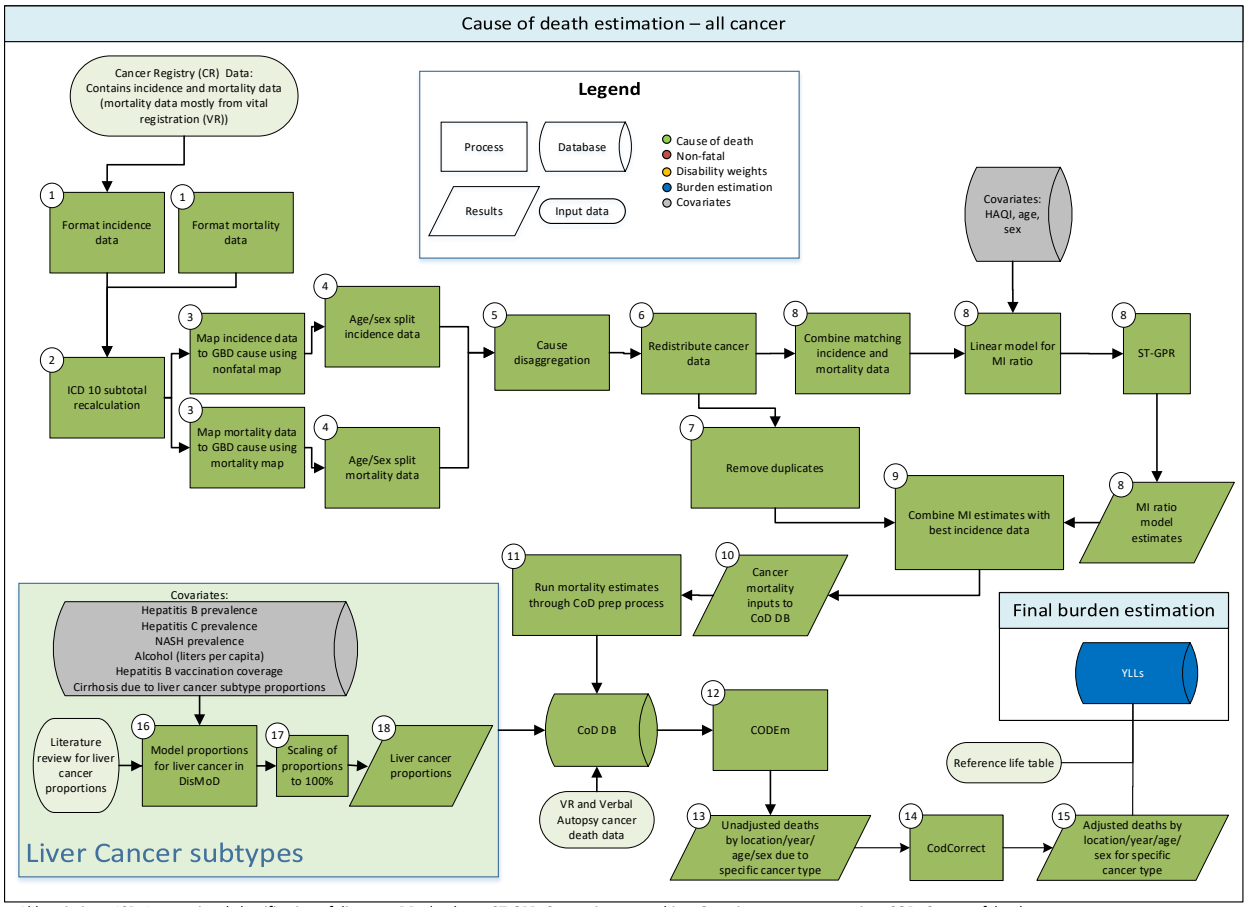
Cancers

Input data and methodological summary for all cancers except for non-melanoma skin cancer



Abbreviations: ICD: International classification of diseases; DB: database, ST-GPR: Space-time smoothing, Gaussian process regression, COD: Causes of death

Data

The cause of death (COD) database contains multiple sources of cancer mortality data. These sources include vital registration, verbal autopsy, and cancer registry data. The cancer registry mortality estimates that are uploaded into the COD database stem from cancer registry incidence data that have been transformed to mortality estimates through the use of mortality-to-incidence ratios (MIR).

Data-seeking processes

Cancer mortality data in the cause of death database other than cancer registry data

Sources for cancer mortality data other than cancer registry data are described in the COD database description (Appendix Section 2.2).

Cancer registry data

Cancer registry data were used from publicly available sources or provided by collaborators. We used all data from GBD 2017 and added registry data from Argentina, Australia, Austria, Bermuda, Canada, Chile, China, Colombia, Germany, Netherlands, Switzerland, United Kingdom, Uruguay, and Yemen.

Inclusion and exclusion criteria

Only population-based cancer registries were included, and only those that included all cancers (no specialty registries), data for all age groups (except for paediatric cancer registries), and data for both sexes. Pathology-based cancer registries were included if they had a defined population. Hospital-based cancer registries were excluded.

