Appendix 1. GBD data collection, modeling/analysis, and dissemination

The Global Burden of Disease (GBD) study is a comprehensive research initiative aimed at systematically assessing the health status and disease burden of populations worldwide. An international network comprising over 11,500 collaborators from 164 countries and territories contributed to the generation of GBD metrics through data provision, review, and analysis. GBD data collection involves diverse sources, including epidemiological surveys, hospital records, vital registration systems, disease surveillance systems, and additional sources such as academic papers and policy reports (https://ghdx.healthdata.org/gbd-2021/sources). The data is standardized using the International Classification of Diseases (ICD) codes to ensure accuracy and comparability (https://ghdx.healthdata.org/record/ihme-data/gbd-2021-cause-icd-code-mappings).

Sophisticated modeling tools, such as DisMod-MR and Spatiotemporal Gaussian Process Regression (ST-GPR), are employed to estimate prevalence, incidence, and mortality rates. Data processing includes corrections for heterogeneity and biases, as well as uncertainty analysis through Monte Carlo simulations. Key health metrics used are Disability-Adjusted Life Years (DALYs), Years of Life Lost (YLLs), and Years Lived with Disability (YLDs) (https://www.healthdata.org/gbd/methods-appendices-2021/cancers). Dissemination of GBD findings is achieved through scientific publications (https://www.healthdata.org/research-analysis/gbd-publications), and interactive tools like GBD Compare and Viz Hub (https://www.healthdata.org/research-analysis/gbd-data). These tools facilitate the exploration and comparison of health data across regions and time periods. The primary goal of GBD findings is to provide a comprehensive framework for understanding global and local health trends, thereby supporting evidence-based health decision-making and resource allocation.

Table of Contents

List of Tables & Figures

Appendix Figures:

Appendix Tables:

Table 10: GBD 2021 Socio-Demographic Index R-squared values with lags up to 10 years

Table 11: GBD 2021 Socio-Demographic Index quintiles - or basically same as SDI values just listed in order of SDI value rather than by location groupings

Table 12: List of International Classification of Diseases (ICD) codes mapped to non-fatal causes and injuries in the GBD 2021

Section 1: GBD Overview

Section 1.1: Geographic locations of the analysis

We produced estimates for 204 countries and territories that were grouped into 21 regions and seven super regions (table 1). - The seven super-regions are central Europe, eastern Europe, and central Asia; high income; Latin America and the Caribbean; north Africa and the Middle East; south Asia; southeast Asia, east Asia, and Oceania; and sub-Saharan Africa. In GBD 2021 we continue to analyse at subnational levels countries that were added in previous cycles including Brazil, China, Ethiopia, India, Indonesia, Iran, Italy, Japan, Kenya, Mexico, New Zealand, Nigeria, Norway, Pakistan, Russia, the Philippines, Poland, South Africa, Sweden, the UK, and the USA. All analyses are at the first level of administrative organisation within each country except for New Zealand (by Māori ethnicity), Sweden (by Stockholm and non-Stockholm), the UK (by local government authorities), and the Philippines (by provinces). To meet data use requirements, in this publication we present subnational estimates for Brazil, India, Indonesia, Japan, Kenya, Mexico, Sweden, the UK, and the USA); given space constraints, these results are presented in Appendix 2 instead of the main text. Subnational estimates for China are included in maps but are not reported in appendix tables. Subnational estimates for other countries will be released in separate publications.

At the most detailed spatial resolution, we generated estimates for 983 unique locations. As was done in GBD 2019, in GBD 2021 we continue to use the set of locations defined as standard locations and non-standard locations. Standard GBD locations are defined as the set of all subnationals belonging to countries where data quality is high and with populations over 200 million, in addition to all other countries. Standard locations include the subnationals for China, India, the USA, and Brazil, but not Indonesia; data for China, India, the USA, and Brazil are also included at the country level. All other countries with subnational estimates are defined as non-standard locations.

Section 1.2: Time period of the analysis

We estimated numbers and rates of incidence, prevalence, years lived with disability (YLDs), and disability-adjusted life-years (DALYs) for the years 1990–2021; we estimated deaths and years of life lost (YLLs) for 1980–2021.

Section 1.3: GBD cause list

The GBD cause and sequelae list is organized hierarchically (see table 2) to accommodate different purposes and needs of various users.

The first two levels aggregate causes into general groupings. At Level 1 there are three cause groups: communicable, maternal, neonatal, and nutritional diseases (Group 1 diseases); non-communicable diseases (Group 2); and injuries (Group 3). These Level 1 aggregates are subdivided at Level 2 of the hierarchy into 22 cause groupings (eg, neonatal disorders, neurological disorders, and transport injuries). The disaggregation into Levels 3 and 4 contains the finest level of detail for causes captured in GBD 2021. The greatest detail available for some causes, such as anxiety disorders or rheumatoid arthritis, is at Level 3 of the hierarchy, while other specific causes are at Level 4 of the hierarchy with an aggregate category at Level 3 (for example, depressive disorders at Level 3, which encompasses major depressive disorders and dysthymia at Level 4). Sequelae of diseases and injuries are organised at Levels 5 and 6 of the hierarchy. In GBD, sequelae are defined as distinct, mutually exclusive categories of health consequences that can be directly attributed to a cause. For example, both neuropathy and blindness due to diabetic retinopathy are sequelae of diabetes; stroke and ischaemic heart disease are not, as these consequences cannot be categorically ascribed to diabetes in an individual despite good evidence for increased risk of these outcomes. The finest detail for all sequelae estimated in GBD is at Level 6 and is aggregated into summary sequelae categories (Level 5) for causes with large numbers of sequelae. Examples include the grouping of the infectious disease episodes and long-term sequelae of meningitis. For GBD 2021 there are 3499 mutually exclusive and collectively exhaustive sequela, 2089 cause sequelae and 1410 injuries sequelae, and thus our YLD estimates at each level of the hierarchy sum to the total of the level above. Prevalence and incidence aggregation is estimated at the level of individuals who may have more than one sequela or disease and therefore are not additive.

The GBD cause list continues to evolve to reflect the policy relevance, and public health and medical care importance of the causes of major losses of health. The cause and sequelae list expanded based on input from the Scientific Council and GBD collaborator network. For GBD 2021, the causes of death cause listhas increased to 288 causes, from the 286 causes in GBD 2019. The non-fatal cause list has expanded from 364 causes in GBD 2019 to 365 causes in GBD 2021. The total number of fatal and non-fatal causes combined for GBD 2021 is 371. As in GBD 2019, we made no estimates for YLDs for just five causes, either because no disability is possible (as is the case with sudden infant death syndrome); because disability may occur rarely but at levels too low for accurate estimation given the data (as for aortic aneurysm); or because the disability is captured by the complicating causes that led to that cause of death (as for indirect maternal deaths, late maternal deaths, and maternal deaths aggravated by HIV/AIDS).

Section 1.4: Statement of GATHER compliance

This study complies with GATHER recommendations. We have documented the steps in our analytical procedures and detailed the data sources used. See table 3for the GATHER checklist. The GATHER recommendations can be found at the GATHER website under GATHER Statement.

Section 1.5 GBD results overview

Results from GBD 2021 are available through an interactive data downloading tool on the Global Health Data Exchange (GHDx). The GHDx is the world's most comprehensive catalogue of surveys, censuses, vital statistics, and other health-related data. Results are measured in terabytes.

The latest version of the data download tool, available here: http://ghdx.healthdata.org/GBD-results-tool, contains core summary results for GBD 2021. These results include deaths, years of life lost (YLLs), YLDs, disability-adjusted life-years (DALYs), prevalence, incidence, and rate of change. The GHDx includes data for causes, risks, cause-risk attribution, aetiologies, and impairments.

Data above a certain size cannot be viewed online but can be downloaded. Depending on the size of the download, users may need to enter an email address; a download location will be sent to them when the files are prepared.

All GBD 2021 online data visualisations are available at http://www.healthdata.org/gbd/data-visualizations, which provides results for all GBD health metrics.

Section 1.6 Data input sources overview

GBD 2021 synthesises a large and growing number of data input sources including surveys, censuses, vital statistics, and other health-related data sources. The data from these sources are used to estimate morbidity; illness, and injury; and attributable risk for 204 countries and territories from 1990 to 2021; mortality deaths are estimated from 1980 to 2021. The input sources are accessible through an interactive citation tool available in the GHDx.

Citations for specific GBD components, causes and risks, and locations can be found through the Data Input Sources Tool in GHDx: http://ghdx.healthdata.org/gbd/2020/data-input-sources. This tool allows users to view and access GHDx records for input sources and export a comma-separated value (CSV) file that includes metadata, citations, and information about where the data were used in GBD. As required by GATHER, additional metadata for input sources are available through the citation tool as well.

Section 1.7 Funding sources

This publication and the research it presents was funded by the Bill & Melinda Gates Foundation; Queensland Department of Health, Australia; the National Health and Medical Research Council, Australia; Public Health England; the Norwegian Institute of Public Health; St. Jude Children's Research Hospital; the Cardiovascular Medical Research and Education Fund; the National Institute on Ageing of the National Institutes of Health (award P30AG047845); and the National Institute of Mental Health of the National Institutes of Health (award R01MH110163). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all data in the study and had final responsibility for the decision to submit for publication.

Section 1.8: Abbreviations

ARC – annualized rate of change ASFR- age-specific fertility rate
ACMR all-cause mortality rate

BMI Body Mass Index
CMNN Communicable, maternal, neonatal, and nutritional diseases
CoD causes of death
CODEm Cause of Death Ensemble modelling
COMO comorbidity correction
COPD Chronic obstructive pulmonary disease
CSMR cause-specific mortality rates
CV coefficient of variation
DALYs disability-adjusted life-years
DisMod-AT disease model-Bayesian age-time
DisMod-MR disease model-Bayesian meta-regression
DW – disability weight
EDU15+ education for those 15 years old and older
EMR excess mortality rate
GATHER Guidelines for Accurate and Transparent Health Estimates Reporting
GBD Global Burden of Diseases, Injuries, and Risk Factors Study
GHDx Global Health Data Exchange
GPR Gaussian process regression
HALE healthy life expectancy
HAT human African trypanosomiasis
ICD- International Classification of Diseases
ICG- ICD groups
IFD in-facility delivery proportion
IHME Institute for Health Metrics and Evaluation
LASSO least absolute shrinkage and selection operator
LDI lag-distributed income
LOESS locally estimated scatterplot smoothing
MAD median absolute deviation
MCCD Medical Certification of Causes of Death
MEPS Medical Expenditure Panel Survey
MICS Multiple Indicators Survey
MR-BRT Meta-regression—Bayesian, regularised, trimmed
NESARC National Epidemiologic Survey on Alcohol and Related Conditions
NSMHWB Australian National Survey of Mental Health and Wellbeing
NTDs – neglected tropical diseases
RSME root mean square error
SARS-CoV 2 Severe acute respiratory syndrome coronavirus 2
SD Standard deviation
SID HCUP State Inpatient Database
SDI Social Demographic Index
ST-GPR spatiotemporal Gaussian process regression
TFR total fertility rate
TFU25 younger than 25 years old (fertility rate)
UI uncertainty interval
UK United Kingdom
UI uncertainty interval
USA United States of America
WHO World Health Organization

YLDs years lived with disability
YLLs years of life lost

Section 2: Non-fatal outcome estimation

The GBD 2021 non-fatal estimation process describes the steps necessary to estimate incidence, prevalence, and YLDs for disease and injury sequelae in GBD 2021. Conceptually, the estimation effort is divided into eight major components: (1) compiling data sources through data identification and extraction; (2) data adjustments; (3) estimation of prevalence and incidence by cause and sequelae by using DisMod-MR 2.1, or alternative modelling strategies for select cause groups; (4) estimation by impairment; (5) severity distributions; (6) incorporation of disability weights (DWs); (7) comorbidity adjustment; and (8) the estimation of YLDs by sequelae and causes. Section 6 contains additional detail specific to each non-fatal disease, impairment, and injury, and their sequelae. Non-fatal modelling strategies vary significantly between causes.

Section 2.1: Data sources, identification, and extraction

Section 2.1.1: Systematic reviews

For GBD 2021, updated systematic reviews were conducted for 77- causes and risk factors. For other disease sequelae, only a small fraction of the existing data appears in the published literature, and other sources predominate, such as survey data, disease registers, notification data, or hospital inpatient data. As was done in past rounds of GBD, data were systematically screened from household surveys archived in the GHDx (http://ghdx.healthdata.org/), including Demographic and Health Surveys, Multiple Indicator Cluster Surveys (MICS), Living Standards Measurement Surveys, and Reproductive Health Surveys. Other national health surveys were identified on the basis of survey series that had yielded usable data for past rounds of GBD, sources suggested to us by in-country GBD collaborators, and surveys identified in major multinational survey data catalogues such as the International Household Survey Network and the WHO Central Data Catalog, as well as through country Ministry of Health and Central Statistical Office websites. Case notifications reported to the WHO were updated through 2020. Citations for all data sources used for non-fatal estimation in GBD 2021 are provided in searchable form through a web tool (http://ghdx.healthdata.org/). A description of the search terms used for cause-specific systematic reviews are detailed by cause in Section 6

Section 2.1.2: Survey data preparation

For GBD 2021, survey data for which we have access to the unit record data constitute a substantial part of the underlying data used in the estimation process. During extraction, we concentrated on demographic variables (eg, location, sex, age), survey design variables (eg,

sampling strategy and sampling weights), and the variables used to define the population estimate (eg, prevalence or a proportion) and a measure of uncertainty (standard error, confidence interval or sample size, and number of cases).

