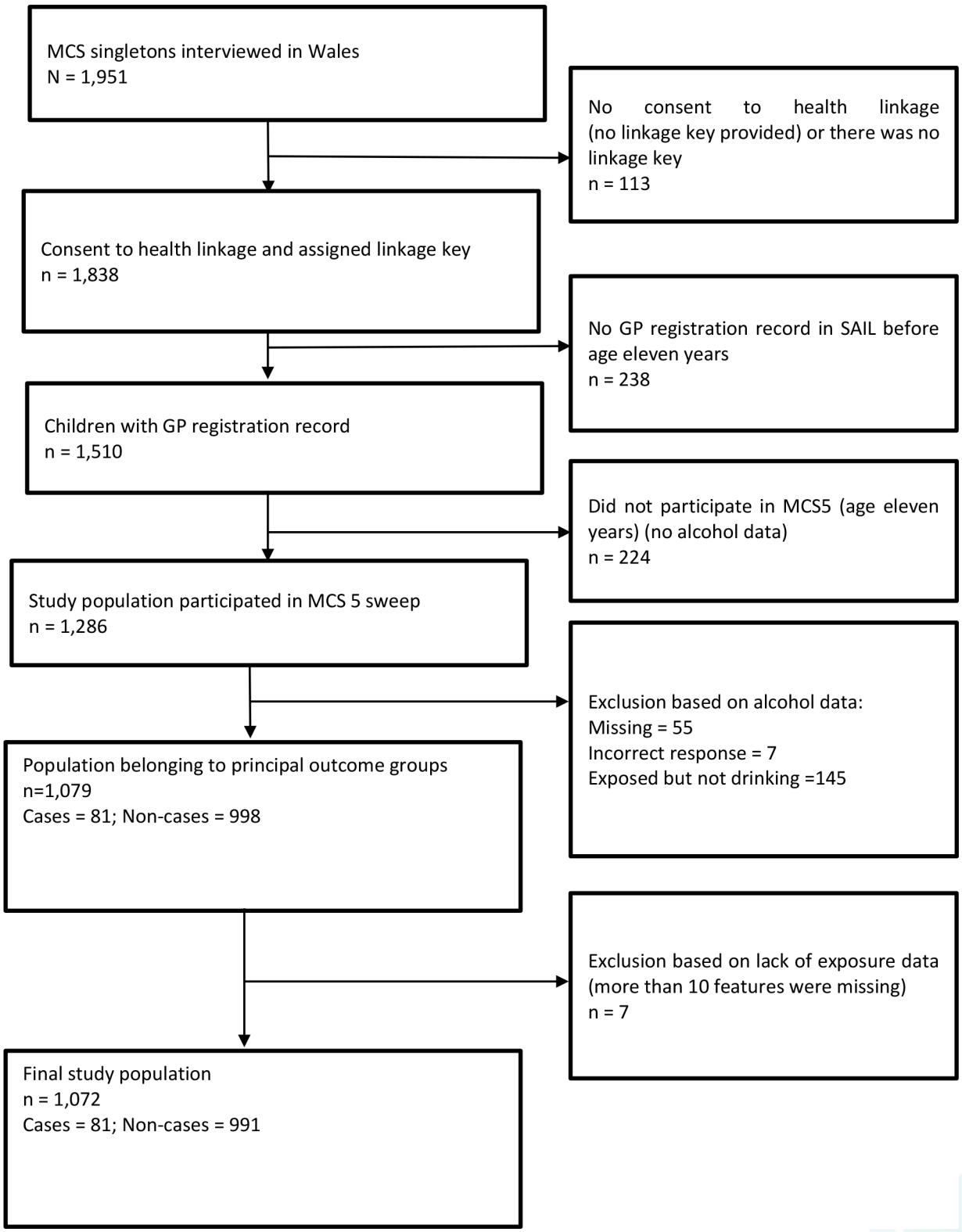
Supplementary table 5: The contingency table for the whole population analysis