Cancer registry data were excluded from either the final incidence data input or the MI model input if a more detailed source (eg, providing more detailed age or diagnostic groups) was available for the same population. Preference was given to registries with national coverage over those with only local coverage, except those from countries where the GBD study provides subnational estimates.

Data were excluded if the coverage population was unknown.

Bias of categories of input data

Cancer registry data can be biased in multiple ways. A high proportion of ill-defined cancer cases in the registry data requires redistribution of these cases to other cancers, which introduces a potential for bias. Changes between coding systems can lead to artificial differences in disease estimates; however, we adjust for this bias by mapping the different coding systems to the GBD causes. Underreporting of cancers that require advanced diagnostic techniques (eg, leukaemia, brain, pancreatic, and liver cancer) can be an issue in cancer registries from low-income countries. On the other hand, misclassification of metastatic sites as primary cancer can lead to overestimation of cancer sites that are common sites for metastases, like the brain or liver. Since many cancer registries are located in urban areas, the representativeness of the registry for the general population can also be problematic. The accuracy of mortality data reported in cancer registries usually depends on the quality of the vital registration system. If the vital registration system is incomplete or of poor quality, the mortality-to-incidence ratio can be biased to lower ratios.

Data for liver cancer aetiology splits

To find the proportion of liver cancer cases due to the five aetiology groups included in GBD (1. Liver cancer due to hepatitis B, 2. Liver cancer due to hepatitis C, 3. Liver cancer due to alcohol, 4. Liver cancer due to non-alcoholic steatohepatitis (NASH), 5. Liver cancer due to other causes), a systematic literature search was performed in PubMed on 10/24/2016 using the following search string: “("liver neoplasms"[All Fields] OR "HCC"[All Fields] OR "liver cancer"[All Fields] OR "Carcinoma, Hepatocellular"[Mesh]) AND (("hepatitis B"[All Fields] OR "Hepatitis B"[Mesh] OR "Hepatitis B virus"[Mesh] OR "Hepatitis B Antibodies"[Mesh] OR "Hepatitis B Antigens"[Mesh]) OR ("hepatitis C"[All Fields] OR "Hepatitis C"[Mesh] OR "hepatitis C antibodies"[MESH] OR "Hepatitis C Antigens"[Mesh] OR "Hepacivirus"[Mesh]) OR ("alcohol"[All Fields] OR "Alcohol Drinking"[Mesh] OR "Alcohol-Related Disorders"[Mesh] OR "Alcoholism"[Mesh] OR "Alcohol-Induced Disorders"[Mesh])) NOT (animals[MeSH] NOT humans[MeSH])”. Also, studies not found through this search but included in the meta-analysis by de Martel and colleagues were included.[10] We also included the study by Hong and colleagues after the authors provided us with additional data on the overlap in risk factors.[11]

Studies were included if the study population was representative of liver cancer for the respective location. For each study, the proportions of liver cancer due to the five specific risk factors were calculated. Cases were considered to be due to NASH when the manuscript explicitly listed the aetiology to be NASH or non-alcoholic fatty liver disease (NAFLD). Cases where the aetiology was listed as “cryptogenic”, “idiopathic”, or “unknown” were included within the “other causes” category. In

manuscripts where the aetiology for a case was not known but major categories could not be ruled out (for example, the study tested for hepatitis B and C, but did not assess alcohol use), these cases were excluded from the numerator of the study (in other words, did not contribute to the proportion of any aetiology). Remaining risk factors were included under a combined "other" group (for example, haemochromatosis, autoimmune hepatitis, Wilson's disease, etc.). If multiple risk factors were reported for an individual patient, these were apportioned proportionally to the individual risk factors. These estimated proportions are then used to split the overall liver cancer estimates into estimates for their respective aetiologies.

Methods

Steps of analysis and data transformation processes

Cancer registry data went through multiple processing steps before integration with the COD database. First, the original data were transformed into standardised files, which included standardisation of format, categorisation, and registry names (#1 in flowchart).

Second, some cancer registries report individual codes as well as aggregated totals (eg, C18, C19, and C20 are reported individually, but the aggregated group of C18-C20 [colorectal cancer] is also reported in the registry data). The data-processing step "subtotal recalculation" (#2 in flowchart) verifies these totals and subtracts the values of any individual codes from the aggregates.

In the third step (#3 in the flowchart), cancer registry incidence data and cancer registry mortality data are mapped to GBD causes. A different map is used for incidence data and for mortality data because of the assumption that there are no deaths for certain cancers. One example is basal-cell carcinoma of the skin. In the cancer registry incidence data, basal-cell carcinoma is mapped to "non-melanoma skin cancer (basal-cell carcinoma)". However, if basal-cell skin cancer is recorded in the cancer registry mortality data, the deaths are instead mapped to "non-melanoma skin cancer (squamous-cell carcinoma)" under the assumption that they were indeed squamous-cell skin cancers that had been misclassified as basal-cell skin cancers. Other examples are benign or in situ neoplasms. Benign or in situ neoplasms found in the cancer registry incidence dataset were simply dropped from that dataset. The same neoplasms reported in a cancer registry mortality dataset were mapped to the respective invasive cancer (eg, melanoma in situ in the cancer registry incidence dataset was dropped from the dataset; melanoma in situ in the cancer registry mortality dataset was mapped to melanoma).