Section 2.1.3: Disease registries

For GBD 2021 non-fatal estimation, disease registries were an important source for a select number of conditions such as cancers, end-stage renal disease, and congenital disorders. Registry data is particularly key in the estimation of neoplasms when we consider the increasing attention to non-communicable diseases, particularly cancers, in low and middle-income areas of the world. The GHDx source tool (http://ghdx.healthdata.org/data-type/disease-registry) provides a comprehensive list of registry data used in GBD estimation processes.

Section 2.1.4: Case notifications

Case notifications, active screening, intervention coverage studies, and surveillance contributed to estimates of infectious diseases. If data were available, we extracted it from survey and administrative microdata; otherwise, data were extracted from published literature and reports. For many infectious diseases and neglected tropical diseases (NTDs), we used cases for which notification was made by countries to the WHO and other global monitoring entities. The causes for which we used WHO case notification data included tuberculosis, measles, yellow fever, rabies, dengue, cholera, whooping cough, human African trypanosomiasis (HAT), meningitis, all sexually transmitted infections, and other infectious diseases and NTDs, such as Ebola.

Section 2.2: Clinical input data and methods summary

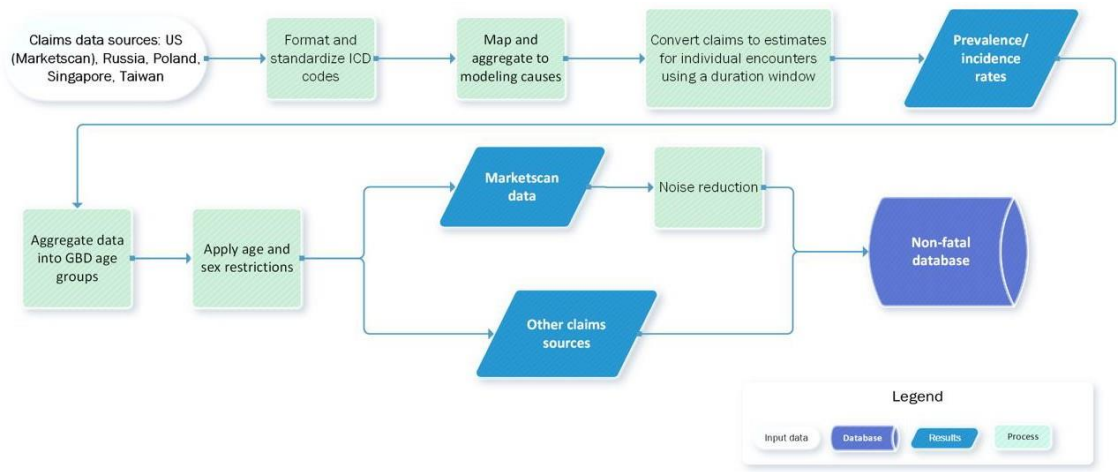
Administrative claims, inpatient hospital, and outpatient data played a key role in the process of estimating many non-fatal causes and injuries in GBD 2021. Data sources were heterogeneous in granularity, comprehensiveness, and level of detail, and the methods described below were used to transform data to be comparable and complete across locations, ages, sexes, years, and causes.

Section 2.2.1: Mapping diagnoses to GBD diseases and injuries

Most clinical sources are coded using the International Classification of Diseases (ICD) system that we map to GBD-defined diagnosis groups. ICD-9 and ICD-10 codes are mapped to what are termed “ICD code groups” (ICGs) with a many-to-one relationship, which simplifies the disease categorization and reduces complexity. ICGs are then mapped to a disease or injury modelling entity used by GBD modelers.

Some ICD codes are not mapped to a clinical modelling entity as some causes in the GBD cause hierarchy do not use clinical data sources. These ICD codes are still included in the sum of all admissions for that location. We also designate whether each modelling entity is processed in terms of incidence or prevalence, depending on the nature of the disease and the expected pattern of treatment. Table 12 shows the ICD codes used for non-fatal modelling by GBD cause and injury.

Figure 1. GBD 2021 Claims Data Processing



Marketscan claims

For GBD 2021, we accessed aggregate data derived from the Merative database of USA private health insurance and Medicare private supplemental insurance for the years 2000 and 2010-2017. The population covered in each year was 3.3 million in 2000, 40.4 million in 2010, 44.4 million in 2011, 40.8 million in 2012, 42.2 million in 2013, 36.4 million in 2014, 22.6 million in 2015, 22.4 million in 2016, and 20.8 million in 2017. For each of these individuals, claims representing every health service encounter were used and all episodes of care were linked to individuals by unique identifiers. For the GBD, we subset the population in the Marketscan database to individuals with a full year of insurance coverage or those who were born or died in the year of interest in order to ensure the sample includes all healthcare utilization for a given individual in that year.

We mapped ICD diagnoses in each source to GBD causes and injuries. GBD conditions are processed as “prevalence” or “incidence” based on the specification of the research team responsible for the cause. Prevalent conditions are identified as any primary or non-primary diagnosis on any inpatient or outpatient claim within the year of interest. To reduce noise from spurious coding practices, a minimum of two outpatient claims for the same individual are required in a calendar year to count as a prevalent case. Incidence of disease or injury was

calculated based on a duration window which varied by cause. Any individual who had multiple diagnoses for the same cause within the duration window are counted as a single incident case, and additional diagnoses outside of the duration window are treated as new incident cases.

After mapping to cause and identifying prevalent and incident cases by cause, we applied a noise reduction model to smooth trends over age and time.

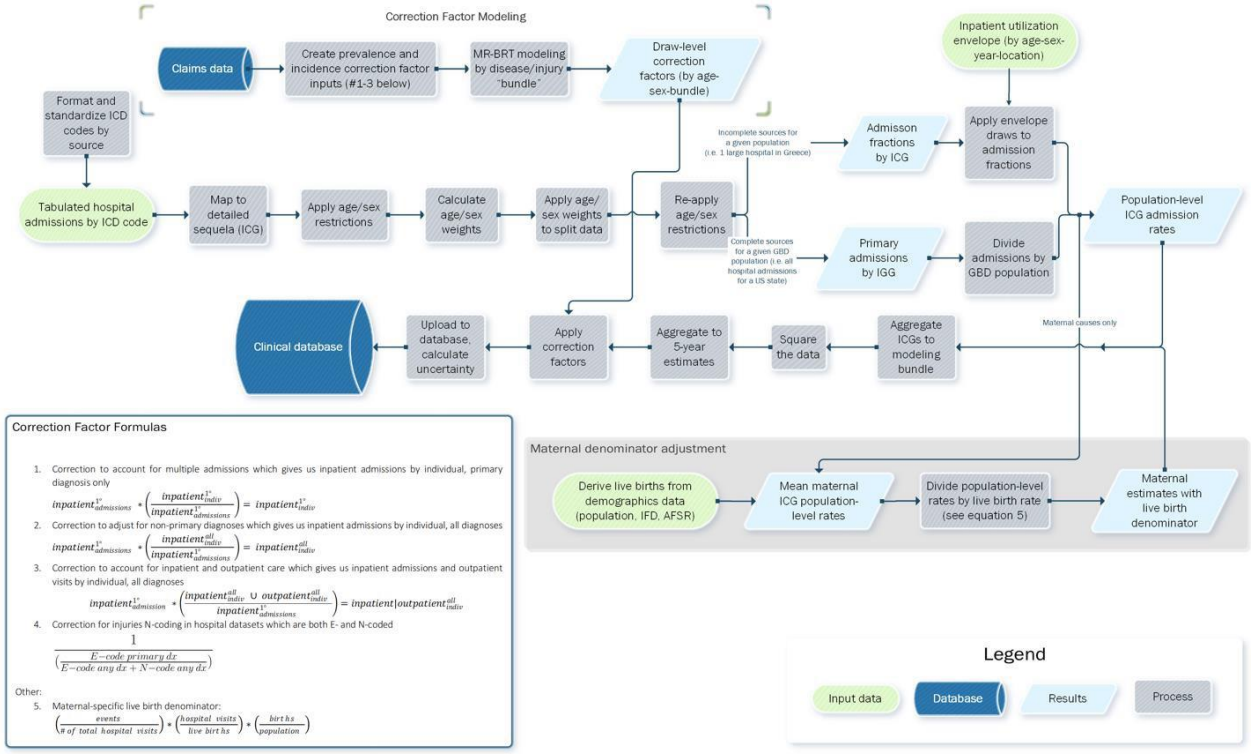
Other claims data

Claims data from Poland, Singapore, and Taiwan (province of China) were also processed for GBD 2021. Anonymized, individual-level claims data from Poland were accessed through an existing collaboration and institutional partnership with the Agency for Health Technology and Tariff System (AOTMiT). The data is derived from the National Health Fund (Narodowy Fundusz Zdrowia) database in Poland and is representative of every service encounter and episode of care in the public health care system (close to 92% population coverage) from 2015 to 2018.

Tabulated inpatient-only claims data from Singapore for the years 1991-2017 were derived from the MediClaim database and provided by the Ministry of Health of Singapore. The MediClaim data processed for the GBD is inclusive of all inpatient admissions in the country's public and private hospital facilities, and for all patients covered under MediShield Life, MediSafe, and MediFund, with admissions aggregated at national level. Similarly, Taiwan (province of China) claims for the year 2016, derived from the National Health Insurance Research Database (NHIRD) and covering all residents in Taiwan under a universal single-payer health care system, was used. The NHIRD is representative of the whole population for Taiwan and covers both inpatient admissions and outpatient encounters.

Section 2.2.3: Inpatient hospital admissions

Figure 2. GBD 2021 Inpatient Hospital Data Processing



Age-sex splitting and processing individual-level data

Inpatient hospital data were extracted from 4,722 location-years in 49 countries. ICD coding was standardized across sources and versions of ICD. Counts of admissions with a primary diagnosis of each cause were extracted from all sources and stepped through the inpatient data processing methods. For inpatient data, a case of disease was defined as an overnight inpatient admission with a primary diagnosis of that cause. We tabulated the incident or prevalent admissions for each source according to the disease or injury. Admissions were then aggregated to create cause fractions, defined as the number of admissions for a given disease/injury divided by total admissions for that age, sex, and year. Secondary diagnostic detail was included in estimation through corrections as described below.

In GBD 2021, 13 inpatient sources with high percentages of live birth diagnosis codes (i.e. Z37.0) in the 0-6 day age group were either removed or swapped from the primary diagnosis position for the subsequent diagnosis in sources with multiple diagnoses.

Deriving population-level estimation

Section 2.2.5 of the appendix describes the modelling process for the inpatient utilisation envelope, an estimate of inpatient admissions per capita for all GBD locations, years, ages, and both sexes. Inpatient sources were assessed for whether or not they capture a complete and representative GBD population, meaning that we would expect all hospital admissions for a given location and year to be present in the data source. Sources that meet this criteria did not

use the inpatient utilisation envelope to derive population-level estimations and used GBD population estimates instead. Changes from GBD 2019 to GBD 2021 in the estimation of the inpatient utilisation envelope are outlined in the dedicated section below.

Corrections

We performed three adjustments on inpatient hospital data to synthesize all inpatient sources to the same definition of care and to account for cases that were not captured in inpatient sources. Data were first adjusted to account for multiple admissions for a single case of disease and then adjusted to account for cases of any disease that were non-primary diagnoses recorded for an admission. Finally, admissions were scaled by the ratio of outpatient cases observed for any inpatient case of disease to account for additional cases that did not warrant an inpatient admission. Combined with the uncorrected incidence and prevalence rates from the inpatient sources (with no scalar applied), this process resulted in four versions of inpatient estimates: (1) un-corrected inpatient admissions by episode, primary diagnosis; (2) inpatient admissions by individual, primary diagnosis only; (3) inpatient hospital admissions, accounting for all diagnoses; and (4) an estimate of inpatient admissions and outpatient visits by individual, accounting for all diagnoses. Estimate 4 was applied to all causes except those where outpatient care or non-primary diagnosis were not used in the modeling strategy given the nature of the disease. Adjustment ratios were calculated using all clinical sources that had patient-level data and primary and non-primary diagnoses.

