Characterisation of chronic hepatitis B virus infection in the UK and risk factors for hepatocellular carcinoma: a large electronic health record-based retrospective cohort study in the QResearch primary care database.

Cori Campbell[1,2], Tingyan Wang[1,2], Iain Gillespie[3],
Eleanor Barnes[2,4], Philippa C Matthews[2,5,6,7]

[1] Nuffield Department of Medicine, University of Oxford, Oxford, UK
[2] NIHR Oxford Biomedical Research Centre, Oxford, UK
[3] GSK, Stevenage, UK
[4] Department of Hepatology, Oxford University Hospitals, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU
[5] The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK
[6] Division of Infection and Immunity, University College London, Gower Street, London, WC1E 6BT, UK
[7] Department of Infectious Diseases, University College London Hospital, Euston Road, London NW1 2BU, UK

Corresponding author: philippa.matthews@crick.ac.uk

Word count: 3929

Key words: hepatitis B virus, epidemiology, liver cancer, statins, electronic health records, big data, viral hepatitis elimination

Abstract

Background: Chronic hepatitis B infection (CHB) is the leading global cause of hepatocellular carcinoma (HCC), however the CHB population in the UK is not well characterised, and few investigations of risk factors for progression to HCC in multiethnic populations have been undertaken.

Methods: We identified CHB individuals from the English primary care database QResearch, containing >35 million patient records. Individuals ≥18 years with a record of CHB between 1999-2019 were included. Risk factors for HCC were investigated using Cox proportional hazards modelling.

Results: We identified 8039 individuals with CHB in whom median follow-up was 3.87 years (interquartile range (IQR) 6.30 years). Most individuals (60.4%) were males of non-White ethnicities (>70%), and 44% were in the fifth (most deprived) Townsend deprivation quintile. In multivariable Cox proportional hazards models (undertaken in a subset of 7029 individuals with longitudinal data), 161 HCC cases occurred. Increased hazards of HCC associated with male sex (adjusted hazards ratio (aHR) 3.44, 95% Confidence Interval (95CI) 2.07-5.73), older age (for age groups 56-55 and ≥66 years of age, compared to 26-35 years, aHRs 7.52 (95CI 4.14-13.67) and 11.89 (95CI 6.26-22.60) respectively), socioeconomic deprivation (aHR for fifth Townsend deprivation quintile 1.69, 95CI 1.01-2.84, compared to third), Caribbean ethnicity (aHR 3.32, 95CI 1.43-7.71, compared to White ethnicity), ascites (aHR 1.85, 95CI 1.02-3.36), cirrhosis (aHR 6.52, 95CI 4.54-9.37) and peptic ulcer disease (aHR 2.20, 95CI 1.39-3.49). Reduced HCC hazards associated with statin use (aHR 0.47, 95CI 0.22-0.99).

Conclusions: We identified novel risk factors for HCC in an ethnically diverse and socioeconomically deprived population of adults living with CHB.

Introduction

The 2022 Global Burden of Disease systematic analysis for hepatitis B virus (HBV) estimates that >300 million people live with chronic infection worldwide (1). Through its progression to primary liver cancer (hepatocellular carcinoma (HCC)), cirrhosis and other late-stage disease, chronic hepatitis B (CHB) accounts for substantial morbidity and mortality, being the leading global cause of HCC death (2), and is third largest cause of cirrhosis death (1). Age-standardised death rates have remained constant or increased for HCC and cirrhosis, respectively, and the total number of worldwide HBV-attributable deaths has increased over recent decades (3). International targets calling for the elimination of HBV infection as a public health threat by the year 2030 have been set (4), with recent investment into the early detection and treatment of HCC (5). Meeting elimination targets relies on a clear understanding of the epidemiology of infection and associated liver disease in order to target resources and interventions to high-risk groups, and to benchmark progress.

CHB prevalence and incidence have not been robustly estimated in many settings, including the U.K. (6), and groups at the highest risk of morbidity and mortality have not been well characterised. Furthermore, even in well-defined CHB populations, treatment coverage and eligibility are often unreported. Regional HBV reports from U.K. public health services (previously Public Health England) have included neither

overall estimates of the proportion of CHB individuals receiving antiviral treatment, nor estimates stratified by relevant subgroups such as age, sex and ethnicity (6–10).

There has been increasing interest in identifying risk factors for progression of CHB to cirrhosis, HCC and other endpoints (11). Age, sex, viral load and viral genotype are recognised determinants of HCC risk (12–18), and recent studies have reported associations between HCC and various comorbidities, including type 2 diabetes mellitus (T2DM) and hypertension (11). Existing risk scores, including PAGE-B (19), REACH-B (20,21), GAG-HCC (22,23) and CU-HCC (24–26), use such risk factors to predict HCC. However, few cohorts have been characterised in European countries and/or in ethnically diverse populations, to validate or inform scoring approaches.

Studies based on electronic health records (EHRs) enable characterisation of large retrospective cohorts, thus enhancing statistical power, and identifying a study sample that is more representative of the disease population compared to clinical trials. Such databases often have longitudinal follow-up, with exposures and outcomes ascertained over time. EHR databases can often be linked to other registries (such as national cancer registries and vital statistics), allowing for identification of relevant endpoints.

Given the substantial evidence gaps concerning HBV epidemiology, disease burden and knowledge of risk factors for progression to end-stage disease, we set out to i) characterise a CHB sample in the QResearch primary care database (27), collating data for individuals with CHB from a large-scale primary care database in England, and investigating risk factors for progression of CHB to HCC.

There has been increasing interest in identifying risk factors for progression of CHB to cirrhosis, HCC and other endpoints (11). Age, sex, viral load and viral genotype are recognised determinants of HCC risk (12–18), and recent studies have reported associations between HCC and various comorbidities, including type 2 diabetes mellitus (T2DM) and hypertension (11). Existing risk scores, including PAGE-B (19), REACH-B (20,21), GAG-HCC (22,23) and CU-HCC (24–26), use such risk factors to predict HCC. However, few cohorts have been characterised in European countries and/or in ethnically diverse populations, to validate or inform scoring approaches.

Studies based on electronic health records (EHRs) enable characterisation of large retrospective cohorts, thus enhancing statistical power, and identifying a study sample that is more representative of the population than clinical trial samples. Such databases often have longitudinal follow-up, with exposures and outcomes ascertained over time. EHR databases can often be linked to other registries (such as national cancer registries and vital statistics), allowing for identification of relevant endpoints.

Given the substantial evidence gaps concerning HBV epidemiology, disease burden and knowledge of risk factors for progression to end-stage disease, we set out to characterise a CHB cohort collating data for individuals with CHB from a large-scale primary care database in England (27), and to investigating risk factors for progression of CHB to HCC.

Methods

Data source and study population/design

We utilised data from the UK primary care database QResearch (version 45), which contains >35 million patient records from >1800 individual practices (28). QResearch was established in 2002 and contains anonymised individual-level patient EHRs. Data are collected prospectively and are linked to hospital episode statistics (HES), National Cancer Registration Analysis Service (NCRAS) and Office for National Statistics (ONS) mortality data.

We first identified individuals from the QResearch database between 01 January 1999 and 31 December 2019 who were age ≥18 years and had a record of CHB based on a diagnostic Systemised Nomenclature of Medicine (SNOMED)/Read or International Classification of Disease (ICD) code, or who had a history of ≥1 hepatitis B surface antigen (HBsAg) or HBV DNA viral load (VL) measurement. From this sample we then selected patients for inclusion in our cohort who satisfied the following inclusion criteria anytime between 01 January 1999 and 31 December 2019: i) age ≥18 years, and; ii) presence of CHB. Patients were identified as having CHB if they met any one of the following conditions: i) record of a diagnostic SNOMED/Read code indicating CHB; ii) record of a diagnostic ICD-9 or -10 code indicating CHB, and/or; iii) Presence of HBsAg or VL on at least two recordings ≥6 months apart (**Supplementary Figure 1)**.

Covariate selection and ascertainment

We identified relevant covariates for extraction *a priori* based on previous literature (11,19,21–26,29) and clinical relevance. Lifestyle factors, demographics and relevant numeric biomarkers were characterised from relevant SNOMED/Read codes. Comorbidities were characterised from relevant SNOMED/Read and ICD-9 and -10 codes. Ethnicity is categorised in QResearch as per 2011 census categories (30). Townsend Deprivation quintile is available as a measure of socioeconomic status in QResearch. Body mass index (BMI, kg/m^2) was categorised (underweight, <18.5 kg/m^2; healthy weight, 18.5-2.49 kg/m^2; overweight, 25.0-29.9 kg/m^2; obese, ≥ 30 kg/m^2) based on World Health Organization (WHO) categories (31). Covariate measurements made within ±3 years of earliest CHB diagnosis and before HCC diagnosis were used as proxy baseline measurements. Where patients had > 1 measurement(s) taken within 3 years of the earliest CHB diagnosis, measurements taken closest to diagnosis date were used.

Outcome ascertainment

Our primary endpoint of interest was HCC. HCC was ascertained via identification of patients with relevant SNOMED/Read or ICD codes corresponding to HCC, and by linkage of the cohort to National Cancer Registry data (32,33). In order to maximise outcome ascertainment, we used a broad definition for HCC including multiple codes relevant to HCC (**Supplementary Table 1**). We performed sensitivity analysis (further details below) whereby all patients with non-HCC neoplasms were excluded, to investigate robustness of main analysis using our broad HCC definition. A tabulation of HCC cases across source of diagnosis is depicted in **Supplementary Table 2**.

Follow-up

Earliest date of CHB diagnosis was regarded as cohort entry and initiation of follow-up for each patient. For patients who experienced HCC, date of HCC diagnosis was regarded as the end of follow-up. For patients who did not develop HCC (i.e., patients who were censored), follow-up ended at patient cohort exit date (either due to leaving their general practice and switching to a practice which does not contribute to QResearch, or death) or 31 December 2019, whichever occurred earlier.