In the fourth data-processing step (#4 in the flowchart) cancer registry data were standardised to the GBD age groups. Age-specific incidence rates were generated using all datasets that include microdata, and datasets that report age groups up to 95+ years of age, while age-specific mortality rates were generated from the CoD data through a method described in Appendix section 2.5. Age-specific proportions were then generated by applying the age-specific rates to a given registry population that required age-splitting to produce the expected number of cases/deaths for that registry by age. The expected number of cases/deaths for each sex, age, and cancer were then normalised to 1, creating final, age-specific proportions. These proportions were then applied to the total number of cases/deaths by sex and cancer to get the age-specific number of cases/deaths.

In the rare case that the cancer registry only contained data for both sexes combined, the now-age-specific cases/deaths were split and reassigned to separate sexes using the same weights that are used for the age-splitting process. Starting from the expected number of deaths, proportions were generated by sex for each age (eg, if for ages 15 to 19 years old there are six expected deaths for males and four expected deaths for females, then 60% of the combined-sex deaths for ages 15-19 years would be assigned to males and the remaining 40% would be assigned to females).

In the fifth step (#5 in the flowchart) data for cause entries that are aggregates of GBD causes were redistributed. Examples of these aggregated causes include some registries reporting ICD10 codes C00-C14 together as, "lip, oral cavity, and pharyngeal cancer." These groups were broken down into sub-causes that could be mapped to single GBD causes. In this example, those include lip and oral cavity cancer (C00-C08), nasopharyngeal cancer (C11), cancer of other parts of the pharynx (C09-C10, C12-C13), and "Malignant neoplasm of other and ill-defined sites in the lip, oral cavity, and pharynx" (C14). To redistribute the data, weights were created using the same "rate-applied-to-population" method employed in age-sex splitting (see step four above). For the undefined code (C14 in the example) an "average all cancer" weight was used, which was generated by adding all cases from SEER/NORDCAN/CI5 and dividing the total by the combined population. Then, proportions were generated by sub-cause for each aggregate cause as in the sex-splitting example above (see step four). The total number of cases from the aggregated group (C00-C14) was then recalculated for each subgroup and the undefined code (C14). C14 was then redistributed as a "garbage code" in step six. Distinct proportions were used for C44 (non-melanoma skin cancer) and C46 (Kaposi's sarcoma). Non-melanoma skin cancer processing is described under section "Input data and methodological summary for non-melanoma skin cancer (squamous-cell carcinoma)." C46 entries were redistributed as "other cancer" and HIV using proportions described in Appendix Section 2.

In the sixth step (#6 in the flowchart) unspecified codes ("garbage codes") were redistributed. Redistribution of cancer registry incidence and mortality data mirrored the process of the redistribution used in the cause of death database (Appendix Section 2.7).

In the seventh step (#7 in the flowchart) duplicate or redundant sources were removed from the processed cancer registry dataset. Duplicate sources were present if, for example, the cancer registry was part of the CI5 database but we also had data from the registry directly. Redundancies occurred and were removed as described in "Inclusion and Exclusion Criteria," where more detailed data were available, or when national registry data could replace regionally representative data. From here, two parallel selection processes were run to generate input data for the MI models and to generate incidence for final mortality estimation. When creating the final incidence input, higher priority was given to registry data from the most standardised source; whereas for the MI model input, only sources that reported both incidence and mortality were used.

In the eighth step (#8 in the flowchart) the processed incidence and mortality data from cancer registries were matched by cancer, age, sex, year, and location to generate MI ratios. These MI ratios were used as input for a three-step modelling approach using ST-GPR, with HAQ Index as a covariate in the linear step mixed effects model using a logit link function. Predictions were made without the random effects. The ST-GPR model has three main hyper-parameters that control for smoothing across time, age, and geography, which were adjusted for GBD 2019. The time adjustment parameter lambda

(λ) aims to borrow strength from neighbouring time points (ie, the exposure in this year is highly correlated with exposure in the previous year but less so further back in time). Lambda was lowered from 2 to 0.05, reducing the weight of more distant years. The age adjustment parameter omega (ω) borrows strength from data in neighbouring age groups and was set to 0.5 (unchanged). The space adjustment parameter zeta (ξ) aims to borrow strength across the hierarchy of geographical locations.[12] Zeta was lowered from 0.95 to 0.01, reducing the weight of more distant geographical data. For the remaining parameters in the Gaussian process regression, we lowered the amplitude from 2 to 1 (reducing fluctuation from the mean function) and reduced the scale value from 15 to 10 (reducing the time distance over which points are correlated). These model specification changes generally led to less smoothing of the data compared to GBD 2017 models.

Data-cleaning steps were similar as for GBD 2017. For each cancer, MI ratios from locations in HAQ quintiles 1-4 were dropped if they were below the median of MI ratios from locations in HAQ quintile 5. We also dropped MI ratios from locations in HAQ quintiles 1-4 if the MI ratios were above the third quartile + 1.5 * IQR (inter-quartile range). We dropped all MIR that were based on less than 15 (this was 25 in 2017) cases to avoid noise due to small numbers, except for mesothelioma and acute myeloid leukaemia, where we dropped MIR that were based on less than ten cases because of lower data availability for these two cancers. We also aggregated incidence and mortality to the youngest five-year age bin where SEER reported at least 50 cases from 1990 to 2015, to avoid unstable MIR predictions in young age groups on too few datapoints. The MIR in the minimum age-bin was used to backfill the MIR down to the lowest age group estimated for that cancer.