Sources of this data include Marketscan and Taiwan claims data as described above; claims and inpatient data from Poland, the Philippines, New Zealand, and the HCUP State Inpatient Database (SID). Only Marketscan, Poland, and Taiwan claims data included a link between inpatient and outpatient care to be used in the fourth estimate described. Ratios from these sources were modelled over age and sex using a mixed-effects model in MR-BRT for each cause. If data for any ratio did not exist for the youngest or oldest age groups, we assumed a uniform tail on the model from the nearest age group with data. All models were conducted in log-space in order to bound the model to be greater than one for any age, sex, and cause. We used the following equations for each of the three scalars:

1) Correction to account for multiple admissions, which gives us inpatient admissions by individual, primary diagnosis only
    a. $inpatient^{1°}_{admin} * (\frac{inpatient^{1°}_{indiv}}{inpatient^{1°}_{admin}}) = inpatient^{1°}_{indiv}$
2) Correction to adjust for non-primary diagnoses, which gives us inpatient admissions by individual, all diagnoses
    a. $inpatient^{1°}_{admin} * (\frac{inpatient^{all}_{indiv}}{inpatient^{1°}_{admin}}) = inpatient^{all}_{indiv}$

3) Correction to account for inpatient and outpatient care, which gives us inpatient admissions and outpatient visits by individual for all diagnoses
   a. $inpatient^{1°}_{admission} * (\frac{inpatient^{all}_{indiv} \cup outpatient^{all}_{indiv}}{inpatient^{1°}_{admissions}}) = inpatient|outpatient^{all}_{indiv}$

Denominators for maternal conditions were adjusted using in-facility delivery proportion (IFD) and age-specific fertility rate (ASFR) covariates to include only those at risk for maternal conditions. After this adjustment, the denominator represents people who gave birth in that year.

Inpatient sources that use the inpatient utilisation envelope:

$$\frac{condition\ specific\ admissions}{all\ cause\ admissions}) * (\frac{estimated\ envelope\ admissions}{GBD\ pop * (IFD * ASFR)})$$

Inpatient sources that do not use the inpatient utilization envelope:

$$\frac{condition\ specific\ admissions}{(GBD\ pop * birth\ adjustment)}$$

Clinical estimates for injuries use a separate correction factor from those described above, which adjusts sources with an insufficient proportion of ecodes (external causes of injury) among all injuries ICD codes. We aggregate total ecodes, in the primary Dx position, by source-GBD location-year, and divide by ecodes and ncodes (nature of injury codes), in any diagnosis position, for the same demographic. Source-GBD location-years that have a proportion less than .15 are dropped. For example, Japan-Yamanashi-2010 has a proportion of .018, interpreted as 1.8% of injuries codes in that demographic are ecodes, and would be removed.

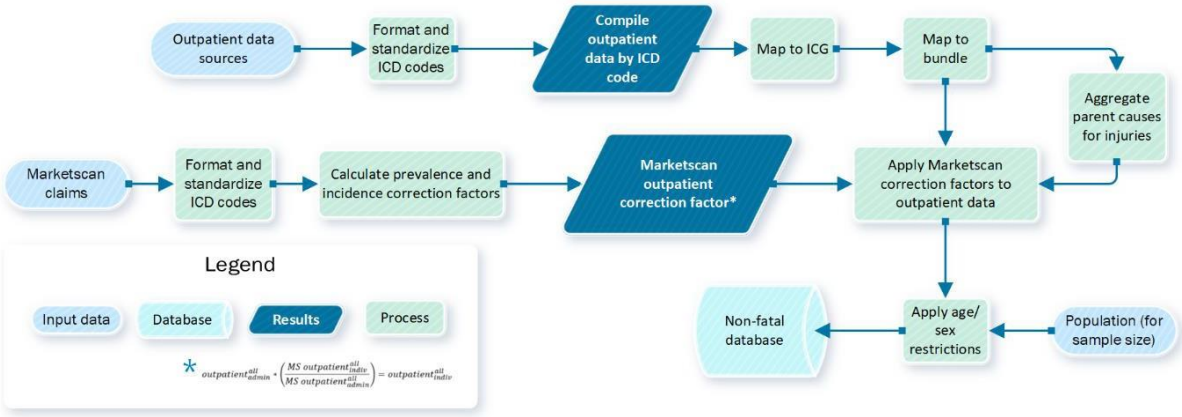
The injuries correction is created directly from this proportion:

$$\frac{1}{ecode\ primary\ Dx\ admissions/(ecode+ncode\ codes,all\ diagnoses)}$$

A final adjustment was applied to each of the above estimates. The HAQ Index was used to account for differences in access and quality of health care across time and space. The HAQ Index adjustment was applied by dividing the above estimates by a scalar ranging from 0 to 100, where 0 represents the first percentile of observed access and quality and 100 the 99th percentile.

Section 2.2.4: Outpatient encounter data

Figure 3. GBD 2021 Outpatient data extraction process



Outpatient encounter data, that could not be linked to inpatient admissions, were processed from the USA and Sweden for 109 location-years. No changes were made in the processing of outpatient data from GBD 2019, except for updates to the ICD mapping.

As with the inpatient hospital data, a scalar was calculated by using Marketscan outpatient claims data to adjust for multiple visits per individual within one year (for prevalent conditions) and within a cause-specific duration (for incident causes).

Calculating uncertainty

Uncertainty in claims estimates was calculated using Wilson's approximation, utilizing sample size derived from enrollment data (i.e. Marketscan) or GBD population estimates (i.e. Poland), depending on the source. Uncertainty in outpatient estimates was also calculated using Wilson's approximation and GBD population. Uncertainty for inpatient sources that are not complete for the population and use the inpatient utilization envelope came from the upper and lower uncertainty intervals of 1000 bootstrapped samples of the envelope and correction factor models. Inpatient sources that are complete for the population derived uncertainty from Wilson's approximation and GBD population.

Wilson's approximation:

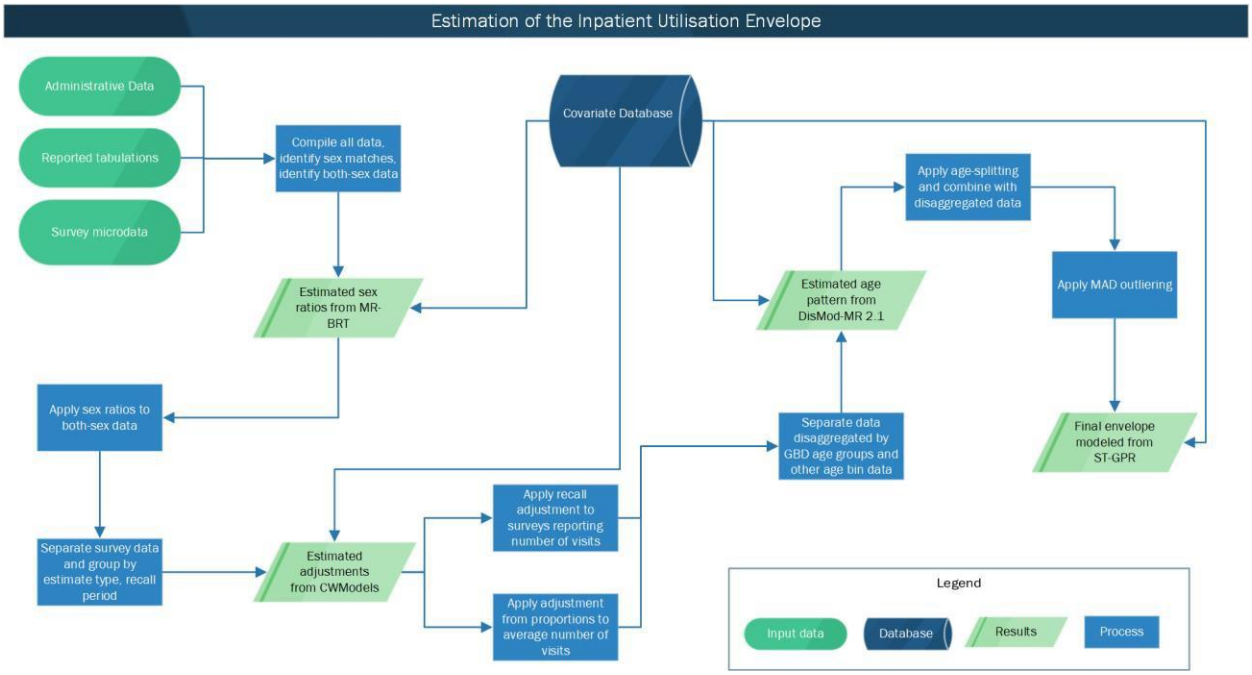
$$\sigma^2 = \frac{cf(1-cf)}{n} + \frac{1.96^2}{4n^2} \Big/ \left(1 + \frac{1.96^2}{n}\right)^2$$

where cf is the cause fraction and n is the sample size.

Section 2.2.5: Estimation of the inpatient utilization envelope

This process utilises administrative data, reported tabulations, and survey microdata to estimate the rates of inpatient admissions per capita for every location and demographic group in the GBD hierarchy.

Figure 4. Overview process of estimation of hospital envelope



Case definition

We defined a hospital admission as admission into a formal health care facility for, at least, an overnight stay. However, we excluded admissions to long-term care facilities (>120 days), nursing care facilities, and facilities staffed by traditional or spiritual healers.

Input data

We searched the GHDx for population surveys, administrative records, and censuses from January 1990 to September 2019. We applied the following keyword filters: “Health care use” OR “Length of stay” AND “Hospitals” OR “Health care services”. We applied no language restrictions to our search and required all returned records to contain either microdata or tabulated reports. We searched the returned records’ metadata for measures of inpatient care. For inclusion, we required all measures to be nationally or subnationally representative. Additionally, we consulted with experts and GBD collaborators to gather data sources that were not within the GHDx. We included 2064 sources for GBD 2021, adding 400 new sources relative to GBD 2019.

Data processing

From data sources for which microdata were available, we extracted and binned the data based on gender and age groups of 0-11 months, 12-23 months, 2 to 4 years, 5-9 years, 10-14 years, and similar increments of years up to 95 years and older. Data was occasionally binned into wider age groups where less detailed age data was available, or where samples were sufficiently small.

Our input data contained a limited number of both-sex data points. We used the MR-BRT modelling tool (see Section 2.5 for details on MR-BRT) to model the ratio of female to male admissions based on

matched sex-specific data. The results of this model were used to split both-sex data points into sex-specific data. The estimated adjustment factor from the MR-BRT analysis is presented below. This factor can be interpreted as the observed ratio between female and male utilisation.

Table 4. Sex-splitting Adjustment Factor

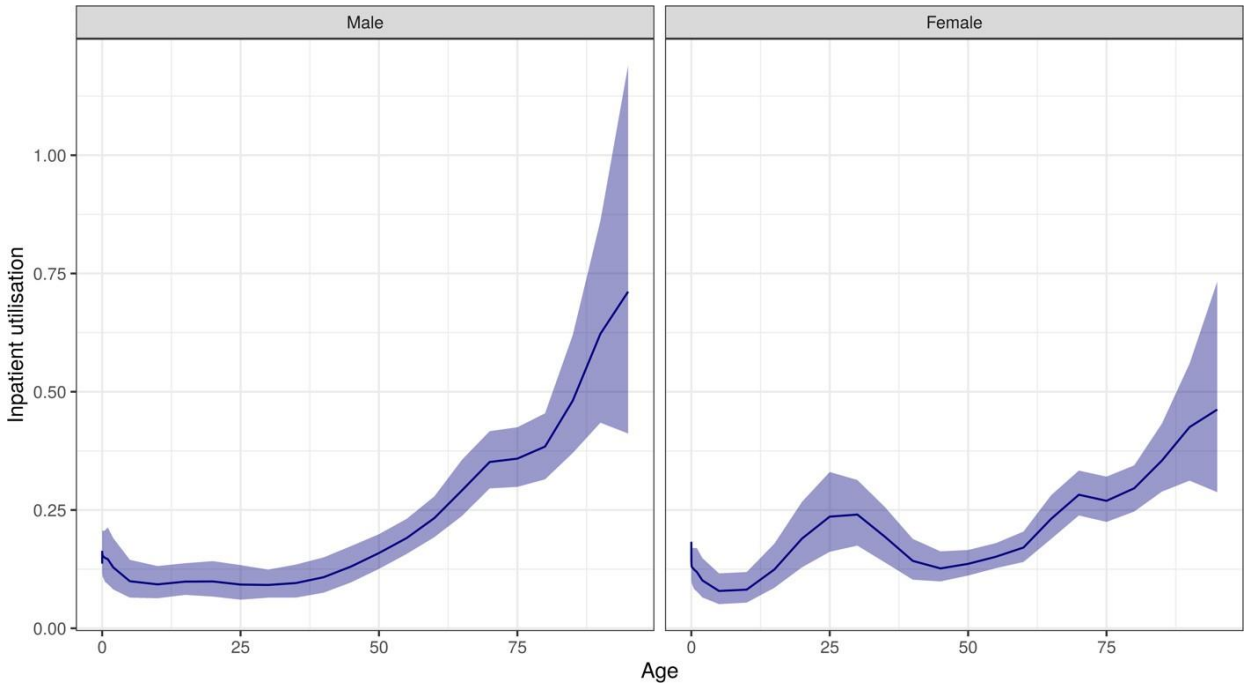
Data input	Beta coefficient, log (95% UI)	Adjustment factor
Sex	-0.056 (-0.559, 0.449)	0.946

We classified each of the accepted data sources into four data types: (1) proportion of survey respondents who were admitted into the hospital in the last 30 days; (2) proportion of survey respondents who were admitted to the hospital in the last year; (3) average number of admissions (utilisation rate) reported by survey respondents in the last year; and (4) average number of visits reported by annual administrative records. We assigned measures reported by annual administrative records as our reference group because these data types were free from recall bias and most closely matched our case definition.

We crosswalked each of the three non-reference (survey) data types to the reference (administrative record) data type via adjustment factors derived from MR-BRT meta-regressions. For each non-reference data type and each sex, we looked for overlap between the non-reference data type and the reference data type based on location, year, age group, and sex. The MR-BRT analyses were performed between each alternative data type and the reference with a spline on age and the covariates of hospital beds per 1000 and lag-distributed income (LDI) to account for non-systematic differences between the data types.