In some patients, HCC diagnosis date or cohort exit date preceded or was equal to CHB diagnosis date (**Supplementary Table 3**). Data from these patients were excluded from analyses of HCC risk factors.

Statistical analysis

Statistical analyses were carried out in R (version 4.1.0). Baseline characteristics were summarised for all CHB patients (regardless of length of follow-up) using descriptive statistics. Means and standard deviations (SDs) or medians and interquartile ranges (IQRs) were presented for continuous measures, and were compared using *t* or Wilcoxon rank-sum tests, respectively. Patient counts and percentages were presented for continuous and binary variables, and were compared using chi-squared or Fisher's exact tests.

Univariable and multivariable Cox proportional hazards models were used to investigate risk factors for progression of CHB to HCC. Variables were included in the multivariable model based on significance of univariable associations (where $P \leq 0.1$) and/or based on biological/clinical relevance and previous literature (11,19,21–26,29). A previous meta-analysis we undertook to investigate risk factors for HCC in CHB was also used to inform variable selection (11). Patients in whom database exit date preceded or was equal to first recorded CHB diagnosis date (n = 1010) whereby follow-up time ≤ 0 years were excluded from longitudinal analysis. Continuous laboratory parameters which were right-skewed were transformed with a natural logarithm for inclusion in multivariable models. Laboratory parameters were divided into quintiles for inclusion in multivariable models. Means and SDs for log AST, log ALT and Plt quintiles are presented in supplement (**Supplementary Table 4**). Hazard ratios and 95% confidence intervals (95% CI) were reported for Cox proportional hazards model outputs. Analysis on the imputed dataset was used for main models.

Sensitivity analyses

Sensitivity analyses were performed for univariable and multivariable Cox proportional hazards whereby models were fitted to complete datasets for which missing values had been imputed. Values were missing for Townsend Deprivation Quintile, ethnicity, alcohol consumption, cigarette consumption, BMI, platelet count (Plt), alanine transaminase (ALT) measurement, aspartate transaminase (AST) measurement, HBsAg, and HBV VL. Missing data are described further in **Supplementary Table 5**.

Multiple imputation by chained equations (MICE) was used to impute missing data across patient characteristics. The assumption of missing at random was made for imputed variables. This is in accordance with previous handling of missing data in cohorts utilising QResearch data (34–37), and current recommendations for imputation of missing data (38). Characteristics with >90% missingness were not imputed. Ten imputed datasets were generated, and results from univariable and multivariable Cox proportional hazards models from each dataset were pooled according to Rubin's rules (39,40). Three sensitivity analyses were performed (**Supplementary Table 6**), as follows (i) all exposures in the complete-case analysis; (ii) exclusion of patients with history of non-HCC neoplasms (presented in **Supplementary Table 7)** to control for unmeasured outcome misclassification whereby secondary liver cancer has been misclassified as primary HCC; (iii) analysis in the imputed dataset additionally including ALT, AST and Plt, as the percentage of missingness in these exposures was too high for them to be included in main analysis.

Results

Cohort characteristics

We identified 8039 adults with CHB. Most (7856/8039, 97.7%) had a SNOMED/Read or ICD CHB diagnostic code with a remaining 2.3% (252/8039) identified by HBsAg/VL measurements (in the absence of a diagnostic code) (**Table 1**). Median follow-up duration was 3.87 years (interquartile range (IQR) 6.30 years), with

differential follow-up between individuals who developed HCC (median follow-up 1.47 years, IQR 5.13 years) and those who did not (median follow-up duration 3.93 years, IQR 6.28 years). Mean age at baseline was 38.3 years (SD 11.6 years), and at baseline >75% were ≤45 years of age. The majority were male (4856/8039, 60.4%).

A large percentage (3553/8039, 44.3%) of patients fell into the fifth Townsend Deprivation quintile (most deprived), while <20% are in the first or second quintiles (least deprived) (**Table 1**, **Figure 1**). Black Africans represented 25.4% of individuals in our cohort, with 12.9% of Chinese ethnicity, 5.9% of Pakistani ethnicity, 3.0% of Indian ethnicity and 28.4% of White ethnicity (**Table 1**). Proportions of Black and ethnic minorities in our cohort are greater than those in both the wider QResearch database (41) and general English population (42) (**Figure 1**).

Data for cigarette consumption, alcohol consumption and BMI were available for 71.5%, 56.1% and 61.3% of the cohort (**Table 1**). At baseline, most (88.0%) of patients had no record of antiviral treatment. Within 1, 2, 3 and ≥ 4 years of CHB diagnoses, cumulatively 2,7%, 4.0 %, 5.0% and 10.1% of patients had record of antiviral treatment initiation, respectively.

Baseline prevalence of T2DM and hypertension were 8.9% and 15.3%, respectively. This differs from the prevalence of T2DM and hypertension in the wider QResearch database of <8% and <3%, respectively (43). Prevalence of other comorbidities (including congestive heart failure, chronic kidney disease, alcohol-related liver disease, ascites, autoimmune hepatitis, cerebrovascular disease, end-stage liver disease, ischaemic heart disease, no-alcoholic fatty liver disease and peptic ulcer

disease; **Table 1**) ranged from 0.1% to 5%. Baseline cirrhosis is difficult to define from primary care EHR due to lack of relevant laboratory and imaging data, but 8.6% of the cohort had a diagnostic code indicating cirrhosis. Medication use was as follows: antidiabetic drugs (9.2%), antihypertensives (5.1%), non-steroidal anti-inflammatory drugs (NSAID) (5.9%) and statins (5.7%). Prevalence of non-HCC neoplasm was 4.7% in the overall cohort.

A description of cohort characteristics stratified by Townsend deprivation quintile is included (**Supplementary Table 8**).

Risk factors for HCC – imputed dataset analysis

Baseline characteristics of the imputed dataset used in analysis of HCC risk factors are presented in **Table 2**. Multivariable Cox proportional hazards models were constructed for 7029 patients in whom 161 HCC cases developed throughout follow-up (**Figure 3, Supplementary Table 9**). This translated to an HCC incidence rate of 4.96 cases per 1000 person-years (95% CI 4.29 to 5.63 cases per 1000 person-years). Hazards of HCC were increased in males (adjusted hazards ratio (aHR) 3.44, 95% CI 2.07 to 5.73), with increasing age (aHR for 36-45 years 1.96, 95% CI 1.10 to 3.48; aHR for 46-55 years 2.73, 95% CI 1.51 to 4.95; aHR for 56-65 years 3.17, 95% CI 1.63 to 6.14; aHR for ≥66 years 4.06, 95% CI 1.92 to 8.58, as compared to 26-35 years reference group) and in the fifth deprivation quintile as compared to the third quintile (aHR 1.69, 95% CI 1.01 to 2.84). Hazards of HCC in the Caribbean ethnicity group were higher than those in the White reference group (aHR 3.32, 95% CI 1.43 to 7.71) but did not differ in any other ethnic category. There were no associations

between alcohol consumption, cigarette smoking, or BMI with hazards of HCC. As expected, increased hazards of HCC were associated with evidence of advanced liver disease, cirrhosis (aHR 6.52, 95% CI 4.54 to 9.37) and ascites (aHR 1.85, 95% CI 1.02 to 3.36). Interestingly, peptic ulcer disease was also associated with increased HCC hazards (aHR 2.20, 95% CI 1.39 to 3.49). Statin use was associated with reduced hazards of HCC (aHR 0.47, 95% CI 0.22 to 0.99). No other medicines, including antiviral treatment, associated with hazards of HCC.

Hazards ratios did not change materially in strength or direction upon sensitivity analysis excluding non-HCC neoplasms or including AST, ALT and Plt at baseline (**Supplementary Table 9**).

Risk factors for HCC – complete case analysis

We undertook sensitivity analysis restricted to the subgroup of patients for whom complete data were available (n=3648 patients in whom 68 cases of HCC occurred (**Supplementary Table 10**)). This translated to 3.69 cases per 1000 person-years (95% CI 2.82 to 4.57 cases per 1000 person-years). Confidence intervals for all aHRs were more precise in the model from the imputed dataset compared to the complete case analysis due to larger sample size in the former. No associations of age, sex, deprivation quintile, ethnicity, BMI or alcohol consumption with hazards of HCC were identified. Hazards were higher in light smokers (aHR 2.29, 95% CI 1.15 to 4.57) as compared to non-smokers. The only comorbidities associated with increased hazards of HCC were chronic kidney disease (aHR 3.32, 95% CI 1.65 to 6.70), cerebrovascular disease (aHR 1.08, 95% CI 0.37 to 3.20), cirrhosis (aHR 8.79,

95% CI 4.85 to 15.93) and peptic ulcer disease (aHR 2.20, 95% CI 1.03 to 7.40). Hazard ratios did not change materially in strength in sensitivity analysis excluding patients with history of non-HCC neoplasms (**Supplementary Table 10**). No aHRs changed direction in this sensitivity analysis.

Discussion

The CHB population in the UK is ethnically and socioeconomically diverse

This is the largest cohort of CHB individuals characterised in the U.K. to date, from either EHR or traditional prospective cohorts. Data collection in QResearch is automatic and systematic, and therefore our findings are likely generalizable to the wider country, and potentially to other settings. The CHB population is ethnically diverse, with higher proportions of black and ethnic minority individuals than either the wider QResearch database (41) or the general English population (42). Importantly, the CHB population is disproportionately socioeconomically deprived, with substantial burdens of comorbid disease.

Risk factors for HCC

We report increased hazards of HCC associated with increasing age, male sex, severe deprivation, Caribbean ethnicity, severe liver disease (ascites and cirrhosis), and comorbid disease (peptic ulcer disease), while statin use was associated with reduced HCC hazards. Age, sex and T2DM have previously been found to positively associate with HCC risk in CHB (11).

It is possible that the positive association of peptic ulcer disease with hazards of HCC is confounded by proton pump inhibitor (PPI) administration for peptic ulcer treatment. Information concerning PPI prescription and usage was not available in our QResearch extract, but previous observational studies have reported increased risks of HCC associated with PPIs (44,45). Pooled risk estimates from meta-analyses have been variable, with some confirming this association (46) and others failing to report associations (47,48), particularly after restricting to studies of participants with viral hepatitis (48).