Since MI ratios can be above 1, especially in older age groups and cancers with low cure rates, we used the 95th percentile (by age group) of the cleaned dataset (detailed above) to cap the MIR input data. This "upper cap" was used to allow MIR over 1 but to constrain the MIR to a maximum level. To run the logit model, the input data were divided by the upper caps to get data from 0 to 1. Model predictions from ST-GPR were then rescaled back by multiplying them by the upper caps.
To constrain the MIRs at the lower end, we used the fifth percentile of the cancer and age-specific cleaned MIR input data to replace all model predictions with this lower cap.

Final MI ratios were matched with the cancer registry incidence dataset in the ninth step (#9 in the flowchart) to generate mortality estimates (Incidence * Mortality/Incidence = Mortality) (#10 in the flowchart). These mortality estimates are then smoothed by a Bayesian noise-reduction algorithm (to deal with problems with zero counts, as also applied to the VR and VA data) and uploaded into the COD database (#11 in the flowchart). Cancer-specific mortality modelling then followed the general CODEm process.

Liver cancer aetiology split models

The proportion data found through the systematic literature review were used as input for five separate DisMod-MR 2.1 models to determine the proportion of liver cancers due to the five subgroups for all locations, both sexes, all years, and all age groups (step #16 in the flowchart). For GBD 2019 we used MR-BRT to split sex-combined input data into sex-specific proportion data. For liver cancer due to hepatitis C and hepatitis B, a prior value of 0 was set between age 0 and 0.01. For liver cancer due to alcohol, a prior value of 0 was set for ages 0 to 5 years. For liver cancer due to hepatitis C, hepatitis C (IgG) seroprevalence was used as a covariate, forcing a positive relationship between the hepatitis C

seroprevalence covariate and the outcome of liver cancer due to hepatitis C proportion. For liver cancer due to hepatitis B, seroprevalence of HBsAg was used as a covariate as well as the population coverage of three-dose Hepatitis B vaccination, forcing a negative relationship between vaccination and the outcome of liver cancer due to hepatitis B proportion. For liver cancer due to alcohol, alcohol (litres per capita) was used as a covariate as well as a covariate for proportion of alcohol abstainers, forcing a negative relationship between the proportion of alcohol abstainers and the outcome of liver cancer due to alcohol proportion. For liver cancer due to NASH, NASH/NAFLD prevalence was used as a covariate as well as a covariate for obesity prevalence and mean body-mass index (BMI), forcing a positive relationship between these covariates and the outcome of liver cancer due to NASH proportion. All covariates used were modelled independently. To ensure consistency between cirrhosis and liver cancer estimates and to take advantage of the data for the respective other related cause (eg, liver cancer due to hepatitis C and the related cause cirrhosis due to hepatitis C), we generated covariates from the liver cancer proportion models that were subsequently used in separate cirrhosis aetiology proportion models. We then created covariates from the cirrhosis aetiology proportion models and used those in final liver cancer aetiology models.

Since the proportion models are run independently of each other, the final proportion models were scaled to sum to 100% within each age, sex, year, and location, by dividing each proportion by the sum of the five (step # 17). For the liver cancer subtype mortality estimates, we multiplied the parent cause "liver cancer" by the corresponding scaled proportions (step # 18). Single cause estimates were adjusted to fit into the separately modelled all-cause mortality envelope in the GBD-wide CoDCorrect process.

Results

Interpretation of results

Cancer mortality estimates for GBD 2019 can differ from the GBD 2017 results for multiple reasons. Updated cancer mortality data were added from vital registration system data, verbal autopsy studies, and cancer registry incidence data. Previously some deaths mapped to liver cancer contained deaths from liver metastases rather than primary liver cancer; for GBD 2019, these deaths were instead mapped as garbage codes and redistributed. The mortality-to-incidence ratio estimation was updated with lower case inclusion criteria and different model hyperparameters compared to GBD 2017, leading to more training data and less smoothing across time and geography. Covariates used in CODEm models were updated for GBD 2019. This included removing or replacing covariates that had been updated by other GBD teams (most of the dietary covariates), assigning a direction of association prior to all covariates (previously covariates such as income and Socio-demographic Index had been allowed to have agnostic direction priors), and changing the minimum age ranges for which the models estimated mortality. Compared to GBD 2017, large differences in the incidence and prevalence estimates for the benign and in-situ neoplasms is due to changes in how the clinical informatics data are processed for these causes. These data are now adjusted for HAQ Index and corrected for outpatient encounters, which should capture significantly more of these cases than before (since that relied on hospital admissions).