After crosswalking all non-reference data to the reference data type, we used DisMod-MR 2.1 model with all data disaggregated by age to estimate countries' age-pattern. This age pattern was then applied the estimated age-pattern to split aggregated age data into the most granular age groups that are necessary for ST-GPR. The age pattern used to split aggregated age data is shown below.

Figure 5. Age-pattern used to age-split wide age bin data



Before modelling, we applied a systematic outliering processes to identify data points that differed substantially from the trend. To do this, we calculated the median absolute deviation (MAD) from the age-standardized mean utilisation for each sex-location-year-source combination. Points that were more than three MADs above or below median utilisation were marked as outliers.

Modelling strategy

The input data were modelled using ST-GPR to allow for smoothing over age, time, and location to produce estimates of utilisation for every age, sex, location, year combination in the GBD. We included three covariates to help explain variation in geographies with little to no data and included random effects on location in the modelling specifications. We used the covariates of the natural log of hospital beds per 1000, natural log of health expenditure per capita, and the HAQ Index for every location. Coefficients for the covariates are presented in the table that follows.

Table 5. Estimated coefficients of the inpatient envelope model.

Covariate	Sex	Coefficient (95% UI)	Exponentiated Coefficient (95% UI)
Log hospital beds per 1000	Male	0.60 (0.57, 0.63)	1.82 (1.77, 1.88)
	Female	0.50 (0.46, 0.53)	1.64 (1.59, 1.70)
HAQ Index	Male	-0.000039 (-0.0012, 0.0012)	1.00 (0.99, 1.00)
	Female	0.00080 (-0.00045, 0.0021)	1.00 (1.00, 1.00)

Log health expenditure per capita	Male	0.21 (0.19, 0.23)	1.24 (1.21, 1.26)
	Female	0.22 (0.20, 0.23)	1.24 (1.22, 1.26)

Changes from GBD 2019 to GBD 2021

Relative to GBD 2019, there were a number of changes to the inpatient utilisation envelope modelling strategy. First was the addition of new input data, both from survey series and new years of administrative data. Second, the crosswalk analysis was done using MR-BRT, where it was previously done via penalised spline regressions. Third was the incorporation of the MAD outliering technique, to help systematically identify implausible estimates of utilisation in the input data prior to modelling. Fourth, we no longer used in-facility delivery estimates as input data for the youngest age group, relying instead on input data from administrative sources and surveys. Finally, we no longer used the all-cause mortality covariate in the ST-GPR model. All together, these changes resulted in more robust estimates of inpatient utilisation across GBD demographics.

Section 2.3 Data Adjustments

Section 2.3.1: Crosswalking

Crosswalking refers to the process of adjusting data for known biases. An observation is considered biased if it differs in a consistent way from the standard GBD definition of the modeled parameter. Examples include self-reported rather than doctor-diagnosed measures of disease incidence, or diagnostic tests with a lower sensitivity or specificity compared to the gold standard diagnostic method. If the difference between an alternative measurement method and the GBD definition is consistent and systematic, we can model it as a function of covariates and use this model to predict the degree of adjustment needed for a given alternative or non-standard observation. The result of crosswalking is that GBD models can incorporate data from a wider range of sources.

Specifically, crosswalking involves:

1. Finding pairs of alternative and reference (e.g. self-reported and measured) observations that match on relevant criteria (e.g. age, sex, location and year);
2. Taking the difference between these observations in log or logit space, to ensure that the crosswalk adjustment remains bounded correctly;
3. Running a meta-regression model that estimates this difference potentially as a function of covariates;
4. Predicting how much each alternative data point in the original dataset should be adjusted; and

5. Applying the adjustment.

Section 2.3.2: Bias adjustment for alternative case definitions and study methods

In GBD 2021 we continued the practice started in GBD 2019 of crosswalking non-fatal and risk exposure data to account for alternative case definitions or study methods. The adjustments were applied prior to entering data into our main analytical tools of DisMod-MR 2.1 and ST-GPR, ensuring that all data inputs were expressed on a consistent scale. We also used this approach to convert data presented for both sexes to a male and female equivalent. The starting point was to explicitly state the reference case definition and study method and identify alternative definitions and study characteristics that fall within our inclusion criteria.

We compiled data from both within-study comparisons (ie, data that used alternative and reference definitions in the same population) and between-study comparisons (ie, data that used an alternative definition in one population and a reference definition in another population that overlap in location, time, age, and sex) of different case definitions. For between-study comparisons, we allowed a maximum calendar year difference between studies of five years. Where validation studies (ie, those carried out at the introduction of a new set of diagnostic criteria comparing to previous criteria) were available, we extracted data on the comparison of alternative to reference. For quantities of interest with multiple alternative definitions/methods we also looked for pairs comparing two alternatives.

If both between and within study pairs were available, we examined whether there was a systematic difference between these. If there was a significant difference, we made judgement call as to whether within-study or between study data comparisons were most appropriate. In general, this was the within-study data. However, there were important measurement or conceptual reasons for choosing between-study data. For example, for crosswalks between self-reported height and weight compared to measured height and weight, between-study comparisons may be preferable if respondents knew they would be measured and, therefore, were less likely to misreport their height and weight.

To quantify the degree of bias for an alternative data source, we calculated the difference between matched pairs of alternative and reference observations and used this quantity as the dependent variable in a mixed effect meta-regression model. The model could include any number of covariates to capture how bias might vary as a function of other variables, like age or sex. Predictions from the model were then used to convert alternative observations to their equivalent reference values. For GBD 2021, we developed an open source Python package to facilitate the process of modeling and applying bias adjustments (ihmeuw-msca , 2023)

To choose covariates for the model, we examined whether there were systematic differences in the adjustments by key demographics (age, sex, geographic location, year) and other potential factors that may lead to variation in the degree of bias adjustment. We did this when there was a strong rationale, eg, biological plausibility, for variation by such characteristics. After fitting

the model, for predicted adjustment factors that were not statistically significant, we still applied the adjustments if there was a conceptual reason to believe that the alternative definition is biased. This expands the variance of data points using a non-standard case definition or study method, effectively reducing their influence in subsequent modeling steps.

Section 2.3.3: Example bias adjustment calculation

As an example, we provide mathematical notation for a bias adjustment to a data source that measures prevalence using a non-standard case definition. We have pairs of alternative and reference observations (denoted i) that match on age, sex, location, and time period combination (denoted j). The degree of bias varies as a function of age and sex. Because the parameter of interest is prevalence, which is bounded by 0 and 1, we calculate the logit-scale difference between alternative and reference observations in a given matched pair:

$$y_{i,j} = logit(p_{i,j}^{alt}) - logit(p_{i,j}^{ref}).$$

In preparing the data for this calculation, if the values of either the reference or alternative were zero, we aggregated values across age groups until both values had non-zero observations. We used the delta method to compute the standard error of the reference and alternative measures in logit space. The standard error of the logit-scale difference was computed as the square root of the sum of the variances of each data point in a pair.

If the parameter had instead been bounded by only 0, like incidence, we would have calculated the log-scale difference. From simulations we found that the two methods provide almost identical results for quantities that after adjustment do not exceed a value of 0.5 (eg, prevalence or proportion). The logit-scale difference method much better dealt with higher values and avoided prevalence or proportions to exceed one.

As a next step in this hypothetical example, we modeled the differences as the dependent variable in a mixed-effects meta-regression model with age and sex as covariates:

$$y_{i,j} = \beta_0 + \beta_1\ age_i + \beta_2\ sex_i + u_j + \epsilon_i$$
$$u_j \sim N(0, \gamma)$$
$$\epsilon_i \sim N(0, \sigma_i^2)$$

We then used the linear predictor of this model to predict the degree of bias adjustment needed for the various age and sex combinations among the alternative observations:

$$\hat{\iota}_{a,s} = f(age, sex) = \beta_0 + \beta_1\ age + \beta_2\ sex.$$

To adjust a particular alternative observation $p_{a,s}^{alt}$ we subtracted the adjustment factor in logit space, and the inverse logit transformation was applied to the result to convert back to natural units:

$$p_{a,s}^{adjusted} = logit^{-1}(logit(p_{a,s}^{alt}) - \delta_{a,s}).$$

The uncertainty for the adjusted logit-scale prevalence includes:

- uncertainty of the original observation in logit space,
- uncertainty from the posterior distribution of the predicted adjustment, and
- random intercepts in the meta-regression model (denoted γ above).

The variances from the three components were summed and then transformed into natural unit space using the delta method.

Section 2.3.4 Network Analysis

When there were multiple alternative case definitions or study methods, we used network analysis to leverage the additional information provided by indirect comparisons. For example, if A is the reference and B and C are two alternatives, the comparison of C versus A would be considered a direct comparison to the reference. This case was the subject of the previous section. In contrast, the combination of A versus B and B versus C provides an indirect comparison of the alternative C against the reference A. Or in other words, the inclusion of B-versus-C comparisons in the dataset provides additional information with which to estimate the difference between C and A.

Implementing a network analysis requires careful construction of the design matrix, or the dataset we pass to the mixed effects meta-regression model. Continuing the example with reference A and alternatives B and C, the design matrix for a network analysis with no covariates is created as follows:

- Create *k* dummy variables where *k* are all definitions/methods other than A (eg, *k* = B, C)
- Code dummy *k* as
  - 1 if the first term of the logit-scale difference is *k*;
  - -1 if *k* is second term of the logit-scale difference;
  - 0 otherwise

For example:

Study	Comparison	DummyB	DummyC
1	logit(B)-logit(A)	1	0
2	logit(B)-logit(A)	1	0
3	logit(C)-logit(A)	0	1
4	logit(C)-logit(A)	0	1
5	logit(C)-logit(B)	-1	1
6	logit(C)-logit(B)	-1	1

The coding structure outlined above assumes that all case definitions are mutually exclusive. In some cases, however, individual case definitions are composed of different sub-components or dimensions. For example, case definitions may vary by the type of symptoms that a respondent experiences as well as the recall period over which those symptoms are experienced. In the presence of sparse data, it may be difficult to find both direct and indirect comparisons of all individual case definitions. In these cases, an alternative approach is to assume different dimensions of case definitions have a multiplicative effect. In other words, the effect of recall period has the same relative effect across different categories of symptoms reported by respondents. To implement this coding scheme:

- Create *k* dummy variable columns for each case definition dimension.
- For each dummy variable k:
    - Add 1 if *k* is a component of the first term in the logit-scale difference.
    - Subtract 1 if *k* is a component of the second term in the logit-scale difference.

Network analysis is a feature of the open source Python package for conducting bias adjustments (ihmeuw-msca , 2023) mentioned earlier. The package abstracts away the need to create the design matrix manually as in this example and can incorporate an arbitrary number of alternative definitions and covariates.

Section 2.3.5 Age sex splitting

Before modelling, we ran a Dismod-MR 2.1 model with data disaggregated by age to estimate countries' age-pattern and then applied the estimated age-pattern to split aggregated all-age data into the 5-year age groups preferred for ST-GPR modelling. This procedure was done by calculating a constant, k, which was the ratio of the aggregated all-age data point, $\mu_{all\ age}$, to the all-age estimated utilisation rate from the DisMod-MR 2.1 model,$\hat{\mu}$

$$k = \frac{\mu_{all\ age}}{\hat{\mu}}$$

The constant, k, was then multiplied by age-specific utilisation rates from the DisMod-MR 2.1 model. Observation-specific uncertainty and uncertainty from the estimated age-pattern were both propagated into the uncertainty for a given post-splitting data point. The split data were then incorporated into the final DisMod-MR 2.1 model.

Section 2.4: Spatiotemporal Gaussian process regression (ST-GPR) modelling

The input data were modelled by using ST-GPR to allow for smoothing over age, time, and location in locations that were missing complete datasets.

The flowchart showing the analytic steps can be found elsewhere (Collaborators, 2020) The approach is a stochastic modelling technique that is designed to detect signals amidst noisy data. It also serves as a powerful tool for interpolating non-linear trends (Vasudevan S, 2009) (CE, 2005). Unlike classical linear models that assume that the trend underlying data follows a definitive functional form, GPR assumes that the specific trend of interest follows a Gaussian process, which is defined by a mean function $m(\cdot)$ and a covariance function $Cov(\cdot)$. For example, let $p_{c,a,s,t}$ be the prevalence, in normal, log, or logit space, observed in country *c*, for age group *a*, and sex *s* at time t:

$$(p_{c,a,s,t}) = g_{c,a,s}(t) + \epsilon_{c,a,s,t}$$

where

$$\epsilon_{c,a,s,t} \sim Normal(0, \sigma_p^2),$$
$$g_{c,a,s}(t) \sim GP\ (m_{c,a,s}(t), Cov\ (g_{c,a,s}(t))).$$

The derivation of the mean and covariance functions, $m_{c,a,s}(t)$ and $Cov\ (g_{c,a,s}(t))$, along with a more detailed description of the error variance (σ_p^2), is described below.