HBV genotype is associated with ethnicity (49,50), and viral genotypes have been associated with both an increased HCC risk (51) and antiviral treatment resistance (52). It is therefore possible that associations of HCC risk with ethnicity are confounded by HBV genotype. However, the association may also be mediated by population genetic factors and/or other exposures (including lifestyle factors) not captured in this dataset. Increased HCC hazards observed in the most deprived Townsend quintile may similarly be confounded by HBV genotype whereby certain genotypes associate with deprivation via region of infection. However, it is more likely the observed association is mediated by socioeconomic factors whereby access to routine care/surveillance is poorer and additional lifestyle or environmental factors which increase HCC risk are more prevalent.

We report a protective association of statin use with HCC hazards. This association has been reported in previous CHB cohorts (11,53,54), in individuals with predisposing HCC risk factors including cirrhosis and T2DM (55,56) and in a general

patient population (54–56). Further analysis, including potential mediation analysis where data availability allows, is warranted to investigate the mechanisms of this association.

HCC risk scores (including PAGE-B (19), REACH-B (20,21), GAG-HCC (22,23) and CU-HCC (24–26) utilise various characteristics, including age, sex, and laboratory parameters to predict HCC risk. The utility of existing risk scores in homogenous patient subgroups has been demonstrated (57). Validation and potentially modification of such scores in heterogenous samples which are ethnically and clinically diverse is warranted.

Caveats of primary care EHR analysis – missing data

We identified 8039 CHB patients in the QResearch database between 1999 and 2019 from a database-wide denominator of ~35 million individuals, which translates to a CHB prevalence of 0.023%. This contrasts with previous U.K. prevalence estimates ranging from 0.27% to 0.73% (58). Therefore, although we have characterised our cohort in the largest UK primary care EHR database, we highlight systematic data missingness due to poor primary care coverage and HBV representation in such data sources. Missingness likely associated with characteristics not measured in our dataset including thereby excluding marginalised subgroups (including undocumented migrants, populations who are highly mobile population subgroups, and people who do not speak English).

Most patients were identified by coding, with a minority (3.0%) having confirmatory laboratory tests accessible in QResearch. This is unsurprising given that specialist referral is recommended following a positive test result (59) thus second confirmatory HBV lab data (VL and/or HBsAg) would typically be performed in a secondary or tertiary care centre, and these results are not reliably transferred into primary care EHR systems.

Length of follow-up is shorter than anticipated for a chronic disease despite a 20-year study period from 1999 to 2019. This is not uncommon in EHR databases in which there may be time lags between notification of a patient characteristic/disease and input into electronic record systems. Additionally, given the neglected nature of CHB, it is likely that individuals present to primary care late in the infection course infection and/or diagnoses are primary managed in secondary/tertiary services without linkage to primary care records. Follow-up was differential between individuals who did and did not develop HCC, with shorter median follow-up duration in HCC cases. This may be due to HCC cases only presenting to a GP practice when their liver disease has progressed to a symptomatic, more advanced stage.

In our sample, 12% of individuals initiated antiviral treatment during follow-up. However, treatment data are most comprehensively captured in secondary/tertiary care records where HBV prescribing is based (60). In the NIHR Health Informatics Collaborative (HIC) secondary care EHR database, 16% of CHB patients were receiving antiviral treatment (61). Clinical guidelines recommend criteria for treatment, but it is not possible for us to determine the extent to which the 88% of

patients with no treatment recorded are treatment ineligible, or are receiving treatment in a secondary/tertiary care setting that is not captured in QResearch data.

Analytical corrections for missing data

We have undertaken imputation of missing baseline data, in line with previous investigations in QResearch (34–37), but were unable to impute important HBV biomarkers (including VL and HBsAg) as >90% of participants were missing any measurement at baseline. We therefore excluded these variables from analysis. High missingness was observed for additional biomarkers, specifically AST, ALT and Plt, which serve as indicators of liver function and can be used to derive proxy scores for staging fibrosis and cirrhosis. Replication of our present analysis in secondary care datasets is a future opportunity to investigate the utility of these biomarkers as markers of disease endpoints. Similarly, we were unable to time-update our models for changes in alcohol and cigarette consumption throughout follow-up due to lack of repeated measurements (>90% of the cohort had a one-off record for alcohol and smoking). Due to missing data, we were also unable to validate and/or modify such scores. Future investigation in data sources where variables are more complete and remeasurement is more comprehensive is warranted to enable these analytical improvements, and allow for more precise estimation of effect sizes.

Characterisation of comorbid disease in primary care EHR data sources is also limited due to missing information. For example, cirrhosis is inferred via biopsy, hepatic imaging or laboratory scores (62). However, these data are usually collected in secondary/tertiary care and are not transferred into primary care EHRs. Although

we identified a subset of individuals with a diagnostic code indicative of cirrhosis, cirrhotic prevalence in CHB was likely underestimated, and due to this missingness we were unable to investigate risk factors for cirrhosis as an endpoint. Similarly, the prevalence of comorbidities, including renal and cerebrovascular disease, is likely underestimated for this reason. and we are underpowered to detect associations with HCC risk.

Conclusions

The CHB population in the England is ethnically diverse and socioeconomically deprived. We have identified risk factors for HCC, and validated associations observed previously in different CHB populations. Future analysis warrants validation and/or modification of existing CHB-HCC risk scores in diverse patient samples. This may require establishment of a prospective cohort study with long-term follow-up of CHB-individuals sampled from a diverse population. Missingness in various data fields an important observation: it is a limitation both for identification of individuals with CHB, and for robust description of those individuals whose diagnosis is recorded. Enhanced capture of such data is crucial to provide an evidence base for interventions, including diagnostic screening, treatment and surveillance, modification of risk factors for HCC, and monitoring progress towards elimination targets.

Key findings

- The CHB population in the England is young, ethnically diverse and socioeconomically deprived.
- Increased hazards of HCC associated with male sex, older age, socioeconomic deprivation, Caribbean ethnicity (compared to White ethnicity), ascites, cirrhosis and peptic ulcer disease.
- Reduced HCC hazards associated with statin use
- Missing data suggest systematic underrepresentation of CHB individuals in primary care.

Ethics approval

The QResearch database consists of anonymised electronic health records. Ethics approval was granted by the East Midlands Derby Research Ethics Committee (reference 18/EM/0400). The protocol for this study (27) was reviewed by the QResearch Scientific Advisory Committee before approval was provided and access to relevant data for this project was accordingly granted, in accordance with ethical approval.

Acknowledgements

We acknowledge the contribution of EMIS practices who contribute to QResearch® and EMIS Health and the Chancellor Masters & Scholars of the University of Oxford for expertise in establishing, developing and supporting the QResearch database. QRESEARCH® is a registered trademark of Egton Medical Information Systems Limited and the University of Nottingham.

This project involves data derived from patient-level information collected by the NHS, as part of the care and support of cancer patients. The data is collated, maintained and quality assured by the National Cancer Registration and Analysis Service, which is part of Public Health England (PHE). Access to the data was facilitated by the PHE Office for Data Release.

This publication is based on data derived from the Intensive Care National Audit & Research Centre (ICNARC) Case Mix Programme Database. The Case Mix Programme is the national, comparative audit of patient outcomes from adult critical care coordinated by ICNARC. We thank all the staff in the critical care units participating in the Case Mix Programme. For more information on the representativeness and quality of these data, please contact ICNARC. Disclaimer: The views and opinions expressed therein are those of the authors and do not necessarily reflect those of ICNARC.

The Office of National Statistics, Public Health England and NHS Digital bear no responsibility for the analysis or interpretation of the data.

Funding

PCM is funded by Wellcome (grant ref 110110/Z/15/Z), UCLH NIHR Biomedical Research Centre and the Francis Crick Institute. CC's doctoral project is jointly funded by the Nuffield Department of Medicine, University of Oxford and by GlaxoSmithKline. EB is funded by the Oxford NIHR Biomedical Research Centre and is an NIHR Senior Investigator. The views expressed in this article are those of the author and not necessarily those of the NHS or the NIHR. PCM, EB and CC are

supported by the DeLIVER program “The Early Detection of Hepatocellular Liver Cancer” project is funded by Cancer Research UK (Early Detection Programme Award, grant reference: C30358/A29725). EB, TW, CC and PCM acknowledge support from the NIHR Health Informatics Collaborative.

Conflicts of interest

IAG is a full-time GSK employee and holds GSK shares. CC’s doctoral project is jointly funded by the Nuffield Department of Medicine, University of Oxford and by GSK.

Acknowledgements

Access to the data was facilitated by the PHE Office for Data Release. We acknowledge the contribution of EMIS practices who contribute to QResearch® and EMIS Health and the CHANCELLOR MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD for expertise in establishing, developing and supporting the QResearch database. United Kingdom Health Security Agency bear no responsibility for the analysis or interpretation of the data.

We would like to thank Professor Julia Hippisley-Cox and Rebekah Burrow in the Nuffield Department of Primary Care Health Sciences at the University of Oxford for their founding/direction of and facilitation of access to the QResearch database (respectively), and for their helpful feedback on analysis.