The other group producing country-level cancer mortality estimates is the International Agency for Research on Cancer (IARC) with their GLOBOCAN database. Significantly different methods between the GBD study and GLOBOCAN can lead to differences in results. Whereas estimates in GLOBOCAN are based on the assumption that there are "In theory, […] as many methods as countries,"[13] the cancer

estimation process for the GBD study follows a coherent, well-documented method for all cancers, which allows cross-validation of models as well as determination of uncertainty. Another major difference is the ability in the GBD study to adjust single cause estimates to the all-cause mortality, which is being determined independently. This also allows us to adjust individual causes of death to the all-cause mortality envelope, which permits us to correct for the underdiagnosis of cancer in countries with inadequate diagnostic resources. Redistribution of a fraction of undefined causes of death to certain cancers is another methodological advantage the GBD study has over GLOBOCAN, and estimates for cancer mortality can therefore differ substantially in countries with a large proportion of undefined causes of deaths in their vital registration data or a large proportion of undefined cancer cases in their cancer registry data.

Limitations

There are certain limitations to consider when interpreting the GBD mortality cancer estimates. First, even though every effort is made to include the most recently available data for each country, data-seeking resources are not limitless and new data cannot always be accessed as soon as they are made available. It is therefore possible that the GBD study does not include all available data sources for cancer incidence or cancer mortality. Second, different redistribution methods can potentially change the cancer estimates substantially if the data sources used for the estimated location contain a large number of undefined causes; however, neglecting to account for these undefined deaths would likely introduce an even greater bias in the disease estimates. Third, using mortality-to-incidence ratios to transform cancer registry incidence data to mortality estimates requires accurate MIR. For GBD 2019 we have made further changes to the MIR estimation, but the method remains sensitive to underdiagnosis of cancer cases or under-ascertainment of cancer deaths. However, given that the majority of data used for the cancer mortality estimation come from vital registration data and not cancer registry data, this is not a major limitation.

Non-melanoma skin cancer (squamous-cell carcinoma)

Data

Data-seeking processes

Since squamous-cell carcinomas are only very infrequently recorded by cancer registries, only vital registration system data were used as input for the squamous-cell carcinoma mortality modelling.

Inclusion and exclusion criteria

Inclusion and exclusion criteria followed the same methods as described for the vital registration data sources (Appendix Section 2).

Bias of categories of input data

The potential biases of the input data are the same as for other cancers (see above).

Methods

Overall methodological process

Vital registration system data were used as input to model deaths due to squamous-cell skin cancer.

Steps of analysis and data transformation processes

Since mortality estimates for non-melanoma skin cancer are only produced for squamous-cell carcinoma

under the assumption that basal-cell carcinoma causes almost no deaths, all mortalities reported as "C44" or "173" were mapped to the "squamous-cell carcinoma" GBD cause.

Model selection

The modelling strategy for non-melanoma skin cancer (squamous-cell carcinoma) followed the general CODEm process.

Model performance and sensitivity

The modelling performance and sensitivity for non-melanoma skin cancer (squamous-cell carcinoma) mirrored that of the general CODEm process.

Uncertainty intervals

Uncertainty was determined using standard CODEm methodology.

Results

Interpretation of results

Non-melanoma skin cancer mortality estimates are not available from other sources. GLOBOCAN, for example, does not report deaths due to non-melanoma skin cancer. Even though the data availability for non-melanoma skin cancer is poor, the fact that it is the most common incident cancer, with rates expected to rise, makes it a necessity to include the disease in the GBD framework.

Limitations

Cancer registry data for non-melanoma skin cancer incidence have to be interpreted with caution due to a substantial amount of underreporting or rules that only the first non-melanoma skin cancer has to be registered. Many cancer registries therefore do not include non-melanoma skin cancers at all. However, the information if registries capture NMSC or not is not consistently available. Therefore, no cancer registry data were used to estimate deaths due to squamous-cell carcinoma of the skin. For vital registration data, we make the assumption that there are no deaths due to basal-cell non-melanoma skin cancer, and therefore all deaths attributed to basal-cell carcinoma were included instead as squamous-cell carcinoma.

Covariates by cancer:

Lip and oral cavity cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Tobacco (cigarettes per capita)	+
	Log-transformed SEV scalar: Mouth Cancer	+
2	*Age- and sex-specific SEV for high red meat*	+
	Age- and sex-specific SEV for low vegetables	+
	Age- and sex-specific SEV for low fruit	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Nasopharynx cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Tobacco (cigarettes per capita)	+
	Log-transformed SEV scalar: Nasopharynx Cancer	+
2	*Age- and sex-specific SEV for low vegetables*	+
	Population density (over 1000 ppl/sqkm, proportion)	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	Age- and sex-specific SEV for low fruit	+
	LDI (I$ per capita)	-
	Socio-demographic Index	+

Oesophageal cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Log-transformed age-standardised SEV scalar: Oesophageal Cancer	+
	Mean BMI	+
	Smoking prevalence	+
2	*Indoor air pollution (all cooking fuels)*	+
	Tobacco (cigarettes per captia)	+
	Age- and sex-specific SEV for low vegetables	+
	Age- and sex-specific SEV for low fruit	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	Sanitation (proportion with access)	–
	Improved water source (proportion with access)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Other pharynx cancer