	Actual negative	Actual positive	Total
Predicted negative	2,182 (true negative [TN])	73 (false negative [FN])	2,255
Predicted positive	1,367 (false positive [FP])	101 (true positive [TP])	1,468
Total	3,549	174	3,723

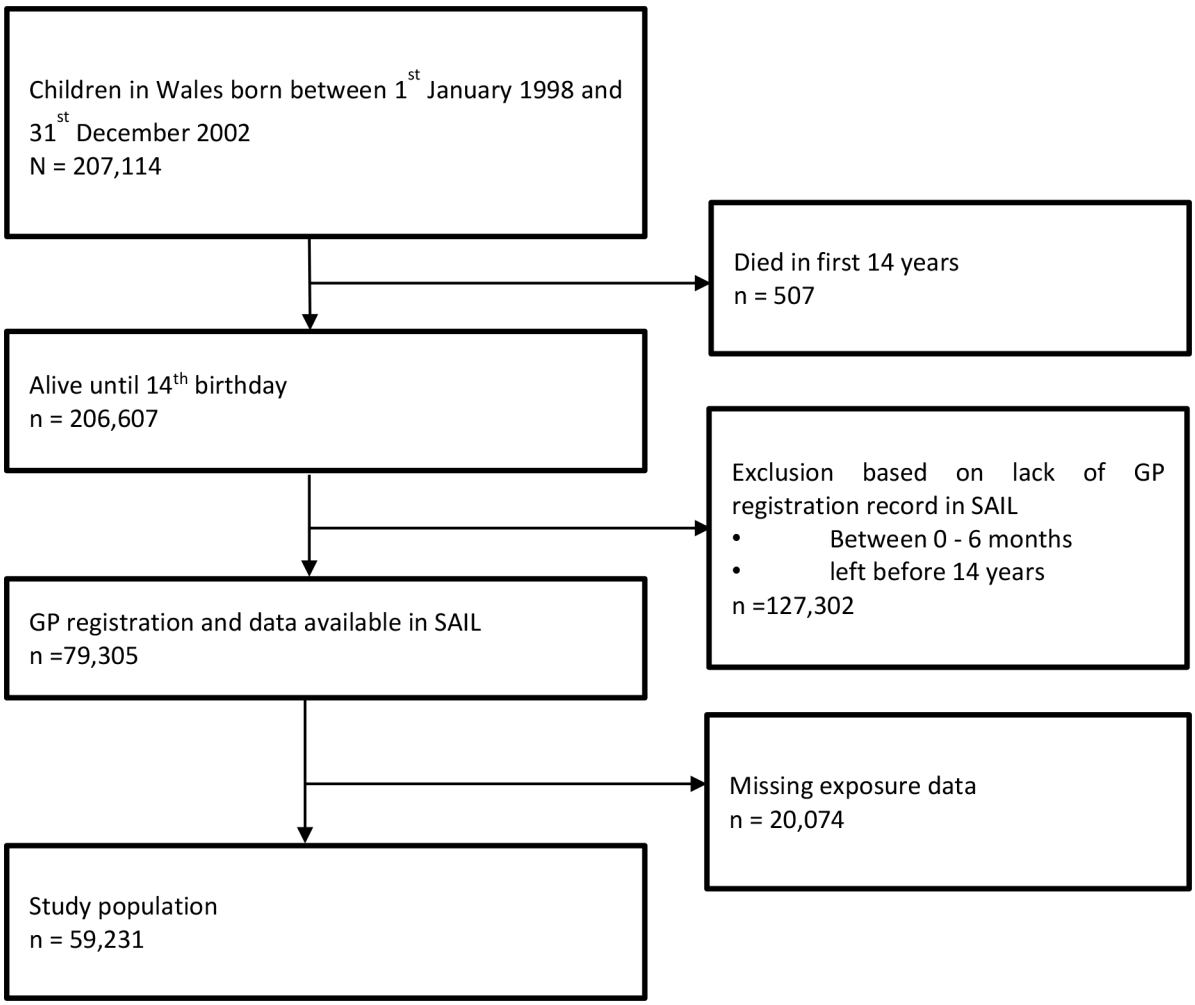
Supplementary table 6: Model prediction results