Accessibility of protocol, raw data, and programming code

Further details regarding the study protocol are available on the study webpage (27), and a full protocol can be made available upon request to the corresponding author. Researchers interested in utilising raw QResearch data are encouraged to seek further guidance on the QResearch webpage (63). Programming code can be made available upon request to the corresponding author

References

1. Sheena BS, Hiebert L, Han H, Ippolito H, Abbasi-Kangevari M, Abbasi-Kangevari Z, et al. Global, regional, and national burden of hepatitis B, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Gastroenterol Hepatol [Internet]. 2022 Jun 1 [cited 2022 Aug 10];7(9):796–829. Available from: http://www.thelancet.com/article/S2468125322001248/fulltext
2. Fitzmaurice C, Akinyemiju T, Abera S, Ahmed M, Alam N, Alemayohu MA, et al. The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level results from the global burden of disease study 2015. JAMA Oncol. 2017 Dec 1;3(12):1683–91.
3. Stanaway JD, Flaxman AD, Naghavi M, Fitzmaurice C, Vos T, Abubakar I, et al. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. The Lancet. 2016 Sep 10;388(10049):1081–8.
4. World Health Organization. Global Health Sector Strategies on viral hepatitis 2016–21. 2016.
5. DeLIVER: Early detection of hepatocellular liver cancer — OxCODE [Internet]. [cited 2022 Aug 11]. Available from: https://www.oxcode.ox.ac.uk/research-showcase/liver-cancer
6. Campbell C, Wang T, Burrow R, Mandal S, Hippisley-Cox J, Barnes E, et al. Estimating the epidemiology of chronic Hepatitis B Virus (HBV) infection in the UK: what do we know and what are we missing? Wellcome Open Research 2022 7:203 [Internet]. 2022 Aug 8 [cited 2022 Aug 10];7:203. Available from: https://wellcomeopenresearch.org/articles/7-203
7. Public Health England. Hepatitis B in the South East: 2016 data [Internet]. 2019 [cited 2019 Nov 4]. Available from: www.facebook.com/PublicHealthEngland
8. Public Health England. Hepatitis B in London: 2016 data [Internet]. 2019 [cited 2019 Nov 4]. Available from: www.facebook.com/PublicHealthEngland
9. Public Health England. Hepatitis B in the South West: 2017 data [Internet]. 2019 [cited 2019 Nov 4]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/842049/Hepatitis_B_in_the_South_West_2017_data__3_.pdf
10. Public Health England. Hepatitis B in the West Midlands: 2017 data [Internet]. 2019 [cited 2019 Nov 4]. Available from: www.facebook.com/PublicHealthEngland
11. Campbell C, Wang T, McNaughton AL, Barnes E, Matthews PC. Risk factors for the development of hepatocellular carcinoma (HCC) in chronic hepatitis B virus (HBV)

infection: a systematic review and meta-analysis. J Viral Hepat [Internet]. 2021 Mar 1 [cited 2022 Mar 3];28(3):493–507. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/jvh.13452

12. Lee MH, Yang HI, Liu J, Batrla-Utermann R, Jen CL, Iloeje UH, et al. Prediction models of long-term Cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: Risk scores integrating host and virus profiles. Hepatology [Internet]. 2013 Aug [cited 2020 Jun 8];58(2):546–54. Available from: https://aasldpubs.onlinelibrary.wiley.com/doi/full/10.1002/hep.26385
13. Yang HI, Yuen MF, Chan HLY, Han KH, Chen PJ, Kim DY, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): Development and validation of a predictive score. Lancet Oncol [Internet]. 2011 [cited 2020 Jun 8];12(6):568–74. Available from: https://pubmed.ncbi.nlm.nih.gov/21497551/
14. Wong GLH, Wong VWS. Risk prediction of hepatitis B virus-related hepatocellular carcinoma in the era of antiviral therapy. World J Gastroenterol [Internet]. 2013 Oct 21 [cited 2020 Jun 8];19(39):6515–22. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3801362/
15. Yuen MF, Tanaka Y, Fong DYT, Fung J, Wong DKH, Yuen JCH, et al. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. J Hepatol [Internet]. 2009 Jan [cited 2020 Jun 8];50(1):80–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18977053/
16. Chan HLY, Hui AY, Wong ML, Tse AML, Hung LCT, Wong VWS, et al. Genotype C hepatitis B virus infection is associated with an increased risk of hepatocellular carcinoma. Gut [Internet]. 2004 Oct [cited 2020 Jun 8];53(10):1494–8. Available from: https://gut.bmj.com/content/53/10/1494
17. Chan HLY, Tse CH, Mo F, Koh J, Wong VWS, Wong GLH, et al. High viral load and hepatitis B virus subgenotype Ce are associated with increased risk of hepatocellular carcinoma. Journal of Clinical Oncology [Internet]. 2008 Jan 10 [cited 2020 Jun 8];26(2):177–82. Available from: https://pubmed.ncbi.nlm.nih.gov/18182659/
18. Wong VWS, Chan SL, Mo F, Chan TC, Loong HHF, Wong GLH, et al. Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers. Journal of Clinical Oncology [Internet]. 2010 Apr 1 [cited 2020 Jun 8];28(10):1660–5. Available from: https://pubmed.ncbi.nlm.nih.gov/20194845/
19. Papatheodoridis G, Dalekos G, Sypsa V, Yurdaydin C, Buti M, Goulis J, et al. PAGE-B predicts the risk of developing hepatocellular carcinoma in Caucasians with chronic hepatitis B on 5-year antiviral therapy. J Hepatol [Internet]. 2016 Apr 1 [cited 2022 Jul 22];64(4):800–6. Available from: http://www.journal-of-hepatology.eu/article/S0168827815007953/fulltext
20. Yang HI, Yuen MF, Chan HLY, Han KH, Chen PJ, Kim DY, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. Lancet Oncol [Internet]. 2011 [cited 2022 Jul 22];12(6):568–74. Available from: https://pubmed.ncbi.nlm.nih.gov/21497551/
21. Lee MH, Yang HI, Liu J, Batrla-Utermann R, Jen CL, Iloeje UH, et al. Prediction models of long-term cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: risk scores integrating host and virus profiles. Hepatology [Internet]. 2013 Aug [cited 2022 Jul 22];58(2):546–54. Available from: https://pubmed.ncbi.nlm.nih.gov/23504622/
22. Wong GLH, Wong VWS. Risk prediction of hepatitis B virus-related hepatocellular carcinoma in the era of antiviral therapy. World J Gastroenterol [Internet]. 2013 Oct 21 [cited 2022 Jul 22];19(39):6515–22. Available from: https://pubmed.ncbi.nlm.nih.gov/24151375/

23. Yuen MF, Tanaka Y, Fong DYT, Fung J, Wong DKH, Yuen JCH, et al. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. J Hepatol [Internet]. 2009 Jan [cited 2022 Jul 22];50(1):80–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18977053/
24. Chan HLY, Hui AY, Wong ML, Tse AML, Hung LCT, Wong VWS, et al. Genotype C hepatitis B virus infection is associated with an increased risk of hepatocellular carcinoma. Gut [Internet]. 2004 Oct [cited 2022 Jul 22];53(10):1494–8. Available from: https://pubmed.ncbi.nlm.nih.gov/15361502/
25. Chan HLY, Tse CH, Mo F, Koh J, Wong VWS, Wong GLH, et al. High viral load and hepatitis B virus subgenotype ce are associated with increased risk of hepatocellular carcinoma. J Clin Oncol [Internet]. 2008 Jan 10 [cited 2022 Jul 22];26(2):177–82. Available from: https://pubmed.ncbi.nlm.nih.gov/18182659/
26. Wong VWS, Chan SL, Mo F, Chan TC, Loong HHF, Wong GLH, et al. Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers. J Clin Oncol [Internet]. 2010 Apr 1 [cited 2022 Jul 22];28(10):1660–5. Available from: https://pubmed.ncbi.nlm.nih.gov/20194845/
27. Characteristics of Chronic Hepatitis B associated with Cirrhosis and Cancer [Internet]. [cited 2020 Aug 18]. Available from: https://www.qresearch.org/research/approved-research-programs-and-projects/characteristics-of-chronic-hepatitis-b-associated-with-cirrhosis-and-cancer/
28. QResearch Database [Internet]. [cited 2020 Aug 18]. Available from: https://www.qresearch.org/
29. Ferreira G, Stuurman AL, Horsmans Y, Cattaert T, Verstraeten T, Feng Y, et al. Hepatitis B virus infection and the risk of liver disease progression in type 2 diabetic patients with potential nonalcoholic fatty liver disease: a retrospective, observational, cohort study in the United Kingdom Clinical Practice Research Datalink. Eur J Gastroenterol Hepatol. 2020 Jan 1;32(1):101–9.
30. List of ethnic groups - GOV.UK [Internet]. [cited 2022 Aug 12]. Available from: https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups#2001-census
31. A healthy lifestyle - WHO recommendations [Internet]. [cited 2022 Aug 12]. Available from: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations
32. Summary Analysis of Linked Cancer Data - QResearch [Internet]. [cited 2022 Feb 6]. Available from: https://www.qresearch.org/data/summary-analysis-of-linked-cancer-data/
33. Data - QResearch [Internet]. [cited 2022 Feb 6]. Available from: https://www.qresearch.org/data/
34. Saatci D, Ranger TA, Garriga C, Clift AK, Zaccardi F, Tan PS, et al. Association Between Race and COVID-19 Outcomes Among 2.6 Million Children in England. JAMA Pediatr [Internet]. 2021 Sep 1 [cited 2022 Feb 6];175(9):928–38. Available from: https://jamanetwork.com/journals/jamapediatrics/fullarticle/2780966
35. Hippisley-Cox J, Young D, Coupland C, Channon KM, Tan PS, Harrison DA, et al. Risk of severe COVID-19 disease with ACE inhibitors and angiotensin receptor blockers: cohort study including 8.3 million people. Heart [Internet]. 2020 Oct 1 [cited 2022 Feb 6];106(19):1503–11. Available from: https://heart.bmj.com/content/106/19/1503
36. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of hormone replacement therapy and risk of venous thromboembolism: nested case-control studies using the QResearch

and CPRD databases. BMJ [Internet]. 2019 Jan 9 [cited 2022 Feb 6];364:4810. Available from: https://www.bmj.com/content/364/bmj.k4810

37. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of combined oral contraceptives and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. [cited 2022 Feb 6]; Available from: www.cprd.com

38. van Buuren S. Flexible Imputation of Missing Data. Second Edition. [Internet]. Boca Raton, FL: Chapman & Hall/CRC. ; 2018 [cited 2022 Aug 16]. Available from: https://stefvanbuuren.name/fimd/