Level	Covariate	Direction
1	Litres of alcohol consumed per capita	+
	Smoking prevalence	+
	Log-transformed SEV scalar: Other Pharynx Cancer	+
2	Cumulative cigarettes (5 years)	+
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low vegetables	+
	Population density (over 1000 ppl/sqkm, proportion)	+
	Population density (under 150 ppl/sqkm, proportion)	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Stomach cancer

Level	Covariate	Direction
1	Diet high in sodium	+
	Tobacco (cigarettes per capita)	+
	Log-transformed SEV scalar: Stomach Cancer	+
	Log-transformed SEV scalar: Stomach Cancer	+
2	Cumulative cigarettes (20 years)	+
	Age- and sex-specific SEV for unsafe water	+
	Age- and sex-specific SEV for unsafe sanitation	+
	Mean BMI	+
	Sanitation (proportion with access)	–
	Improved water source (proportion with access)	–
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	Age- and sex-specific SEV for low fruits	+
	Age- and sex-specific SEV for low vegetables	+
	LDI (I$ per capita)	+
	Socio-demographic Index	–

Testicular cancer

Level	Covariate	Direction
2	Cumulative cigarettes (5 years)	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (15 years)	+
	Cumulative cigarettes (20 years)	+
	Tobacco (cigarettes per capita)	+
	Smoking prevalence	+
	Age- and sex-specific SEV for low fruits	+
	Age- and sex-specific SEV for low vegetables	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Liver Cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	HIV age-standardised prevalence	+
	Hepatitis B seroprevalence (HBsAg) age-standardised	+
	Hepatitis C seroprevalence (anti-HCV) age-standardised	+
	Log-transformed SEV scalar: Liver Cancer	+
2	*Hepatitis B 3-dose coverage (proportion)*	–
	Hepatitis B vaccine coverage (proportion), aged through time	–
	Intravenous drug use (age-standardised proportion)	+
	Cumulative cigarettes (20 years)	+
	Mean BMI	+
	Tobacco (cigarettes per capita)	+
	Healthcare Access and Quality Index	–
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+

Liver cancer (continued)

Level	Covariate	Direction
3	*Education (years per capita)*	–
	Age- and sex-specific SEV for high red meat	+
	LDI (I$ per capita)	–
	Socio-demographic Index	–

Gallbladder and biliary tract cancer

Level	Covariate	Direction
1	*Log-transformed SEV scalar: Gallbladder Cancer*	+
	Mean BMI	+
2	*Litres of alcohol consumed per capita*	+
	Cumulative cigarettes (5 years)	+
	Cumulative cigarettes (10 years)	+
	Smoking prevalence	+
	Tobacco (cigarettes per capita)	+
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low vegetables	+
	Diabetes age-atandardised prevalence (proportion)	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	-

Pancreatic cancer

Level	Covariate	Direction
1	*Cumulative cigarettes (10 years)*	+
	Cumulative cigarettes (20 years)	+
	Tobacco (cigarettes per capita)	+
	Log-transformed SEV scalar: Pancreas Cancer	+
	Mean BMI	+
2	*Age- and sex-specific SEV for high red meat*	+
	Litres of alcohol consumed per capita	+
	Age- and sex-specific SEV for low vegetables	+
	Energy unadjusted (kcal)	+
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+
	Diabetes age-standardised prevalence (proportion)	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	Age- and sex-specific SEV for low fruit	+
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Larynx cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Log-transformed SEV scalar: Larynx Cancer	+
2	*Smoking prevalence*	+
	Asbestos consumption (metric tons per year per capita)	+
	Age- and sex-specific SEV for low vegetables	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Population density (over 1000 ppl/sqkm, proportion)	+
	Healthcare Access and Quality Index	–
3	*Age- and sex-specific SEV for low fruit*	+
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Tracheal, bronchus, and lung cancer

Level	Covariate	Direction
1	*Asbestos consumption (metric tons per year per capita)*	+
	Smoking prevalence	+
	Secondhand smoke	+
	Log-transformed SEV scalar: Lung Cancer	+
	Log-transformed age-standardised SEV scalar: Lung Cancer	+
2	*Indoor air pollution (all cooking fuels)*	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Outdoor air pollution ($PM_{2.5}$)	+
	Residential radon	+
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Malignant skin melanoma

Level	Covariate	Direction
1	Litres of alcohol consumed per capita	+
2	Latitude under 15 (proportion)	−
	Latitude 15 to 30 (proportion)	−
	Latitude 30 to 45 (proportion)	−
	Latitude over 45 (proportion)	−
	Healthcare Access and Quality Index	−
3	Education (years per capita)	−
	LDI (I$ per capita)	−
	Socio-demographic Index	+

Non-melanoma skin cancer

Level	Covariate	Direction
1	Cumulative cigarettes (5 years)	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (15 years)	+
	Smoking prevalence	+
2	Average latitude	−
	Healthcare Access and Quality Index	−
3	Education (years per capita)	−
	LDI (I$ per capita)	−
	Socio-demographic Index	+