Section 2.4.1 Estimating mean functions

We estimated mean functions by using a two-step approach. To be more specific, $m_{c,a,s}(t)$ can be expressed, depending on the prevalence transformation, as:

$$log(p_{c,a,s}(t)) = X_{c,a,s}\beta + h(r_{c,a,s,t})$$

$$logit(p_{c,a,s}(t)) = X_{c,a,s}\beta + h(r_{c,a,s,t})$$

$$p_{c,a,s}(t) = X_{c,a,s}\beta + h(r_{c,a,s,t})$$

where $X\beta$ is the summation of the components of a hierarchical mixed-effects linear regression, including the intercept and the product of covariates with their corresponding fixed-effect coefficients. Some models were run as hierarchical mixed-effects linear regressions with random effects on the levels of the location hierarchy. For most mixed-effects models, random effects were only used in the fit, not in the prediction. The second part of the equation, $h(r_{c,a,s,t})$, is a smoothing function for the residuals, $r_{c,a,s,t}$, derived from the linear model.[44] Cause-specific methods details can be found in appendix sections 6.

Although the linear component captures general trends over time, much of the data variability may still not be adequately accounted for. To address this, we fit a locally weighted polynomial regression (locally estimated scatterplot smoothing, or LOESS) function $h(r_{c,a,s,t})$ to systematically estimate this residual variability by borrowing strength across time, age, and space patterns (the spatiotemporal component of ST-GPR) (Ng M, 2014) (Ng M) The time adjustment parameter, defined by λ , aims to borrow strength from neighboring time points (ie, the prevalence in this year is highly correlated with prevalence in the previous year but less so

further back in time). The age-adjustment parameter, defined by ω, borrows strength from data in neighboring age groups. The space-adjustment parameter, defined by ξ, aims to borrow strength across the hierarchy of geographical locations. The spatial and temporal weights are combined into a single space-time weight to allow the amount of spatial weight given to a particular point $r_{c,a,s,t}$ to fluctuate given the data availability at each time *t* and location-level *l* in the location hierarchy.

Let $w_{c,a,s,t}$ be the final weight assigned to observation $r_{c,a,s,t}$ with reference to a focal observation r_{c_0,a_0,s_0,t_0}. We first generated a temporal weight $t.w_{c,a,s,t}$ for smoothing over time, which was based on the scaled distance along the time dimension of the two observations (Ng M) :

$$t.w_{c,a,s,t} = \frac{1}{e^{\lambda|t-t_0|}}$$

Next, we generated a spatial weight to smooth over geography. Specifically, we defined a geospatial relationship by categorizing data based on the GBD location hierarchy (table 1). *zeta* acts as a scalar on a given datapoint given its proximity to the target location:

$$t.w_{c,a,s,t} = \zeta^{|c-c_0|}$$

For example, estimating a country, would use the following weighting scheme:

- Country data: ζ^0 = 1
- Regional data not from the country being estimated: ζ^1
- Data from other regions in the same super region: ξ^2
- Global data from other super regions: ζ^3

Under the spatial weighting specification, typical values of ζ range from [0.001, 0.2], where ζ can be interpreted as the amount to downweight regional datapoints compared to country datapoints for a given estimating country. For example, for a given datapoint $r_{c,a,s,t}$ and ζ = 0.01, a datapoint not within country *c* but within the same region *r* as $r_{c,a,s,t}$ would be assigned $\frac{1}{100}$ the weight of a datapoint within the country.

The spatial and temporal weights were then multiplied and summed across each level of the location hierarchy and normalised for each time period *t* . This procedure allowed the space-time weight to implicitly take into account the amount of data available at the country vs. region vs. super-region level and attribute spatial weight accordingly.

Given a normalisation constant,

$$K_i = \sum_{c\epsilon C} s.w_{c,t} * t.w_{c.t} + \sum_{c\epsilon R} s.w_{c,t} * t.w_{c.t} + \sum_{c\epsilon SR} s.w_{c,t} * t.w_{c.t}$$

the final space-time weight would then equal

$$w'_{c,a,s,t} = \frac{s.w_{c,t} * t.w_{c,t}}{K_i}$$

Finally, we calculated the weight $w''_{c,a,s,t}$ to smooth over age, which is based on a distance along the age dimension of two observations. For a point between the age a of the observation $r_{c,a,s,t}$ and a focal observation r_{c_0,a_0,s_0,t_0}, the weight is defined as follows:

$$w''_{c,a,s,t} = \frac{1}{e^{\omega|a-a_0|}}$$

The final weights were then computed by simply multiplying the space-time weights and age weights and normalising so all weights for a given time period t sum to 1. A full derivation of weights for each category, assuming the location being estimated was a country, follows:

1) If the observation $r_{c,t}$ belongs to the same country c_0 of the focal observation r_{c_0,t_0} :

$$w_{c,a,s,t} = \frac{(w'_{c,a,s,t} \quad w''_{c,a,s,t})}{\sum_{c=c_0}(w_{c,a,s,t} \quad w_{c,a,s,t})} \qquad \forall c = c_0$$

2) If the observation $r_{c,t}$ belongs to a different country than the focal observation r_{c_0,t_0}, but both belong to the same region *R*:

$$w_{c,a,s,t} = \frac{(w'_{c,a,s,t} \quad w''_{c,a,s,t})}{\sum_{c\neq c_0}(w_{c,a,s,t} \quad w_{c,a,s,t})} \qquad \forall c \neq c_0 \cap R[c] = R[c_0]$$

3) If the observation $r_{c,t}$ belongs to the same super region *SR* but to both a different country c_0 and a different region $R[c_0]$ than the focal observation r_{c_0,t_0}:

$$w_{c,a,s,t} = \frac{(w'_{c,a,s,t} \quad w''_{c,a,s,t})}{\sum_{c\neq c_0}(w_{c,a,s,t} \quad w_{c,a,s,t})} \qquad \forall c \neq c_0 \cap R[c] \neq R[c_0] \cap SR[c] = SR[c_0]$$

4) If the observation $r_{c,t}$ is from a different super region than the focal observation r_{c_0,t_0} (Ie, all other data currently not receiving a weight):

$$w_{c,a,s,t} = \frac{(w'_{c,a,s,t} \quad w''_{c,a,s,t})}{\sum_{c\neq c_0}(w_{c,a,s,t} \quad w_{c,a,s,t})} \qquad \forall c \neq c_0 \cap R[c] \neq R[c_0] \cap SR[c] \neq SR[c_0]$$

Observations could be downweighted by a factor of 0.1, usually because they were not geographically representative at the unit of estimation. Details of reasons for downweighting can be found in cause-specific modeling summaries. The final weights were then normalised such that the sum of weights across age, time, and geographic hierarchy for a reference group was 1.

Section 2.4.1: Estimating error variance

σ_p^2 represents the error variance in normal or transformed space including the sampling variance of the estimates and prediction error from any crosswalks performed. First, variance was systematically imputed if the data extraction did not include any measure of uncertainty. When some sample sizes for data were available, missing sample sizes were imputed as the 5th percentile of available sample sizes. Missing variances were then calculated as $\sigma_p^2 = \frac{p*(1-p)}{n}$ for proportions or were predicted from the mean by using a regression for continuous values. When sample sizes were entirely missing and could not be imputed, the 95th percentile of available variances at the most granular geographic level (ie, first country, then region, etc.) were used to impute missing variances. For proportions where *p*n* or *(1-p)*n* is <20, variance was replaced by using the Wilson Interval Score method.

Next, if prevalence was modelled as a log transformation, the error variance was transformed into log-space by using the delta method approximation as follows:

$$\sigma_p^2 \cong \frac{\sigma_{p'}^2}{p_{c,a,s,t}^2}$$

where $\sigma_{p'}^2$ represents the error variance in normal space. If prevalence was modelled as a logit transformation, the error variance was transformed into logit-space by using the delta method approximation as follows:

$$\sigma_p^2 \cong \frac{\sigma_{p'}^2}{(p_{c,a,s,t} * (1 - p_{c,a,s,t}))^2}$$

Finally, prior to GPR, an approximation of non-sampling variance was added to the error variance. Calculations of non-sampling variance were done on normal-space variances. Non-sampling variance was calculated as the variance of inverse-variance weighted residuals from the space-time estimate at a given location-level hierarchy. If there were <10 data points at a given level of the location hierarchy, the non-sampling variance was replaced with that of the next highest geography level with >10 data points.

Section 2.4.2: Estimating the covariance function

The final input into GPR is the covariance function, which defines the shape and distribution of the trends. Here, we have chosen the Matern-Euclidian covariance function, which offers the flexibility to model a wide spectrum of trends with varying degrees of smoothness. The function is defined as follows:

$$M(t, t') = \sigma^2 \frac{2^{1-v}}{\Gamma(v)} \left(\frac{d(t, t')\sqrt{2v}}{l} \right)^v K_v \left(\frac{d(t, t')\sqrt{2v}}{l} \right)$$

where $d(\cdot)$is a distance function; σ^2, v, l, and K_v are hyperparameters of the covariance function—specifically σ^2 is the marginal variance, v is the smoothness parameter that defines the differentiability of the function, l is the length scale, which roughly defines the distance between which two points become uncorrelated, and K_v is the Bessel function. We approximated σ^2 by taking the normalised median absolute deviation $MADN(r_\partial')$ of the difference, which is the normalised absolute deviation of the difference of the first-stage linear regression estimate from the second-stage spatiotemporal smoothing step for each country. We then took the mean of these country-level MADN estimates for all countries with 10+ country-years of data to ensure that differences between first- and second-stage estimates had sufficient data to truly convey meaningful information on model uncertainty. We used the parameter specification $v = 2$ for all models. The scale parameter l used for each cause is reported in appendix sections 3.4 and 4.12.

Section 2.4.3: Prediction using GPR

We integrated over $g_{c,t}(t_*)$ to predict a full time series for country c, age a, sex s, and prediction time t_*as follows:

$$p_{c,a,s}(t_*) \sim N\left(m_{c,a,s,t}(t_*), \sigma_p^2 I + Cov\left(g_{c,a,s,t}(t_*)\right)\right)$$

Random draws of 1000 samples were obtained from the distributions above for every country for a given indicator. The final estimated mean for each country was the mean of the draws. In addition, 95% UIs were calculated by taking the 2.5 and 97.5 percentile of the sample distribution. The linear modelling process was implemented by using the lmer4 package in R, and the ST-GPR analysis was implemented through the PyMC2 package in Python.

Section 2.4.4: Subnational scaling and aggregation

To ensure internal consistency of the estimates between countries and their respective subnational locations, national estimates were either created by population-weighted aggregation or subnational estimates were adjusted by population-weighted scaling to the national estimates, depending on the data coverage of a given country compared to that of its subnational locations. For example, if data coverage was better at the national level than at its corresponding subnational locations for a given country and cause across age, sex, and time, estimates were rescaled to be consistent with the national level. Conversely, if data coverage was better at the subnational level, estimates for its parent country were generated through population-weighted aggregation of subnational estimates.

Estimates can also be scaled within logit space. Scaling in logit space ensures that subnational estimates of proportion models do not exceed one after being rescaled to the national estimate.

Section 2.5: MR-BRT meta-regression modelling

Section 2.5.1 MR-BRT Overview

MR-BRT is a meta-regression modeling tool developed at IHME. In contrast to other types of regression, meta-regression incorporates uncertainty in the dependent variable; each observation comes with its own standard error. This characteristic is important when the input data are results of scientific studies that are reported with uncertainty. Observations with greater uncertainty are given less weight in the model. To describe variation in the parameter of interest, MR-BRT can incorporate both fixed and random effects. Fixed effects include binary and continuous covariates as in a traditional regression model. Random effects describe group-level variation and are often used to characterize differences between studies beyond what is captured by measured covariates.

Section 2.5.2 MR-BRT Formula

Formally, a linear mixed effects meta-regression as implemented in MR-BRT can be described as:

$$y_{ij} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + u_j + \epsilon_{ij}.$$

The variable y_{ij} refers to the value of observation i in study j; it is typically expressed in log or logit space to ensure that model predictions remain within logical constraints, for example that relative risks cannot be negative. The terms $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$comprise the linear predictor, including both the intercept and the effects of any number of covariates. The term u_j is a random intercept corresponding to study j. The full set of random intercepts is assumed to follow a Normal distribution where γ is the variance of between-study heterogeneity. Random

effects can be estimated for continuous covariates as well, in which case they are called random slopes. The term ϵ_{ij} refers to the stochastic error corresponding to observation i in study j, and the set of values are assumed to follow a Normal distribution in which observation-specific standard errors are known prior to modeling. This linear mixed effects formulation of the model covers most features MR-BRT.Features that involve nonlinear optimization techniques like the ratio model (described below) extend this framework and are described formally elsewhere (Zheng, 2021).