39. Rubin DB. Multiple Imputation for Nonresponse in Surveys [Internet]. 1st ed. Vol. 1. Cambridge, Massachusetts: John Wiley & Sons, Ltd; 1987 [cited 2022 Feb 6]. i–xxix. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/9780470316696.fmatter

40. Marshall A, Billingham LJ, Bryan S. Can we afford to ignore missing data in cost-effectiveness analyses? Eur J Health Econ [Internet]. 2009 Feb [cited 2022 Feb 6];10(1):1–3. Available from: https://pubmed.ncbi.nlm.nih.gov/18936993/

41. Summary of GP population characteristics - QResearch [Internet]. [cited 2022 Jul 19]. Available from: https://www.qresearch.org/data/summary-of-gp-population-characteristics/

42. Population estimates by ethnic group, England and Wales - Office for National Statistics [Internet]. [cited 2022 Jul 19]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/datasets/populationestimatesbyethnicgroupenglandandwales

43. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. BMJ Open [Internet]. 2014 Aug 1 [cited 2022 Aug 12];4(8):e005809. Available from: https://bmjopen.bmj.com/content/4/8/e005809

44. Shao YHJ, Chan TS, Tsai K, Wu SY. Association between proton pump inhibitors and the risk of hepatocellular carcinoma. Aliment Pharmacol Ther [Internet]. 2018 Aug 1 [cited 2022 Jul 20];48(4):460–8. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/apt.14835

45. Kim S, Jeong S, Park SJ, Chang J, Choi S, Cho Y, et al. Association between Proton Pump Inhibitor Use and Risk of Hepatocellular Carcinoma: A Korean Nationally Representative Cohort Study. Journal of Clinical Medicine 2022, Vol 11, Page 2865 [Internet]. 2022 May 19 [cited 2022 Jul 20];11(10):2865. Available from: https://www.mdpi.com/2077-0383/11/10/2865/htm

46. Song HJ, Jiang X, Henry L, Nguyen MH, Park H. Proton pump inhibitors and risk of liver cancer and mortality in patients with chronic liver disease: a systematic review and meta-analysis. European Journal of Clinical Pharmacology 2020 76:6 [Internet]. 2020 Mar 14 [cited 2022 Jul 20];76(6):851–66. Available from: https://link.springer.com/article/10.1007/s00228-020-02854-8

47. Wang J, Wu Y, Bi Q, Zheng X, Zhang J, Huang W. Adverse outcomes of proton pump inhibitors in chronic liver disease: a systematic review and meta-analysis. Hepatology International 2020 14:3 [Internet]. 2020 Jan 7 [cited 2022 Jul 20];14(3):385–98. Available from: https://link.springer.com/article/10.1007/s12072-019-10010-3

48. Kao WY, Su CW, Chia-Hui Tan E, Lee PC, Chen PH, Tang JH, et al. Proton Pump Inhibitors and Risk of Hepatocellular Carcinoma in Patients With Chronic Hepatitis B or C. Hepatology [Internet]. 2019 Mar 1 [cited 2022 Jul 20];69(3):1151–64. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/hep.30247

49. Lin CL, Kao JH. The clinical implications of hepatitis B virus genotype: Recent advances. J Gastroenterol Hepatol [Internet]. 2011 [cited 2022 Jul 22];26(SUPPL.

1):123–30. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1440-1746.2010.06541.x

50. Lin CL, Kao JH. Hepatitis B Virus Genotypes and Variants. Cold Spring Harb Perspect Med [Internet]. 2015 May 1 [cited 2022 Jul 22];5(5):a021436. Available from: http://perspectivesinmedicine.cshlp.org/content/5/5/a021436.full

51. Wong GLH, Chan HLY, Yiu KKL, Lai JWY, Chan VKK, Cheung KKC, et al. Meta-analysis: the association of hepatitis B virus genotypes and hepatocellular carcinoma. Aliment Pharmacol Ther [Internet]. 2013 Mar 1 [cited 2022 Jul 22];37(5):517–26. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/apt.12207

52. Araujo NM, Teles SA, Spitz N. Comprehensive Analysis of Clinically Significant Hepatitis B Virus Mutations in Relation to Genotype, Subgenotype and Geographic Region. Front Microbiol. 2020 Dec 14;11:3241.

53. Goh MJ, Sinn DH, Kim S, Woo SY, Cho H, Kang W, et al. Statin Use and the Risk of Hepatocellular Carcinoma in Patients With Chronic Hepatitis B. Hepatology [Internet]. 2020 Jun 1 [cited 2022 Jul 22];71(6):2023–32. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/hep.30973

54. Facciorusso A, el Aziz MAA, Singh S, Pusceddu S, Milione M, Giacomelli L, et al. Statin Use Decreases the Incidence of Hepatocellular Carcinoma: An Updated Meta-Analysis. Cancers 2020, Vol 12, Page 874 [Internet]. 2020 Apr 3 [cited 2022 Jul 22];12(4):874. Available from: https://www.mdpi.com/2072-6694/12/4/874/htm

55. Islam MM, Poly TN, Walther BA, Yang HC, Li YC. Statin Use and the Risk of Hepatocellular Carcinoma: A Meta-Analysis of Observational Studies. Cancers 2020, Vol 12, Page 671 [Internet]. 2020 Mar 13 [cited 2022 Jul 22];12(3):671. Available from: https://www.mdpi.com/2072-6694/12/3/671/htm

56. Kim G, Jang SY, Nam CM, Kang ES. Statin use and the risk of hepatocellular carcinoma in patients at high risk: A nationwide nested case-control study. J Hepatol. 2018 Mar 1;68(3):476–84.

57. Voulgaris T, Papatheodoridi M, Lampertico P, Papatheodoridis G v. Clinical utility of hepatocellular carcinoma risk scores in chronic hepatitis B. Liver International [Internet]. 2020 Mar 1 [cited 2022 Aug 17];40(3):484–95. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/liv.14334

58. Campbell C, Wang T, Burrow R, Mandal S, Hippisley-Cox J, Barnes E, et al. Estimating the epidemiology of chronic Hepatitis B Virus (HBV) infection in the UK: what do we know and what are we missing? 2022 Apr 13 [cited 2022 Aug 12]; Available from: https://europepmc.org/article/PPR/PPR482357

59. National Institute for Health and Care Excellence. Hepatitis B (chronic): diagnosis and management Clinical guideline. 2013 [cited 2022 Jul 20]; Available from: www.nice.org.uk/guidance/cg165

60. Wang T, Smith DA, Campbell C, Freeman O, Moysova Z, Noble T, et al. Cohort Profile: The National Institute for Health Research Health Informatics Collaborative: Hepatitis B Virus (NIHR HIC HBV) research dataset. Int J Epidemiol [Internet]. 2022 Jun 16 [cited 2022 Jul 6];2022:1–11. Available from: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyac127/6609198

61. Campbell C, Wang T, Smith DA, Freeman O, Noble T, Várnai KA, et al. Impact of the COVID-19 pandemic on routine surveillance for adults with chronic hepatitis B virus (HBV) infection in the UK. [cited 2022 Aug 16]; Available from: https://doi.org/10.1101/2021.11.10.21265651

62. Cirrhosis in over 16s Assessment and management NICE guideline NG50.

63. Information for Researchers - QResearch [Internet]. [cited 2022 Aug 22]. Available from: https://www.qresearch.org/information/information-for-researchers/

1

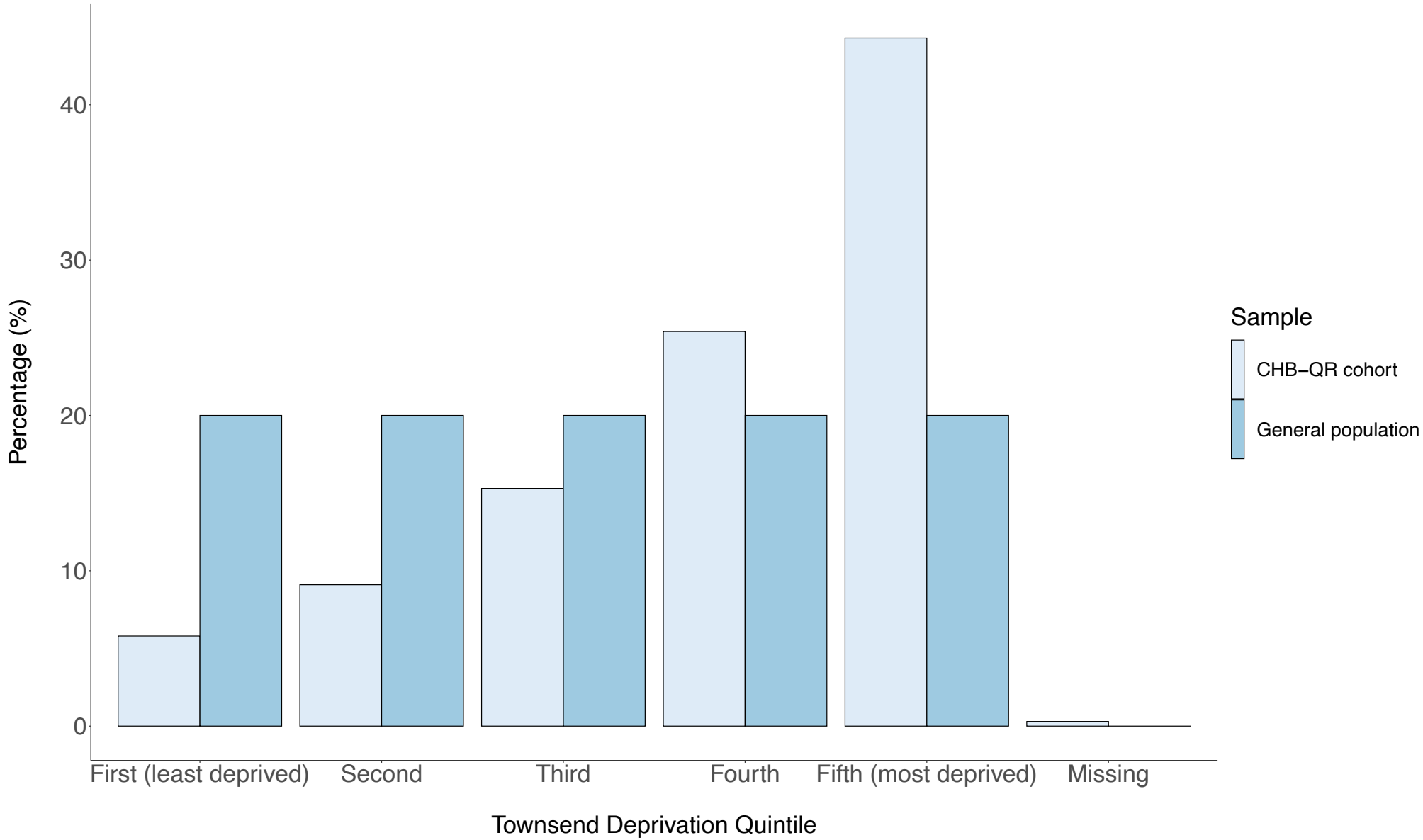


Figure 1. Townsend deprivation quintile breakdown in 8039 adults with chronic hepatitis B virus infection derived from the QResearch primary care database (England) versus the United Kingdom general population.