Breast cancer

Level	Covariate	Direction
1	Litres of alcohol consumed per capita	+
	Mean BMI	+
	Log-transformed SEV scalar: Breast Cancer	+
2	Age-specific fertility rate	−
	Total fertility rate	−
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low vegetables	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Smoking prevalence	+
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+
	Healthcare Access and Quality Index	−
3	LDI (I$ per capita)	−
	Socio-demographic Index	+

Cervical cancer

Level	Covariate	Direction
1	Cumulative cigarettes (5 years)	+
	HIV age-standardised prevalence	+
2	Age-specific fertility rate	+
	Total fertility rate	+
	Smoking prevalence	+
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low vegetables	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	–
	Socio-demographic Index	–

Uterine cancer

Level	Covariate	Direction
1	Log-transformed SEV scalar: Uterus Cancer	+
	Mean BMI	+
2	Cumulative cigarettes (5 years)	+
	Cumulative cigarettes (10 years)	+
	Smoking prevalence	+
	Tobacco (cigarettes per capita)	+
	Diabetes age-standardized prevalence (proportion)	+
	Total fertility rate	–
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low vegetables	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Prostate cancer

Level	Covariate	Direction
1	Log-transformed SEV scalar: Prostate Cancer	+
2	Smoking prevalence	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	–
	Socio-demographic Index	+

Kidney cancer

Level	Covariate	Direction
1	*Cumulative cigarettes (5 years)*	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (15 years)	+
	Mean BMI	+
	Log-transformed SEV scalar: Kidney Cancer	+
2	*Litres of alcohol consumed per capita*	+
	Diabetes age-standardised prevalence (proportion)	+
	Systolic blood pressure (mmHg)	+
	Smoking prevalence	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Bladder cancer

Level	Covariate	Direction
1	*Schistosomiasis prevalence (proportion)*	+
	Cumulative cigarettes (10 years)	+
	Smoking prevalence	+
	Log-transformed SEV scalar: Bladder Cancer	+
2	*Litres of alcohol consumed per capita*	+
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+
	Age- and sex-specific SEV for low vegetables	+
	Healthcare Access and Quality Index	–
3	*Age- and sex-specific SEV for low fruits*	+
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Brain and nervous system cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Cumulative cigarettes (10 years)	+
	Smoking prevalence	+
2	*Cholesterol (total, mean per capita)*	+
	Systolic blood pressure (mmHg)	+
	Age- and sex-specific SEV for high red meat	+
	Age- and sex-specific SEV for low vegetables	+
	Age- and sex-specific SEV for low fruit	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Thyroid cancer

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Log-transformed SEV scalar: Thyroid Cancer	+
2	*Age- and sex-specific SEV for low vegetables*	+
	Age- and sex-specific SEV for high red meat	+
	Tobacco (cigarettes per capita)	+
	Mean BMI	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	Sanitation (proportion with access)	–
	Improved water source (proportion with access)	–
	Age- and sex-specific SEV for low fruits	+
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Mesothelioma

Level	Covariate	Direction
1	*Asbestos consumption (metric tons per year per capita)*	+
	Cumulative cigarettes (5 years)	+
	Log-transformed SEV scalar: Mesothelioma	+
	Log-transformed age-standardized SEV scalar: Mesothelioma	+
	Smoking prevalence	+
2	*Gold production (binary)*	+
	Indoor air pollution (all cooking fuels)	+
	Population density (over 1000 ppl/sqkm, proportion)	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	–
	Socio-demographic Index	+

Hodgkin lymphoma

Level	Covariate	Direction
2	*Healthcare Access and Quality Index*	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	–
	Socio-demographic Index	–

Non-Hodgkin lymphoma

Level	Covariate	Direction
2	*Cumulative cigarettes (5 years)*	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (15 years)	+
	Cumulative cigarettes (20 years)	+
	Litres of alcohol consumed per capita	+
	Smoking prevalence	+
	Mean BMI	+
	Healthcare Access and Quality Index	–
3	*Total fertility rate*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Multiple myeloma

Level	Covariate	Direction
1	*Litres of alcohol consumed per capita*	+
	Smoking prevalence	+
	Tobacco (cigarettes per capita)	+
2	*Age- and sex-specific SEV for low vegetables*	+
	Age- and sex-specific SEV for low fruits	+
	Age- and sex-specific SEV for high red meat	+
	Mean BMI	+
	Sanitation (proportion with access)	–
	Improved water source (proportion with access)	–
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Leukaemia

Level	Covariate	Direction
1	*Log-transformed age-standardised SEV scalar: Leukaemia*	+
	Log-transformed SEV scalar: Leukaemia	+
2	*Litres of alcohol consumed per capita*	+
	Mean BMI	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Tobacco (cigarettes per capita)	+
	Healthcare Access and Quality Index	–
3	*Education (years per capita)*	–
	LDI (I$ per capita)	+
	Socio-demographic Index	–

Myelodysplastic, myeloproliferative, other haemopoietic neoplasms

Level	Covariate	Direction
1	Log-transformed age-standardised SEV scalar: Leukaemia	+
	Log-transformed SEV scalar: Leukaemia	+
2	Litres of alcohol consumed per capita	+
	Cumulative cigarettes (5 years)	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (15 years)	+
	Cumulative cigarettes (20 years)	+
	Smoking prevalence	+
	Tobacco (cigarettes per capita)	+
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Other malignant cancers