Section 2.5.3 MR-BRT Features

MR-BRT – as suggested by its full name "Meta-Regression with Bayesian priors, Regularization and Trimming" – comes equipped with several capabilities that expand upon the classical mixed effects meta-regression model:

- Bayesian priors can be applied to any estimated coefficient, enabling information from outside the dataset to be considered in the process of fitting the model. A Uniform prior sets hard bounds on the allowed values of an estimated coefficient. A Gaussian prior acts as a suggestion for the estimated value of a coefficient, with the standard deviation of the specified Gaussian distribution determining the strength of the prior.
- LASSO variable selection, also known as L1 regularization, can be implemented by specifying Laplace priors with mean 0 on the β coefficients. Similarly, ridge regression, also known as L2 regularization, can be implemented by specifying Gaussian priors with mean 0 on the β coefficients.
- Trimming is a method for identifying and removing the effects of outliers. Users define the proportion of points to be excluded and the algorithm determines which ones to exclude. Because the trimming algorithm is an integrated part of the model's likelihood function, MR-BRT identifies outliers and estimates the β coefficients simultaneously during the fitting process.
- A spline term may be used to describe the nonlinear effect of a covariate. MR-BRT implements a B-spline, or basis spline. Users have control over the flexibility of the estimated curve by specifying the number of knots, location of knots, spline degree (i.e. cubic or quadratic), linearity in the tail segments, convexity, concavity, or a monotonicity constraint requiring the spline to be non-decreasing or non-increasing.
- Pairs of exposure intervals may be used as an independent variable using a method known as the "ratio model". This feature is most often used when the epidemiological literature reports relative risks corresponding to a reference exposure range (e.g. BMI = [18,22)) and an alternative exposure range (e.g. BMI = [30,35)). It is usually used in conjunction with a spline to capture the nonlinear effect of the exposure. The ratio model works by integrating over the span of each interval and taking the ratio as part of the likelihood function (Zheng, 2021).

The source code for MR-BRT is publicly available on GitHub as the Python package mrtool (ihmeuw-msca , 2023) The mrtool package builds upon the open source mixed effects package LimeTr (https://github.com/zhengp0/limetr). For a full technical description of MR-BRT and the underlying mathematics (Zheng, 2021)

Section 2.6: DisMod-MR 2.1 estimation

Section 2.6.1: Estimation of sequelae and causes

The most extensively used estimation method is the Bayesian meta-regression method DisMod-MR 2.1. For some causes, such as HIV/AIDS or measles, disease-specific natural history models have been used for which the underlying three-state model in DisMod-MR 2.1 (susceptible, cases, dead) is insufficient to capture the complexity of a disease process. For some diseases with a range of sequelae differentiated by severity, such as COPD or diabetes mellitus, DisMod-MR 2.1 was used to meta-analyse the data on overall prevalence with separate DisMod-MR 2.1 models of the proportions of cases with different severity levels or sequelae. Likewise, DisMod-MR 2.1 was used to meta-analyse data on the proportions of liver cancer and cirrhosis due to underlying aetiologies such as hepatitis B, hepatitis C, and alcohol use disorders.

Section 2.6.2: DisMod-MR 2.1 description

Until GBD 2010, non-fatal estimates in burden of disease assessments were based on a single data source on prevalence, incidence, remission, or a mortality risk selected by the researcher as most relevant to a particular location and time. For GBD 2010, we set a more ambitious goal: to evaluate all available information on a disease that passes a minimum quality standard. That required a different analytical tool that would be able to pool disparate information presented for varying age groupings and from data sources by using different case definitions. The DisMod-MR 1.0 tool used in GBD 2010 evaluated and pooled all available data, adjusted data for systematic bias associated with case ascertainment methods that varied from the reference and produced estimates by world regions with UIs by using Bayesian statistical methods. For GBD 2013, the improved DisMod-MR 2.0 increased computational speed, which allowed computations to be consistent between all disease parameters at the country rather than the region level. The hundred-fold increase in speed of DisMod-MR 2.0 was partly due to a more efficient rewrite of the code in C++, but also due to switching to a model specification of log rates rather than a negative binomial model used in DisMod-MR 1.0. In cross-validation tests, the log rates specification worked as well as or better than the negative binomial specification.[39] The sequence of estimation occurs at five levels: global, super-region, region, country and, where applicable, subnational location. The super-region priors are generated at the global level with mixed-effects, non-linear regression by using all available data; the super-region fit, in turn, informs the region fit, and so on down the cascade. Analysts can choose to branch the cascade in terms of time and sex at different levels depending on data density. The default used in most models is to branch by sex after the global fit but to retain all years of data until the lowest level in the cascade is reached.

The computational engine is limited to three levels of random effects; we differentiate estimates at the super-region, region, and country level. In GBD 2013, the subnational units of China, the United Kingdom and Mexico were treated as “countries” to enable a random effect to be estimated for every location with contributing data. However, the lack of a hierarchy between country and subnational units meant that the fit to country data contributed as much to the estimation of a subnational unit as the fits for all other countries in the region. We found inconsistency between the country fit and the aggregation of subnational estimates when the country’s epidemiology varied from the average of the region. Adding an additional level of random effects required a prohibitively comprehensive rewrite of the underlying DisMod-MR engine. Instead, we added a fifth layer to the cascade, with subnational estimation informed by the country fit and country covariates, plus an adjustment based on the average of the residuals between the subnational location’s available data and its prior. This technique mimicked the impact of a random effect on estimates among subnationals.

In GBD 2015, we also improved how country covariates differentiate non-fatal estimates for diseases with sparse data. The coefficients for country covariates are re-estimated at each level of the cascade. For a given location, country coefficients are calculated by using both data and prior information available for that location. In the absence of data, the coefficient of its parent location is used to utilise the predictive power of our covariates in data-sparse situations.
For GBD 2016, the computational engine (DisMod-MR 2.1) remained substantively unchanged from GBD 2015. We updated the age prediction sets to include age groups 80–84 years, 85–89 years, 90–94 years, and 95 years and older to comply with changes across all functional areas of the GBD.

In GBD 2017, we continued to use DisMod-MR 2.1 because no substantial changes were made. Updates to computation include extending the terminal prediction year to 2017 and additional subnational units in Ethiopia, Iran, New Zealand, Norway, and the Russian Federation.
In GBD 2019 and 2021, no substantial changes were made to DisMod-MR 2.1, but we made more substantial changes to how we use the tool. First, we added the years 2019, 2020, and 2021 as additional years of estimation. Second, we also included the option again to have random effects on cause-specific mortality rates (CSMR) and EMR. This functionality had been dropped a couple of GBD rounds earlier. Third, as we did all our adjustments for alternative case definitions and study methods as well as adjustments to combined-sex data points prior to entering data into DisMod-MR 2.1, we no longer used the functionality in DisMod-MR 2.1 to estimate coefficients for study and sex covariates. Fourth, based on simulation testing conducted in GBD 2019 we found that coverage improved, and errors reduced when passing down priors with a wider setting of minimum coefficient of variation (which determines the uncertainty around priors and hence how 'informative' the priors are) than had generally been used in past GBD iterations. We settled on a default value of 0.8 where in the past values of 0.4 or less had been more commonly used. We made some exceptions for highly prevalent conditions where a lower minimum coefficient of variation (CV) setting achieved the task of making priors less informative, but not completely uninformative.

In GBD 2017 and 2019 GBD rounds we calculated priors on excess mortality and entered these as data points by matching sex-specific prevalence data with an age width of 20 or less with the corresponding CSMR for the same location and year. Forstability, we excluded calculation of EMR for prevalence data points of less than 1 in a million. EMR is simply calculated as CSMR divided by prevalence. As with previous GBD years, for diseases with an average duration of less than a year (as indicated by a setting of remission greater than one), we ran an initial global model to get an equivalent prevalence and used the following formula to calculate EMR:

$$EMR = {CSMR * (remission + (ACMR - CSMR) + EMR\_pred)}/{incidence}$$

where,

ACMR is the all-cause mortality rate
EMR_pred is the EMR fit from an initial global DisMod model

Despite using the log of LDI or the HAQ Index as a covariate with a prior that the coefficient had to be negative, we found many disease models with an implausible distribution of mortality to prevalence (or incidence) ratios implying lower case fatality in locations with lower HAQ Index than in countries with higher HAQ Index. This likely signals an inconsistency between fatal and non-fatal data inputs. For GBD 2019, we decided to run regressions on EMR data (calculated as described above) first using MR-BRT with HAQ Index as a predictor. In general, we tend to think that CSMR estimates are more robust than non-fatal data because of much greater data availability and a lesser task in adjusting cause death data for garbage coding than the complex task of adjusting non-fatal data sources for alternative case definitions and study methods. To indicate that we would reduce the random effects on EMR and the minimum coefficient of variation for priors on EMR being created at each next level down the cascade. However, there were exceptions. For drug use disorders, the risk of overdose deaths is less a function of a country's quality of health services but driven more by the availability of harm reduction strategies, such as opioid substitution therapy, and the availability of highly potent opioids such as fentanyl, which have been an important contributor to the large increase in overdose deaths in the USA in the last decade. We settled on a model for opioid use disorder with wider random effects and higher minimum coefficient of variation to give less emphasis on CSMR when enforcing consistency with prevalence data. In a next round, we will work to find covariates that are more relevant to drug overdose deaths such as a grading of harm reduction strategies by country and over time. In the case of COPD, we noted that following the data on CSMR and EMR led to large increases in prevalence estimates in east Asia, Oceania and, to a lesser extent, south Asia. In the oldest age groups, prevalence estimates would be higher than the prevalence data for these locations and reach a level of close to 80% in the oldest age groups. In these locations, we will pay attention to how garbage codes are being redistributed onto COPD in the next round of GBD.

Section 2.6.3: DisMod-MR 2.1 likelihood estimation

Analysts have the choice of using a Gaussian, log-Gaussian, Laplace, or Log-Laplace likelihood function in DisMod-MR 2.1. The default log-Gaussian equation for the data likelihood is

$$-log[p(y_j|\Phi)] = \log(\sqrt{2\pi}) + \log(\delta_j + s_j) + \frac{1}{2}\left(\frac{\log(a_j + \eta_j) - \log(m_j + \eta_j)}{\delta_j + s_j}\right)^2$$

Where,

y_j is a "measurement value" (ie, data point)

Φ denotes all model random variables

η_j is the offset value, *eta*, for a particular "integrand" (prevalence, incidence, remission, excess mortality rate, with-condition mortality rate, cause-specific mortality rate, relative risk, or standardised mortality ratio)

a_j is the adjusted measurement for data point j, defined by

$$a_j = e^{(-u_j - c_j)} y_j$$

Where:

u_j is the total "area effect" (ie, the sum of the random effects at three levels of the cascade: super-region, region and country) and

c_j is the total covariate effect (ie, the mean combined fixed effects for sex, study level, and country level covariates), defined by

$$c_j = \sum_{k=0}^{K[I(j)]-1} \beta_{I(j),k} \hat{X}_{k,j}$$

with SD

$$s_j = \sum_{l=0}^{L[I(j)]-1} \zeta_{I(j),l} \hat{Z}_{k,j}$$

Where:

k denotes the mean value of each data point in relation to a covariate (also called x-covariate)

I(j) denotes a data point for a particular integrand, *j*

$\beta_{I(j),k}$ is the multiplier of the k^{th} x-covariate for the i^{th} integrand

$\hat{X}_{k,j}$ is the covariate value corresponding to the data point *j* for covariate *k*;

l denotes the SD of each data point in relation to a covariate (also called z-covariate)

$\zeta_{I(j),k}$ is the multiplier of the l^{th} z-covariate for the i^{th} integrand

δ_j is the SD for adjusted measurement *j*, defined by:

$$\delta_j = log[y_j + e^{(-u_j - c_j)} \eta_j + c_j] - log[y_j + e^{(-u_j - c_j)} \eta_j]$$

Where:

m_j denotes the model for the j^{th} measurement, not counting effects or measurement noise, and defined by:

$$m_j = \frac{1}{B(j) - A(j)} \int_{A(j)}^{B(j)} I_j(a)\, da$$

Where:

A(j) is the lower bound of the age range for a data point
B(j) is the upper bound of the age range for a data point
I_j denotes the function of age corresponding to the integrand for data point *j*

Section 2.7: Impairment and underlying cause estimation

For GBD 2021, as in GBD 2019, GBD 2017 and GBD 2016, we estimated the country-age-sex-year prevalence of nine impairments. Impairments in GBD are conditions or specific domains of functional health loss that are spread across many GBD causes as sequelae and for which there are better data to estimate the occurrence of the overall impairment than for each sequela based on the underlying cause. These impairments included anaemia, epilepsy, hearing loss, heart failure, intellectual disability, infertility, vision loss, Guillain-Barré syndrome, and pelvic inflammatory disease. Overall impairment prevalence was estimated by using DisMod-MR 2.1. We constrained cause-specific estimates of impairments, as in the 19 causes of blindness, to sum to the total prevalence estimated for that impairment. Anaemia, epilepsy, hearing loss, heart failure, and intellectual disability were estimated at different levels of severity. Estimates were made separately for primary infertility (those unable to conceive), secondary infertility (those having trouble conceiving again), and whether the impairment affected men and/or women. In the case of epilepsy, we determined the proportions with idiopathic and secondary epilepsy as well as the proportions with severe and less severe epilepsy by using mixed effects regressions. The sparse data for the proportion of seizure-free, treated epilepsy were pooled in a random effects meta-analysis. DisMod-MR 2.1 models produced country-, age-, sex-, and year-specific severity levels of hearing loss and vision loss. Because of limited information on the severity levels of intellectual disability, we assumed a similar distribution of severity globally based on random effects meta-analysis of IQ-specific data for the overall impairment. This assumption was supplemented by cause-specific severity distributions for chromosomal causes and iodine deficiency; the severity of intellectual disability included in the long-term sequelae of causes including neonatal disorders, meningitis, encephalitis, neonatal tetanus, and malaria was estimated in combined health states of multiple impairments such as motor impairment, blindness, and/or seizures (R, 2015).We changed the name of the intellectual disability impairment to specify that estimates reflect cases arising during the developmental period, which we have defined as ages under 20 years. The severity of heart failure was derived from our Medical Expenditure Panel Surveys (MEPS) analysis and therefore was not specific for country, year, age, or sex.