Measurement	Formula	Value
Accuracy	TP + TN / TP + TN + FP + FN	61.32
Sensitivity	TP / TP + FN	58.05
Specificity	TN / TN + FP	61.48
Positive predictive value	TP / TP + FP	6.88
Negative predictive value	TN / TN + FN	96.76

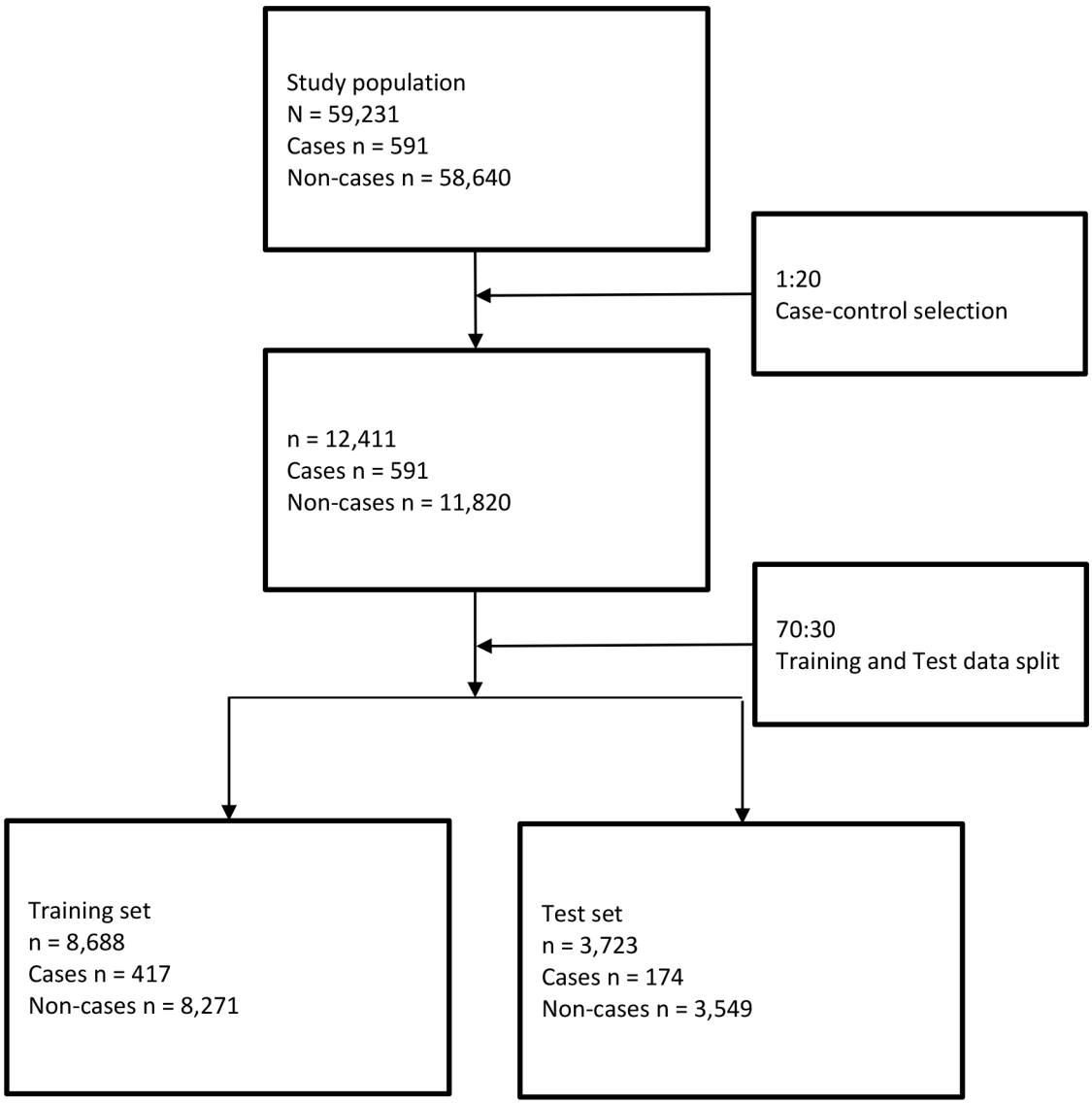
Supplementary Figure 1: Flow diagram of the MCS participants