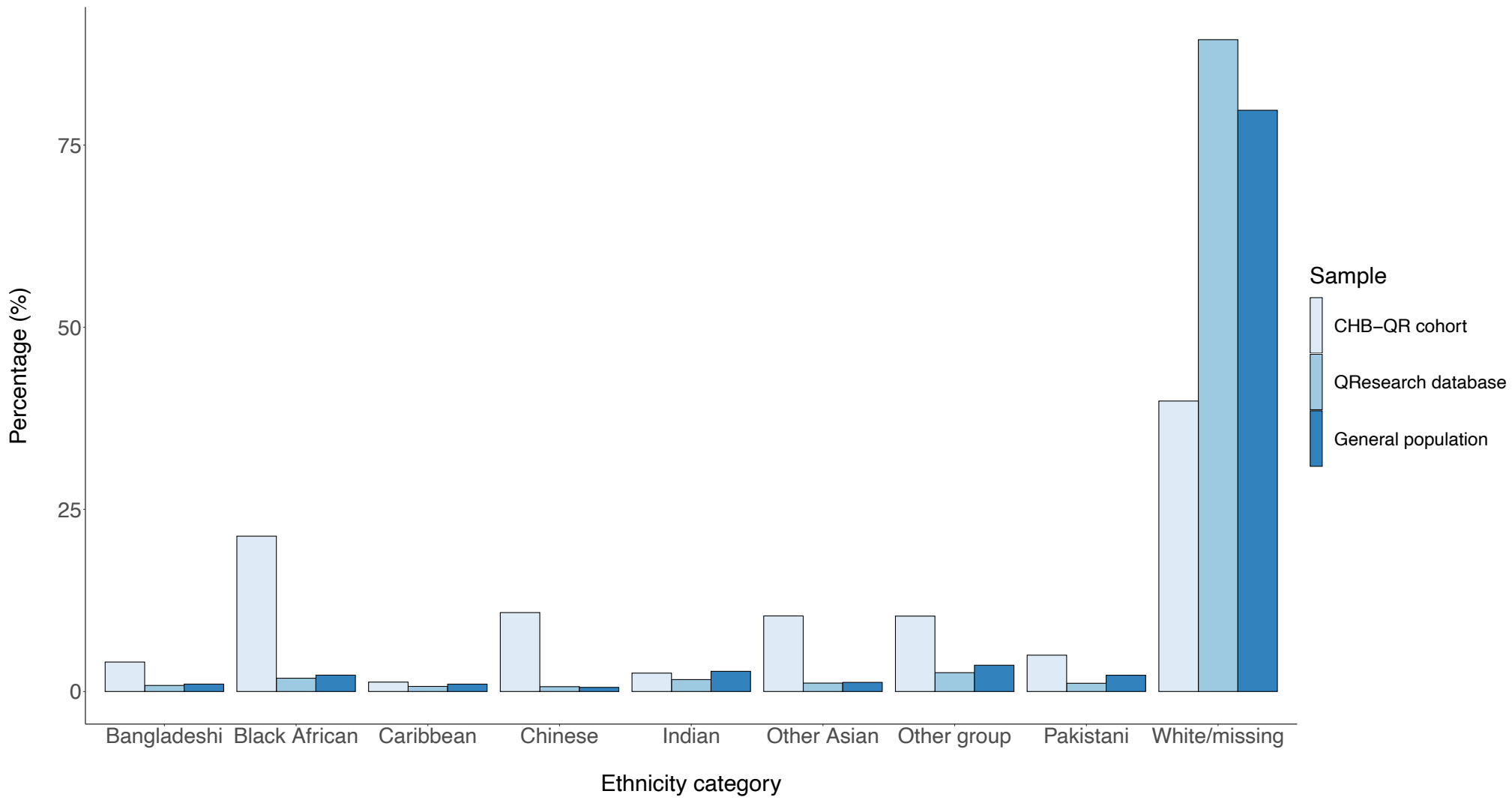
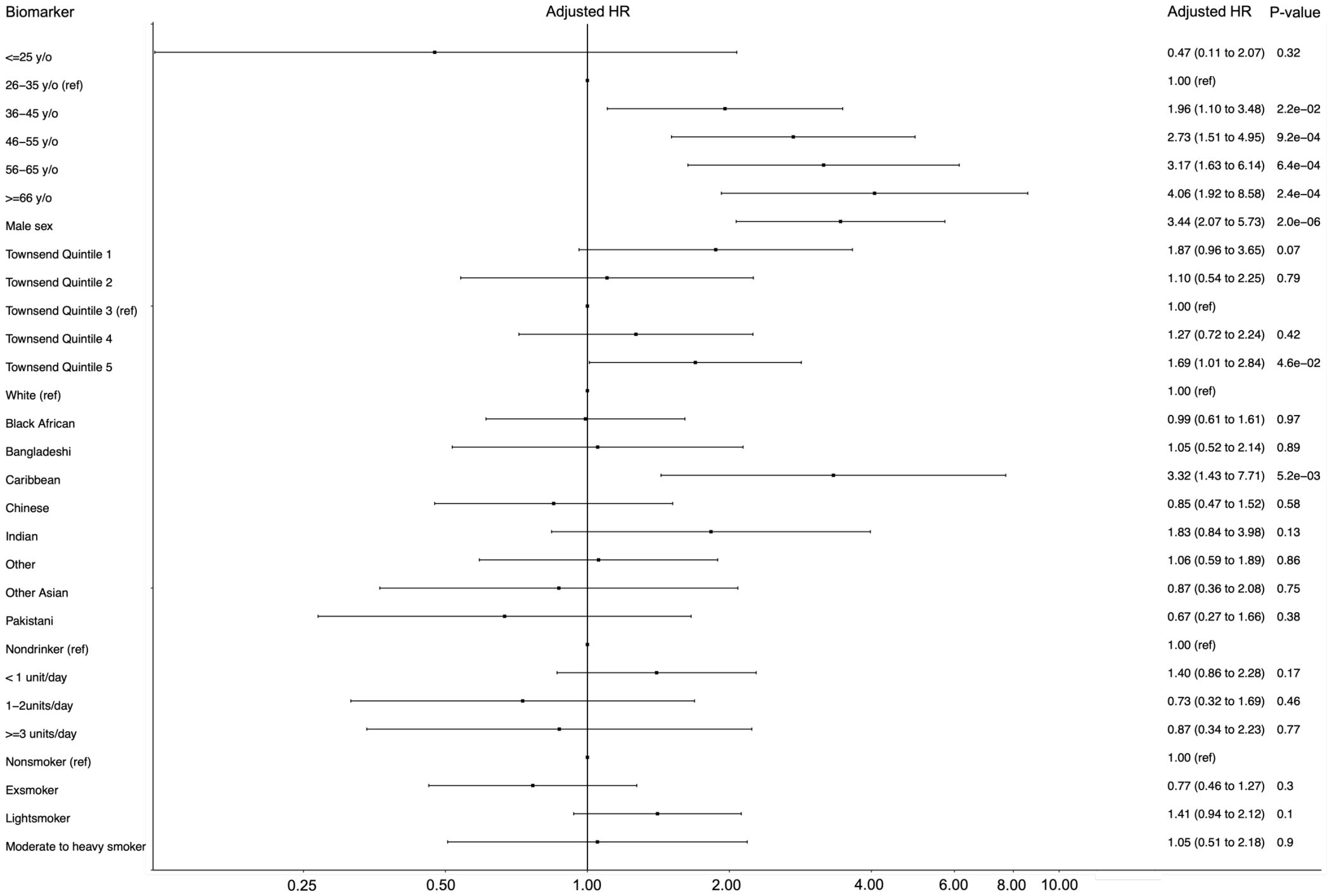


Figure 2. Ethnicity breakdown in 8039 adults in a chronic hepatitis B virus cohort characterised from the QResearch primary care database (England) versus the United Kingdom general population. General population estimates obtained from 2019 estimates from the Office for National Statistics.