A detailed description of the methods of each impairment can be found at the end of Section 4.12 of this appendix.

Section 2.7.1: Impairment squeeze

For the impairments epilepsy, intellectual disability, and blindness, mentioned above in Step 4, we often have better information regarding the total prevalence of the impairment rather than the prevalence of said impairment due to its various causes. For example, we have more data and a better idea of the total number of blind individuals (which we refer to herein as the blindness "envelope") in the world than we do the number of individuals who are blind due to a specific cause like retinopathy of prematurity or cataract. We achieve this consistency by either squeezing or inflating the individual sequela prevalence values so that their sums fit into each appropriate envelope. Blindness, epilepsy, and/or intellectual disability appear in various combinations with motor impairment levels as sequelae for a number of neonatal disorders and infectious diseases like malaria and neonatal tetanus ("Moderate motor impairment with blindness and epilepsy due to neonatal tetanus", for example). This presents an extra challenge because any squeeze or inflation of one of the impairments making up a sequela affects the others.

We set rules on how to do these adjustments sequentially. First, when the envelope of an impairment is smaller than the sum of all contributing causes, we redistribute the excess prevalent cases of combined impairment sequelae onto the sequelae that only have motor impairment (at a mild, moderate, or severe level) within the same cause grouping. Second, we apply the adjustments in a particular order such that we always fit at least one of the envelopes exactly where the other one or two envelopes may be exceeded by some amount. We first enforce a fit to the epilepsy impairment envelope, then intellectual disability, and last, blindness. Thus, the epilepsy envelope always matches exactly, whereas the intellectual disability and blindness envelopes may occasionally be exceeded on a draw-by-draw basis.

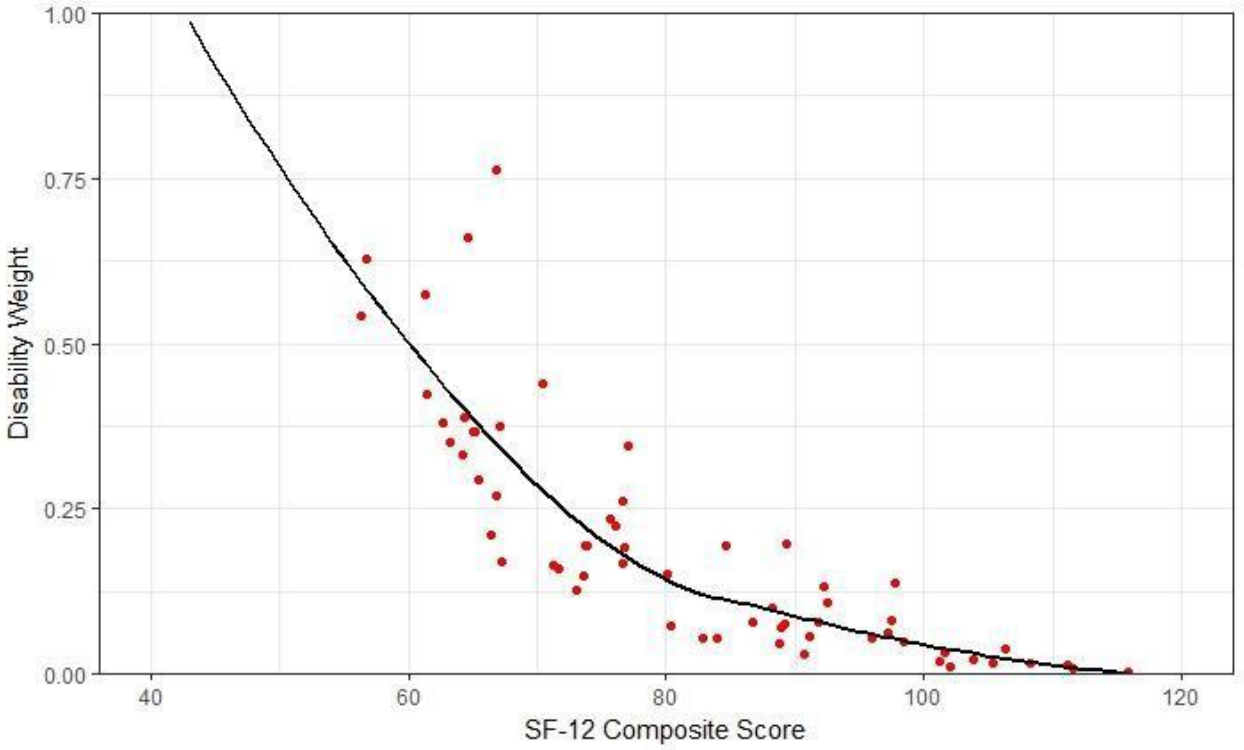
Section 2.8: Severity distributionSequelae were defined in terms of severity for 236 causes. We generally followed the same approach for estimating the distribution of severity we used in GBD 2019. In cases in which severity was related to a particular impairment, such as mild, moderate, and severe heart failure due to ischaemic heart disease or pulmonary arterial hypertension, the analysis was driven by impairment estimation methods. Severity levels for causes such as chronic kidney disease, epilepsy and COPD were modelled using DisMod-MR 2.1 or ST-GPR, whereas we performed meta-analyses to estimate the allocation of severity for causes such as rheumatoid arthritis, and multiple sclerosis. For dementia, we changed from using meta-analysis of three age categories to a more flexible model in MR-BRT using a spline on age. That allowed us to increase the number of studies informing severity from 7 to 67. For gallbladder and biliary diseases, we performed a meta-analysis of six community-based studies of the proportion of cases of gallbladder disease identified by ultrasonography who are symptomatic. In previous rounds, inpatient admission for gall bladder and biliary disease as a primary diagnosis were taken to represent symptomatic cases.

For many causes, we continue to have inadequate data on severity from surveys or the epidemiological literature. For those diseases, we made use of three population surveys: the MEPS 2000–2014, the [US] National Epidemiological Survey on Alcohol and Related Conditions (NESARC) 2000–2001 and 2004–2005, and the Australian National Survey of Mental Health and

Wellbeing of Adults (NSMHWB) 1997 (Medical Expenditure Panel Survey Home, n.d.) (Mental Health and Wellbeing: Profile of Adults, Australia, 1998) Each dataset contained individual-level measurements of functional health status made by using the 12-Item Short Form Health Survey (SF-12) as well as diagnostic information on the causes affecting each individual.
To use the data collected by measuring the distribution of severity with the SF-12, the individual SF-12 summary scores were mapped to an equivalent DW. A convenience sample of respondents was asked to complete SF-12 for the hypothetical individual living in a health state described by using a selection of 60 of the 235 health states with their lay descriptions from the GBD DW surveys reflecting the full range of severity. Each of these health states has a measured DW associated with it on a zero to one scale. We collected 2783 usable responses in total.

The final relationship between SF-12 score and DW is depicted in figure 8:

Figure 8. SF-12 composite scores and disability weights for 60 health states with fitted loess regression



To generate a smooth mapping from SF-12 combined scores to the GBD DW space, we used locally estimated scatterplot smoothing regression on the random effects for each health state. DWs were capped to remain between 0 and 1. All SF-12 survey data were thus transformed into DW space.
The second stage of the analysis was to build models predicting the transformed SF-12 scores as a function of the number of causes suffered by each individual. First, variable selection was performed by using least absolute shrinkage and selection operator (LASSO) regression to

penalize the regression coefficients of highly correlated causes. The tuning parameter, λ, controls the strength of the least-squares penalty. When λ=0, LASSO regression returns the same results as ordinary least-squares regression. Higher values of λ impose a stronger penalty and constrain a greater number of model parameters to 0. A ten-fold cross-validation was used to find the value of the λ that minimized the mean cross-validated error. This process resulted in a λ value of 0.0013 and eliminated 10 causes from the analysis. Transformed SF-12 scores into the DW scale for the remaining 190 causes were then modelled for each measure *m* of each individual *i* over *n* total causes in the survey as follows:

$$logit(DW)_{im} = \beta_0 + \beta_1 Condition1_{im} + \cdots + \beta_n Condition\, n_{im}$$

This equation effectively assumes that comorbid causes act to change SF-12 scores in a multiplicative fashion rather than an additive fashion.
To estimate the comorbidity-corrected effect of each cause (ie, in isolation) on total disability, we compared the predicted DW without the cause of interest (counterfactual DW) with the predicted DW including the cause of interest. Following the multiplicative comorbidity equation, the joint effect can be written

$$Condition\ specificDW = 1 - \frac{1 - predictedDW_m}{1 - counterfactualDW_m}$$

The mean of this cause-specific effect over all observations is the population marginal effect of a cause.

Using the model above, we estimate a counterfactual DW – the total individual DW excluding the effect of the cause of interest. We compared the observed distribution of functional health status with this counterfactual distribution to determine the marginal effect of the cause of interest. In other words, we estimated the health state for each individual and for each cause as the cumulative individual weight minus the effects of all comorbid causes.

$$Health\ state\ DW = 1 - \frac{1 - individual\ cumulativeDW_m}{1 - counterfactualDW_m}$$

The estimation strategy for health state-specific severity distributions for which there are multiple severity categories involved binning individuals' weights into severity cut-offs (eg, mild, moderate, and severe) for which DWs were derived. These bins were defined by using results from the GBD Disability Weights Studies (JA, 2015) for causes that had multiple health states defined. Cut-offs for the severity group were the midpoints between DWs of the health state and cases distributed into severity bins accordingly. For example, individuals with a health state DW above the mid-point between the mild DW and the moderate DW for a particular condition

would be assigned the moderate sequela. Cases were considered asymptomatic if the counterfactual weight was equal to or greater than the individual cumulative weight. The proportion of cases of a condition assigned to each level of severity for that condition was then used as the severity distribution of the condition for prevalence estimates to be apportioned accordingly into severity-specific prevalence estimates.

Section 2.9: Disability weights

To compute YLDs for a particular health outcome in a given population, the number of people living with that outcome is multiplied by a disability weight (DW) that represents the magnitude of health loss associated with the outcome. DWs are measured on a scale from 0 to 1; 0 implies a state equivalent to full health, and 1, a state equivalent to death.

Section 2.9.1: GBD 2010 Disability Weights Measurement Study

For GBD 2010, a primary data collection effort focused on measuring health loss rather than welfare loss by using a standardised approach of simple comparison questions directed to the general public across diverse communities.

Multi-country household surveys were conducted between Oct 28, 2009 and June 23, 2010 in five countries (Bangladesh, Indonesia, Peru, Tanzania, and the USA) selected to provide diversity across culture, language, and socioeconomic status.

Personal face-to-face computer-assisted interviews were conducted for all household surveys except for the survey in the US, which was conducted by computer-assisted telephone interview. Households were randomly selected by using a multistage stratified sampling design for which the probability of selection was proportional to the population size. In all cases, samples were designed to be representative of a given geographical area and, in the USA, to provide national representation.

For every contacted household, an adult respondent age 18 years or older was randomly selected by the survey program by means of the Kish approach. For face-to-face interviews, as many as three visits were made to selected households to establish contact. When a respondent was identified, as many as three return visits were made to do the survey at a time when the respondent was available. For the US telephone surveys, repeated calls were made up to seven times.

A web-based survey was posted at a dedicated URL between July 26, 2010 and May 16, 2011. The survey was initially available in English and subsequently available in Spanish and Mandarin. Recruitment of respondents occurred through several channels, such as news items and editorials in scientific journals, announcements at scientific meetings, postings on websites of institutions participating in the GBD, and social networking and communication mobilisation channels as well as direct contact with individuals and groups with known global health

interests by tapping into the professional networks of the study investigators and their colleagues. Participants in the web-based survey were required to be ages 18 or older. Household surveys obtained oral informed consent from all participants; written informed consent was obtained from participants in the web survey. Ethical review board approval was obtained from each household survey site and the University of Washington, Seattle, WA.

Standardised survey instruments were developed to obtain comparative assessments of the full array of disease and injury sequelae, parsimoniously captured in 220 unique health states. Lay descriptions of health states formed the basis for all comparisons. These descriptions used simple, non-clinical vocabulary that emphasised the major functional consequences and symptoms associated with each health state. Development of these descriptions involved an iterative process of detailed consultation with experts participating in the GBD 2010 study; the goals was to capture the most relevant details of each health state while avoiding ambiguity and ensuring consistency. When possible, health states were grounded in standard clinical classifications systems. For example, the Canadian Cardiovascular Society grading scale was referenced for descriptions of stages of angina (L., 2002), and the New York Heart Association functional classification was referenced for severity of heart failure (M., 1994). Pilot testing indicated that the lay descriptions in face-to-face interviews should not exceed 30 words.