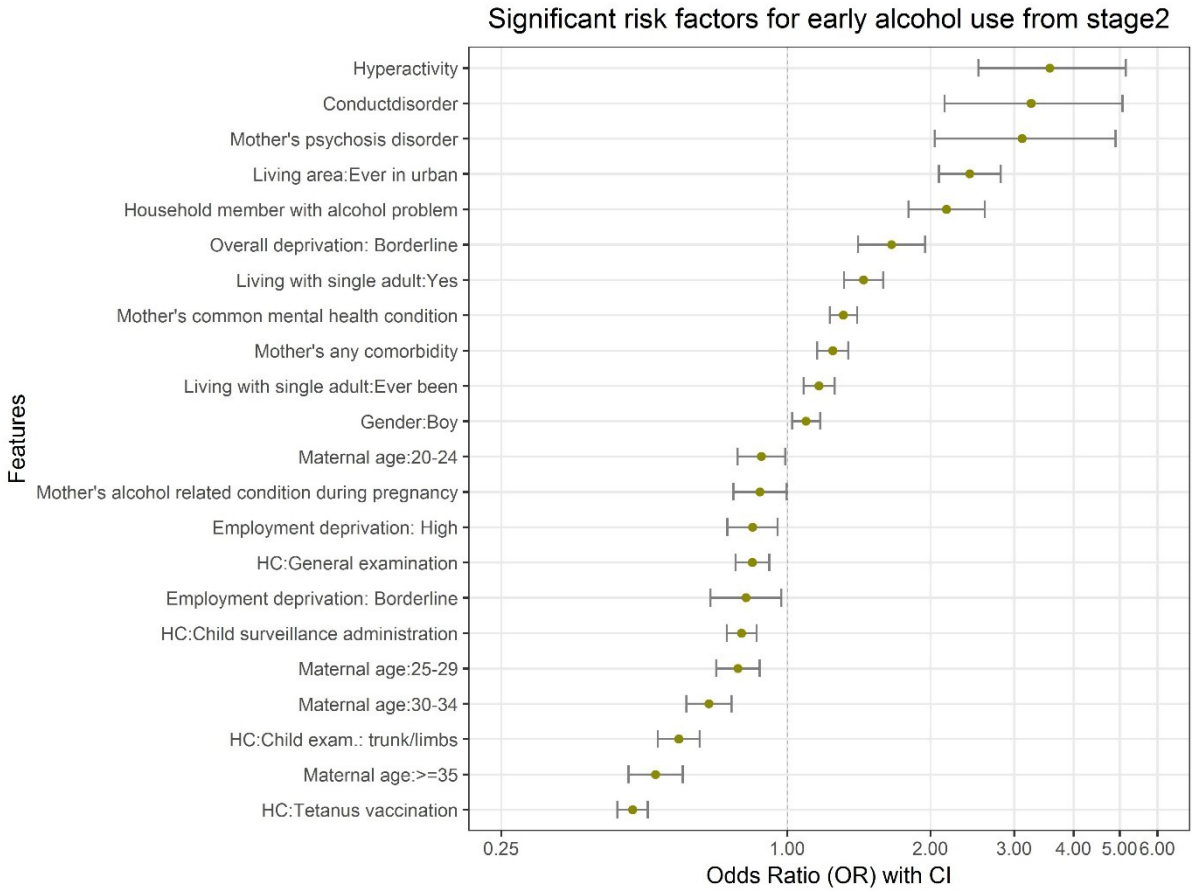
Supplementary Figure 2: Flow diagram of the whole population participants



Supplementary Figure 3: Flow diagram for the final study population



Supplementary Figure 4: Significant risk factors associated with higher and lower risk of early alcohol-related health outcomes from whole population analysis (stage 2)



HC: Health code from EHRs