Biomarker
Adjusted HR
Adjusted HR
P-value
<=25 y/o
0.47 (0.11 to 2.07)
0.32
26–35 y/o (ref)
1.00 (ref)
36–45 y/o
1.96 (1.10 to 3.48)
2.2e–02
46–55 y/o
2.73 (1.51 to 4.95)
9.2e–04
56–65 y/o
3.17 (1.63 to 6.14)
6.4e–04
>=66 y/o
4.06 (1.92 to 8.58)
2.4e–04
Male sex
3.44 (2.07 to 5.73)
2.0e–06
Townsend Quintile 1
1.87 (0.96 to 3.65)
0.07
Townsend Quintile 2
1.10 (0.54 to 2.25)
0.79
Townsend Quintile 3 (ref)
1.00 (ref)
Townsend Quintile 4
1.27 (0.72 to 2.24)
0.42
Townsend Quintile 5
1.69 (1.01 to 2.84)
4.6e–02
White (ref)
1.00 (ref)
Black African
0.99 (0.61 to 1.61)
0.97
Bangladeshi
1.05 (0.52 to 2.14)
0.89
Caribbean
3.32 (1.43 to 7.71)
5.2e–03
Chinese
0.85 (0.47 to 1.52)
0.58
Indian
1.83 (0.84 to 3.98)
0.13
Other
1.06 (0.59 to 1.89)
0.86
Other Asian
0.87 (0.36 to 2.08)
0.75
Pakistani
0.67 (0.27 to 1.66)
0.38
Nondrinker (ref)
1.00 (ref)
< 1 unit/day
1.40 (0.86 to 2.28)
0.17
1–2units/day
0.73 (0.32 to 1.69)
0.46
>=3 units/day
0.87 (0.34 to 2.23)
0.77
Nonsmoker (ref)
1.00 (ref)
Exsmoker
0.77 (0.46 to 1.27)
0.3
Lightsmoker
1.41 (0.94 to 2.12)
0.1
Moderate to heavy smoker
1.05 (0.51 to 2.18)
0.9
0.25
0.50
1.00
2.00
4.00
6.00
8.00
10.00

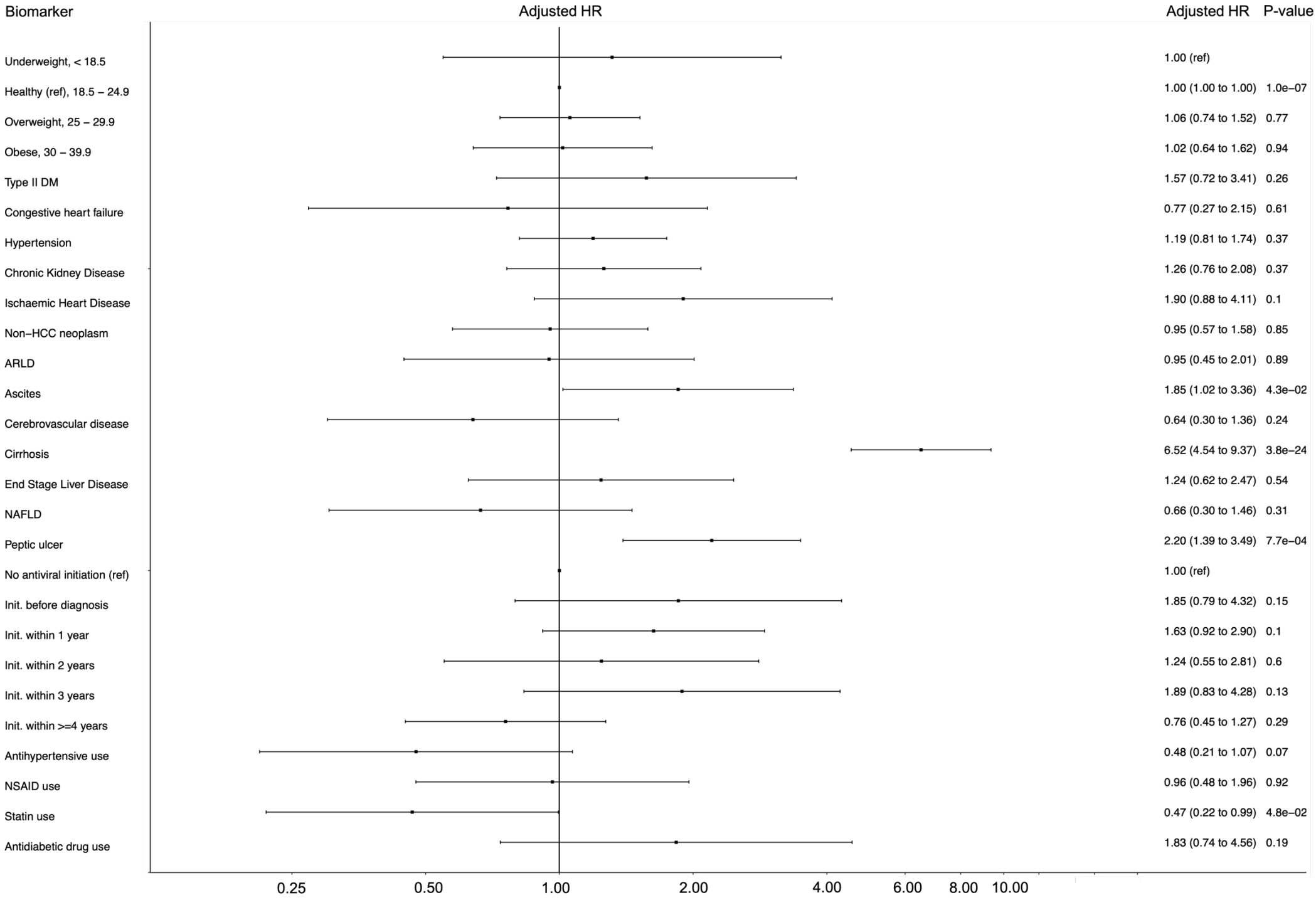

Biomarker
Adjusted HR
Adjusted HR
P-value
Underweight, < 18.5
1.00 (ref)
Healthy (ref), 18.5 – 24.9
1.00 (1.00 to 1.00) 1.0e−07
Overweight, 25 – 29.9
1.06 (0.74 to 1.52) 0.77
Obese, 30 – 39.9
1.02 (0.64 to 1.62) 0.94
Type II DM
1.57 (0.72 to 3.41) 0.26
Congestive heart failure
0.77 (0.27 to 2.15) 0.61
Hypertension
1.19 (0.81 to 1.74) 0.37
Chronic Kidney Disease
1.26 (0.76 to 2.08) 0.37
Ischaemic Heart Disease
1.90 (0.88 to 4.11) 0.1
Non−HCC neoplasm
0.95 (0.57 to 1.58) 0.85
ARLD
0.95 (0.45 to 2.01) 0.89
Ascites
1.85 (1.02 to 3.36) 4.3e−02
Cerebrovascular disease
0.64 (0.30 to 1.36) 0.24
Cirrhosis
6.52 (4.54 to 9.37) 3.8e−24
End Stage Liver Disease
1.24 (0.62 to 2.47) 0.54
NAFLD
0.66 (0.30 to 1.46) 0.31
Peptic ulcer
2.20 (1.39 to 3.49) 7.7e−04
No antiviral initiation (ref)
1.00 (ref)
Init. before diagnosis
1.85 (0.79 to 4.32) 0.15
Init. within 1 year
1.63 (0.92 to 2.90) 0.1
Init. within 2 years
1.24 (0.55 to 2.81) 0.6
Init. within 3 years
1.89 (0.83 to 4.28) 0.13
Init. within >=4 years
0.76 (0.45 to 1.27) 0.29
Antihypertensive use
0.48 (0.21 to 1.07) 0.07
NSAID use
0.96 (0.48 to 1.96) 0.92
Statin use
0.47 (0.22 to 0.99) 4.8e−02
Antidiabetic drug use
1.83 (0.74 to 4.56) 0.19
0.25
0.50
1.00
2.00
4.00
6.00
8.00
10.00

Figure 3. Forest plot for Cox proportional hazards model to identify risks of hepatocellular carcinoma (HCC) in an adult population with chronic Hepatitis B virus infection derived from the QResearch primary care database. Analysed using a dataset generated by multiple imputation with chained equations (n = 7029, HCC cases = 161). First (least deprived) to fifth (most deprived) Townsend Deprivation Quintiles are denoted by SES1-5, respectively.

Table 1: Cohort baseline characteristics

Characteristic	**Total, n (%)**	**HCC cases, n (%)**	**Non-HCC cases, n (%)**
N	8039	210	7829
Follow-up time (years), median (IQR)	3.87 (6.30)	1.47 (5.13)	3.93 (6.28)
CHB diagnosis source, n (%) *			
Any ICD	3952 (49.2%)	47 (22.4%)	3789 (48.4%)
Any SNOMED/Read	4425 (55.0%)	68 (32.4%)	4357 (55.7%)
Any laboratory test	240 (3.0%)	<5	238 (3.0%)
Age (years), mean (SD)	38.3 (11.6)	50.3 (12.4)	37.9 (11.4)
Age group, n (%)			
18-25 years	860 (10.7%)	<5	857 (10.9%)
26-35 years	2930 (36.4%)	22 (10.5%)	2908 (37.1%)
36-45 years	2346 (29.2%)	52 (24.8%)	2294 (29.3%)
46-55 years	1182 (14.7%)	63 (30.0%)	1119 (14.3%)
56-65 years	502 (6.2%)	41 (19.5%)	461 (5.9%)
≥66 years	219 (2.7%)	29 (13.8%)	190 (2.4%)
Sex, n (%)			
Female	3183 (39.6%)	24 (11.4%)	3159 (40.3%)
SMale	4856 (60.4%)	186 (88.6%)	4670 (59.7%)
Townsend Deprivation Quintile, n (%)			
First (least deprived)	467 (5.8%)	27 (12.9%)	440 (5.6%)
Second	731 (9.1%)	20 (9.5%)	711 (9.1%)
Third	1226 (15.3%)	27 (12.9%)	1199 (15.3%)
Fourth	2040 (25.4%)	48 (22.9%)	1992 (25.4%)
Fifth (most deprived)	3553 (44.3%)	87 (41.4%)	3466 (44.3%)
Missing	22 (0.3%)	<5	21 (0.3%)
Ethnicity, n (%)			
Black African	1714 (25.4%)	35 (22.3%)	1679 (25.5%)
Bangladeshi	325 (4.8%)	11 (7.0%)	314 (4.8%)
Caribbean	104 (1.5%)	7 (4.5%)	97 (1.5%)
Chinese	871 (12.9%)	21 (13.4%)	850 (12.9%)
Indian	203 (3.0%)	7 (4.5%)	196 (3.0%)
Other	832 (12.3%)	15 (9.6%)	817 (12.4%)
Other Asian	384 (5.7%)	5 (3.2%)	379 (5.7%)