Level	Covariate	Direction
1	Smoking prevalence	+
	Tobacco (cigarettes per capita)	+
2	Age- and sex-specific SEV for low vegetables	+
	Age- and sex-specific SEV for low fruits	+
	Age- and sex-specific SEV for low nuts and seeds	+
	PUFA adjusted (percent)	–
	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Other neoplasms

Level	Covariate	Direction
2	Healthcare Access and Quality Index	–
3	Education (years per capita)	–
	LDI (I$ per capita)	+
	Socio-demographic Index	–

Colon and rectum cancer

Level	Covariate	Direction
1	Mean BMI	+
	Tobacco (cigarettes per capita)	+
	Total physical activity (MET-min/week), age-specific	–
	Log-transformed SEV scalar: Colorectal Cancer	+
	Age- and sex-specific SEV for high red meat	+
2	Litres of alcohol consumed per capita	+
	PUFA adjusted (percent)	–
	Age- and sex-specific SEV for low vegetables	+
	Age- and sex-specific SEV for low fibre	+
	Age- and sex-specific SEV for low calcium	+
	Cumulative cigarettes (5 years)	+
	Diabetes fasting plasma glucose (mmol/L), age-standardised 25+	+
3	Education (years per capita)	–
	Age- and sex-specific SEV for low milk	+
	Age- and sex-specific SEV for low fruit	+
	Age- and sex-specific SEV for low nuts and seeds	+
	Healthcare Access and Quality Index	–
	LDI (I$ per capita)	+
	Socio-demographic Index	+

Ovarian cancer

Level	Covariate	Direction
1	Litres of alcohol consumed per capita	+
	Cumulative cigarettes (10 years)	+
	Cumulative cigarettes (20 years)	+
	Contraception (modern) prevalence (proportion)	–
	Log-transformed SEV scalar: Ovary Cancer	+
2	Asbestos consumption (metric tons per year per capita)	+
	Smoking prevalence	+
	Total fertility rate	–
	Energy unadjusted (kcal)	+
	Mean BMI	+
	Diabetes age-standardized prevalence (proportion)	+
	Healthcare Access and Quality Index	`–
3	Education (years per capita)	–
	Age- and sex-specific SEV for low fruits	+
	Age- and sex-specific SEV for low vegetables	+
	LDI (I$ per capita)	–
	Socio-demographic Index	+

References

1 Waterhouse J, Muir C, Shanmugaratnam K, Powell J. Cancer Incidence in Five Continents IV. Lyon: IARC, 1982.

2 Curado M, Edwards B, Shin H, *et al.* Cancer Incidence in Five Continents IX. Lyon: IARC, 2007 http://www.iarc.fr/en/publications/pdfs-online/epi/sp160/CI5vol9-A.pdf.

3 Muir C, Mack T, Powell J, Whelan S. Cancer Incidence in Five Continents V. Lyon: IARC, 1987.

4 Parkin D, Muir C, Whelan S, Gao Y, Ferlay J, Powell J. Cancer Incidence in Five Continents VI. Lyon: IARC, 1992.

5 Parkin D, Whelan S, Ferlay J, Raymond L, Young J. Cancer Incidence in Five Continents VII. Lyon: IARC, 1997.

6 Parkin D, Whelan S, Ferlay J, Teppo L, Thomas D. Cancer Incidence in Five Continents VIII. Lyon: IARC, 2002.

7 Forman D, Bray F, Brewster D, *et al.* Cancer Incidence in Five Continents X. 2013. http://ci5.iarc.fr.

8 Engholm G, Ferlay J, Christensen N, *et al.* NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.3 Association of the Nordic Cancer Registries. Danish Cancer Society. 2016; published online Aug 7. http://www.ancr.nu.

9 Steliarova-Foucher E, O'Callaghan M, Ferlay J, Masuyer E, Forman D, Comber H, Bray F. European Cancer Observatory: Cancer Incidence, Mortality, Prevalence and Survival in Europe. Version 1.0 European Network of Cancer Registries, International Agency for Research on Cancer. 2012; published online Sept. http://eco.iarc.fr.

10 de Martel C, Maucort-Boulch D, Plummer M, Franceschi S. World-wide relative contribution of hepatitis B and C viruses in hepatocellular carcinoma. *Hepatology* 2015; **62**: 1190–200.

11 Hong TP, Gow P, Fink M, *et al.* Novel population-based study finding higher than reported hepatocellular carcinoma incidence suggests an updated approach is needed. *Hepatology* 2016; **63**: 1205–12.

12 GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* 2016; **388**: 1659–724.

13 International Agency for Research on Cancer, World Health Organization. GLOBOCAN estimated cancer incidence, mortality, and prevalence worldwide in 2012. Lyon, France: IARC, 2014 http://globocan.iarc.fr/Default.aspx (accessed April 19, 2016).

14 Karagas MR, Greenberg ER, Spencer SK, Stukel TA, Mott LA. Increase in incidence rates of basal cell and squamous cell skin cancer in New Hampshire, USA. New Hampshire Skin Cancer Study Group. *Int J Cancer* 1999; **81**: 555–9.