Characteristic	Total, n (%)	HCC cases, n (%)	Non-HCC cases, n (%)
Pakistani	401 (5.9%)	7 (4.5%)	394 (6.0%)
White	1916 (28.4%)	49 (31.2%)	1867 (28.3%)
Missing	1289 (16.0%)	53 (25.2%)	1236 (15.8%)
Cigarette consumption, n (%)			
Non-smoker	3810 (66.2%)	78 (54.2%)	3732 (66.5%)
Ex-smoker	768 (13.4%)	25 (17.4%)	743 (13.2%)
Light smoker (1-9 cigarettes/day)	908 (15.8%)	31 (21.5%)	877 (15.6%)
Moderate to heavy smoker (≥10 cigarettes per day)	266 (4.6%)	10 (6.9%)	256 (4.6%)
Missing	2287 (28.5%)	66 (31.4%)	2221 (28.4%)
Alcohol consumption, n (%)			
Non-drinker	3697 (81.9%)	100 (83.3%)	3597 (81.9%)
Trivial drinker (<1 unit per day)	485 (10.7%)	13 (10.8%)	472 (10.7%)
Light drinker (1-2 units per day)	209 (4.6%)	<5	206 (4.7%)
Moderate to heavy drinker (≥3 units per day)	121 (2.7%)	<5	117 (2.7%)
Missing	3527 (43.9%)	90 (42.9%)	3439 (43.9%)
BMI (kg/m^2), mean (SD)	26.05 (5.06)	26.13 (4.63)	26.05 (5.07)
BMI, n (%)			
Underweight (<18.5 kg/m^2)	159 (3.2%)	<5	156 (3.2%)
Normal weight (18.5-24.9 kg/m^2)	2092 (42.5%)	49 (39.8%)	2043 (42.5%)
Overweight (25.0-29.9 kg/m^2)	1734 (35.2%)	52 (42.3%)	1682 (35.0%)
Obese (≥30 kg/m^2)	943 (19.1%)	19 (15.4%)	924 (19.2%)
Missing	3111 (38.7%)	87 (41.4%)	3024 (38.6%)
Antiviral initiation, n (%)			
No initiation	7076 (88.0%)	140 (66.7%)	6936 (88.6%)
Before CHB diagnosis	138 (1.7%)	11 (5.2%)	127 (1.6%)

Characteristic	Total, n (%)	HCC cases, n (%)	Non-HCC cases, n (%)
Within 1 year of CHB diagnosis	221 (2.7%)	24 (11.4%)	197 (2.5%)
Within 2 years of CHB diagnosis	120 (1.5%)	7 (3.3%)	113 (1.4%)
Within 3 years of CHB diagnosis	78 (1.0%)	7 (3.3%)	71 (0.9%)
≥4 years after CHB diagnosis	406 (5.1%)	21 (10.0%)	385 (4.9%)
Type 2 DM, n (%)	715 (8.9%)	21 (10.0%)	664 (8.5%)
Congestive heart failure, n (%)	75 (0.9%)	7 (3.3%)	68 (0.9%)
Hypertension, n (%)	1229 (15.3%)	71 (33.8%)	1158 (14.8%)
Chronic kidney disease, n (%)	287 (3.6%)	29 (13.8%)	258 (3.3%)
Alcohol-related liver disease, n (%)	93 (1.2%)	19 (9.0%)	74 (0.9%)
Ascites, n (%)	114 (1.4%)	32 (15.2%)	82 (1.0%)
Autoimmune hepatitis, n (%)	9 (0.1%)	<5	9 (0.1%)
Cerebrovascular disease, n (%)	310 (3.9%)	21 (10.0%)	289 (3.7%)
Cirrhosis, n (%)	689 (8.6%)	121 (57.6%)	568 (7.3%)
End-stage liver disease, n (%)	85 (1.1%)	16 (7.6%)	69 (0.9%)
Ischaemic heart disease, n (%)	255 (3.2%)	20 (9.5%)	235 (3.0%)
Non-alcoholic fatty liver disease, n (%)	404 (5.0%)	8 (3.8%)	396 (5.1%)
Non-HCC neoplasm, n (%)	381 (4.7 %)	345 (4.4%)	36 (17.1%)
Peptic ulcer, n (%)	248 (3.1%)	32 (15.2%)	216 (2.8%)
Antidiabetic drug use, n (%)	737 (9.2%)	49 (23.3%)	688 (8.8%)
Antihypertensive use, n (%)	408 (5.1%)	22 (10.5%)	386 (4.9%)
NSAID use, n (%)	472 (5.9%)	24 (11.4%)	448 (5.7%)
Statin use, n (%)	457 (5.7%)	19 (9.0%)	438 (5.6%)

CHB, chronic hepatitis B virus infection; ICD, International classification of disease; SNOMED, Systemised Nomenclature of Medicine; SD, standard deviation; BMI, body mass index; DM, diabetes mellitus; HCC, hepatocellular carcinoma; NSAID, non-steroidal anti-inflammatory drug.

* Some patients have diagnostic indicators of CHB from multiple sources and therefore there overlap between categories.

Table 2. Baseline characteristics of complete-case and imputed (by multiple imputation with chained equations) datasets used in Cox proportional hazards models.

Characteristic	**Complete-case dataset, n (%)**	**Imputed dataset, n (%)**
N	3648	7029*
HCC cases, n (%)	68 (1.9%)	161 (2.3%)
Age group, n (%)		
18-25 years	292 (8.0)	659 (9.4)
26-35 years	1253 (34.3)	2500 (35.6)
36-45 years	1126 (30.9)	2108 (30.0)
46-55 years	593 (16.3)	1086 (15.5)
56-65 years	270 (7.4)	468 (6.7)
≥66 years	114 (3.1)	208 (3.0)
Sex, n (%)		
Male	2082 (57.1)	4186 (59.6)
Townsend Deprivation Quintile, n (%)		
First (least deprived)	174 (4.8)	419 (6.0)
Second	261 (7.2)	645 (9.2)
Third	523 (14.3)	1094 (15.6)
Fourth	924 (25.3)	1781 (25.3)
Fifth (most deprived)	1766 (48.4)	3090 (44.0)
Ethnicity, n (%)		
Black African	956 (26.2)	1772 (25.2)
Bangladeshi	213 (5.8)	322 (4.6)
Caribbean	58 (1.6)	112 (1.6)
Chinese	428 (11.7)	894 (12.7)
Indian	131 (3.6)	224 (3.2)
Other	441 (12.1)	863 (12.3)
Other Asian	194 (5.3)	397 (5.6)
Pakistani	241 (6.6)	420 (6.0)
White	986 (27.0)	2025 (28.8)
Cigarette consumption, n (%)		
Non-smoker	2484 (68.1)	4661 (66.3)
Ex-smoker	480 (13.2)	924 (13.1)
Light smoker (1-9 cigarettes/day)	531 (14.6)	1110 (15.8)

Characteristic	Complete-case dataset, n (%)	Imputed dataset, n (%)
Moderate to heavy smoker (≥10 cigarettes per day)	153 (4.2)	334 (4.8)
Alcohol consumption, n (%)		
Non-drinker	3013 (82.6)	5784 (82.3)
Trivial drinker (<1 unit per day)	371 (10.2)	737 (10.5)
Light drinker (1-2 units per day)	170 (4.7)	318 (4.5)
Moderate to heavy drinker (≥3 units per day)	94 (2.6)	190 (2.7)
BMI (kg/m^2), mean (SD)	26.14 (5.07)	26.05 (5.02)
BMI, n (%)		
Underweight (<18.5 kg/m^2)	113 (3.1)	222 (3.2)
Normal weight (18.5-24.9 kg/m^2)	1530 (41.9)	2948 (41.9)
Overweight (25.0-29.9 kg/m^2)	1292 (35.4)	2527 (36.0)
Obese (≥30 kg/m^2)	713 (19.5)	1332 (19.0)
Antiviral initiation, n (%)		
No initiation	3168 (86.8)	6092 (86.7)
Before CHB diagnosis	67 (1.8)	121 (1.7)
Within 1 year of CHB diagnosis	116 (3.2)	214 (3.0)
Within 2 years of CHB diagnosis	73 (2.0)	120 (1.7)
Within 3 years of CHB diagnosis	49 (1.3)	78 (1.1)
≥4 years after CHB diagnosis	175 (4.8)	404 (5.7)
Type 2 DM, n (%)	503 (13.8)	678 (9.6)
Congestive heart failure, n (%)	41 (1.1)	71 (1.0)
Hypertension, n (%)	754 (20.7)	1165 (16.6)

Characteristic	Complete-case dataset, n (%)	Imputed dataset, n (%)
Chronic kidney disease, n (%)	175 (4.8)	271 (3.9)
Alcohol-related liver disease, n (%)	39 (1.1)	74 (1.1)
Ascites, n (%)	34 (0.9)	88 (1.3)
Autoimmune hepatitis, n (%)	3 (0.1)	9 (0.1)
Cerebrovascular disease, n (%)	181 (5.0)	290 (4.1)
Cirrhosis, n (%)	301 (8.3)	608 (8.6)
End-stage liver disease, n (%)	28 (0.8)	78 (1.1)
Ischaemic heart disease, n (%)	143 (3.9)	229 (3.3)
Non-alcoholic fatty liver disease, n (%)	248 (6.8)	393 (5.6)
Non-HCC neoplasm, n (%)	167 (4.6)	340 (4.8)
Peptic ulcer, n (%)	113 (3.1)	221 (3.1)
Antidiabetic drug use , n (%)	495 (13.6)	689 (9.8)
Antihypertensive use, n (%)	293 (8.0)	389 (5.5)
NSAID use, n (%)	331 (9.1)	446 (6.3)
Statin use, n (%)	342 (9.4)	443 (6.3)

CHB, chronic hepatitis B virus infection; ICD, International classification of disease; SNOMED, Systemised Nomenclature of Medicine; IQR, interquartile range; SD, standard deviation; BMI, body mass index; DM, diabetes mellitus; HCC, hepatocellular carcinoma; NSAID, non-steroidal anti-inflammatory drug.