A paired comparison question formed the basis of all surveys. The questions in the survey were framed with the following statement, “A person’s health may limit how well parts of his body or mind work. As a result, some people are not able to do all of the things in life that others may do, and some people are more severely limited than others. I am going to ask you a series of questions about different health problems. In each question, I will describe two different people…” Descriptions of two hypothetical people, each with a particular health state, were presented to respondents who were then asked which person they regarded as healthier. Health pairs in all surveys were selected by a randomizing computer algorithm. In the five household surveys, paired comparisons were presented for a subset of 108 health states pertaining to chronic conditions. The framing of chronic and acute conditions is different as they were presented as causing life-long or temporary health loss. We chose to only field health states that could be framed as lasting a lifetime in the household surveys as we hypothesized that presenting differently framed comparisons would be difficult to convey in face-to-face interviews. In the web survey, we considered this more feasible because respondents could read and refer to the framing of the question for each pair-wise comparison. All 220 health states were thus evaluated in the web survey.

In addition, the web survey included questions relating to population health and health programs specifically—such as “Imagine two different health programs. The first program prevented 1000 people from getting an illness that causes rapid death. The second program prevented 2000 people from getting an illness that is not fatal but causes lifelong health

problems resulting in moderate to severe disability. Which program would you say produced the greater overall health benefits?" This information was used to anchor the results from the pair-wise comparisons on the 0–1 DW scale.

Section 2.9.2: GBD 2013 European disability weights measurement study

The GBD 2010 DWs were critically dependent on the ways that outcomes were described to survey respondents. Descriptions for health states were designed to balance validity and parsimony, and this approach necessarily meant that some details of different health states had to be omitted. Because lay descriptions were developed collaboratively through individual expert groups organised around a particular set of health issues, some amount of variability in language and detail inevitably occurred. Criticisms and suggestions for improvement came from a number of commentators on the GBD 2010 DWs measurement study (E., 2013) (Taylor HR, 2013) (Voigt K, 2014)

GBD 2013 expanded the list of disease and injury causes and sequelae mapped to 235 unique health states. Additional data for the European Disability Weights Measurement Study were collected between September 23, 2013, and November 11, 2013, in Hungary, Italy, the Netherlands, and Sweden. The initiation of these surveys was connected to a project sponsored by the European Centre for Disease Prevention and Control (M, 2012)The four selected countries were chosen to be representative of the four regions of Europe (east, south, middle, and north) in terms of age, sex, and education of the respondents. Respondents were recruited from standing internet panels in each country on the basis of quota sampling with reference to age, sex, and education in such a way as to maintain the population representativeness of these characteristics. Eligible participants were 18–65 years old and were preselected in the Netherlands, where the age, sex, and education of respondents were already known, or in the other three countries, invited to participate via a web-link and then selected on the basis of their individual characteristics.

The protocol for the European DWs measurement study followed the protocol that was developed and implemented in the GBD 2010 DWs measurement study. Lay descriptions for some health states that lacked mention of an important symptom or for which consistency of wording across different levels of severity had been noted were reworded. The European DWs measurement study included 255 health states, of which 183 were used in the analyses of GBD 2013. Those 183 consisted of 135 of the 220 health states that were included in the European DWs measurement study with unmodified lay descriptions and 30 from GBD 2010 for which alternative lay descriptions were included. DWs were estimated for additional sequelae that were incorporated into GBD 2013 but had not been included in GBD 2010.

Finding high correlation in resulting DW values between the country surveys and the web survey, we analysed the results of all surveys together. We ran probit regression analyses on

the answers to the pair-wise comparison questions by using dummies for each health state with a value of 1 for the first state in a pair, –1 for the second state in a pair, and 0 for all states other than the pair. This method formalizes the intuition that if two health states in a pair produce similar health loss, the answers are likely to be evenly split; a pair of health states with very different health loss get many more responses favouring one over the other. The statistical methods infer the distances between values attached to different health states based on the frequencies of responses to the paired comparisons.

A second analytic step is needed to anchor the resulting estimates onto the 0–1 DWs scale, where 0 equals no loss of health, with 1 meant to represent loss equivalent to death. We anchored results from the probit regression analysis onto the 0–1 scale by using population health equivalence data from the GBD 2010 web survey by using a linear regression of the probit coefficients from the analysis of paired comparisons on the logit-transformed DW estimates derived from interval regression of the population health equivalence responses. Using numerical integration, we then estimated mean values for DWs on the natural 0–1 scale. Uncertainty was estimated by bootstrapping with 1,000 samples. For a complete listing of the lay descriptions and values for the 440 health states (including combined health states) used in GBD 2021, please refer to Table 6. For a complete overview of disability weights applied to the Global Burden of Disease Study (al, 2015)

Section 2.10: Comorbidity correction (COMO)

The final stage in the estimation of YLDs is a micro-simulation, which adjusts for comorbidity. We refer to this micro-simulation process as “COMO” (for comorbidity correction). For GBD 2019 and 2021, we estimated the co-occurrence of different diseases by simulating 20,000 individuals in each location-age-sex-year combination as exposed to the independent probability of having any of the sequelae included in GBD based on prevalence. We tested the contribution of dependent and independent comorbidity in the US MEPS data and found that independent comorbidity was the dominant factor even though well-known examples of dependent comorbidity exist, such as clustering of conditions like diabetes and stroke or anxiety and alcohol use disorders. Age was the main predictor of comorbidity such that age-specific micro-simulations accommodated most of the required comorbidity correction (Vos T, 2012)

The two components necessary for the computation of YLDs and are the two inputs into COMO: 1) prevalence of each disease sequela and 2) DWs. The prevalence values of causes are primarily produced by using DisMod-MR 2.1 and, for causes with multiple sequelae, subsequently apportioned into sequela-specific prevalence based on available estimates of the severity distribution. The estimation of DWs and severity distributions have been described earlier in this appendix.

The micro-simulation, as performed for each age-sex-location-year, can best be represented as a four-step process. First, simulated individuals (simulants) are exposed to independent

probabilities of having each sequela, where the probability is equal to the prevalence estimate. For each simulant, the probability of having a disease sequela is equal to the estimated prevalence . Each simulant is determined to have or not have the disease sequelae based on a draw from a binomial distribution. From this simulation, simulants end up with any number of sequelae, from 0 up to the theoretical maximum given their demographics. Second, the DW for each simulant is estimated on the basis of the disease sequelae that they have acquired. The formula for the cumulative DW for a simulant is one minus the multiplicative sum of one minus each DW present

$$Simulant\ DW_l = 1 - \prod_{k=i}^{j}(1 - DW_k)$$

Where:

DW_k is the DW for the k^{th} disease sequela that the simulant *l* has acquired.

Once the simulant DW is computed, the DW attributable to each sequela for the simulant is calculated by using the following formula:

$$ADW_{lk} = \frac{DW_k}{\sum_{k=i}^{k=j} DW_k} * Simulant\ DW_l$$

Where:

ADW_{lk} is the attributable DW for disease sequela *k* in simulant *l*

DW_k is the DW for disease sequela *k*

Simulant DW_l is the DW for simulant *l* from the combination of all sequelae that they have acquired.

This formula apportions the overall simulant DW to each condition in proportion to the DW of each condition in isolation.

Finally, YLDs per capita in an age-sex-country-year are computed by taking the sum of the attributable DWs for a disease sequela across simulants.

$$YLD\ Rate_k = \frac{\sum_{l=1}^{n} ADW_{lk}}{n}$$

The actual number of YLDs from disease sequela *k* in an age-sex-location-year is then computed as the YLD rate *k* times the appropriate age-sex-location-year population.

By repeating the simulation process for each age-sex-country-year 500 times, the uncertainty in the prevalence of each disease sequela and the DW is propagated into the final comorbidity corrected YLD results. We selected 20,000 simulants for each age-sex-location-year group on the basis of simulation testing, which has shown that results are stable for YLDs at this number of simulants even in the younger age groups when prevalence is relatively low. Mean results for

YLDs that reflect 10 million simulants (20,000 simulants multiplied by 500 iterations to capture uncertainty) are very stable in each age-sex-location-year. For any given location-year-age-sex group, acause aggregate prevalence values were calculated as 1- ∏(1-prevalence)

Section 2.11: YLD computation, uncertainty, and residual YLDs

For GBD 2021, we computed YLDs by sequela as prevalence multiplied by the DW for the health state associated with that sequela. The uncertainty ranges reported around YLDs incorporate uncertainty in prevalence and uncertainty in the DW. To do this, we take the 500 samples of comorbidity-corrected YLDs and 500 samples of the DW to generate 500 samples of the YLD distribution. We assume no correlation in the uncertainty in prevalence and DWs. The 95% uncertainty interval is reported as the 25^{th} and 975^{th} values of the distribution. UIs for YLDs at different points in time (1990, 1995, 2000, 2005, 2010, 2015, 2020 and 2021) for a given disease or sequela are correlated because of the shared uncertainty in the DW and DW draws are not year specific. For this reason, changes in YLDs over time can be significant even if the UIs of the two estimates of YLDs largely overlap And prevalence UIs are used to determine significance of change in YLDs over time since DW draws are year agnostic.

Section 2.11.1: Residual YLDs

Despite expanding our list of causes and sequelae in successive GBD iterations, many diseases remain for which we do not explicitly estimate disease prevalence and YLDs. Less common diseases and their sequelae were included in 34 residual categories (table 7). For 22 of these residual categories, epidemiological data on incidence or prevalence were available, so these were modelled accordingly. For 13 residual categories, epidemiological data on incidence and prevalence were not available, but sufficient CoD data allowed for CoD estimates. For these residual categories, we estimated YLDs by multiplying their YLL estimates by the ratio of YLDs to YLLs from the Level 3 causes in the same disease category that were explicitly modelled. This scaling was done for each country-sex-year. This approach made the simplifying assumption that the residual diseases caused disability proportionate to the ratio of disability to mortality in explicitly modelled diseases. We did not include causes with large disability but no or little mortality in estimating these ratios. For example, we estimated the YLDs from other neurological disorders from the YLD to YLL ratios for dementia, multiple sclerosis, and Parkinson's disease but did not include the YLDs from headaches and epilepsy in the ratio. Detailed information on how YLDs for residual causes were estimated are available in their respective cause writeups in section 6.

Section 2.12: Birth prevalence

A number of conditions are present at birth, and quantifying them is important in fully describing the epidemiology of diseases within populations. These include many conditions included in the GBD cause group of neonatal disorders, infections that are transmitted from

mother to child either transplacentally or during birth, and congenital birth defects arising either *de novo* or from maternal exposures. Although these conditions were included in the underlying models informing previous GBD iterations, we developed a system for reporting them for the first time in GBD 2017; a list of these causes is reported in table 8.
Mathematically (ie, in the models), conditions present at birth are equivalent to "birth prevalence." However, we report these as "incidence" in recognition of the way that GBD defines incidence as a new case of a disease or injury entering the population. To process these results for publication in GBD, we used a three-step process. First, the number of cases at birth was calculated as birth prevalence rate multiplied by number of live births for each location, sex, and year. Second, the number of cases present at birth were summed with incident cases during the early neonatal period (calculated as the 0-to-6-days incidence rate times the 0-to-6-days population), and the early neonatal incidence rate was recalculated by re-dividing by the 0-to-6-days population. Third, incidence rates for aggregate age groups were re-calculated by using the revised incidence figures for the early neonatal period.
Causes included in reporting are all of those for which birth prevalence has been estimated in GBD 2021 as part of existing modelling processes. Although extensive, this list should not be considered exhaustive of all of the conditions that can be present at birth. Future efforts in GBD will focus on identifying and comprehensively including all conditions present at birth, including revision of model frameworks as necessary. These efforts will also be facilitated by continuing improvements in the resolution of epidemiologic estimates of disease burden during pregnancy. These efforts are also expected to facilitate subsequent analyses derived from GBD that evaluate how maternal interventions, including pregnancy surveillance, can influence patterns of neonatal, infant, and child health.

Section 3: SDI

Section 3.1: SDI definition

The Socio-demographic Index (SDI) is a composite indicator of background social and economic conditions that influence health outcomes in each location. In short, it is the geometric mean of 0 to 1 indices of total fertility rate (TFR) for those younger than 25 years old (TFU25), mean education for those 15 years old and older (EDU15+), and lag-distributed income (LDI) per capita. For GBD 2021 after calculating SDI, values were multiplied by 100 for a scale of 0 to 100.

Section 3.2: Development of revised SDI indicator

SDI was originally constructed for GBD 2015 by using the Human Development Index (HDI) methodology, wherein a 0 to 1 index value was determined for each of the original three covariate inputs (TFR in ages 15 to 49 years, EDU15+, and LDI per capita) by using the observed minima and maxima over the estimation period to set the scales (H, 2016)

In response to feedback from collaborators and the evolution of the GBD, we have refined the indicator with each GBD cycle. Beginning in GBD 2017, along with our expanded estimation of age-specific fertility, we replaced TFR with TFU25 as one of the three component indices. The TFU25 provides a better measure of women's status in society because it focuses on ages at which childbearing disrupts the pursuit of education and entrance into the workforce. In addition, we observed that in highly developed countries, the TFU25 has tended to decline consistently over time despite rebounds in TFR driven by increasing fertility at older ages. The concordance correlation coefficient between SDI based on the GBD 2016 method and the updated method for GBD 2017 was 0.981.

During GBD 2016, we moved from using relative index scales to using absolute scales to enhance the stability of SDI interpretation over time because we noticed that the measure was highly sensitive to the addition of subnational units that tended to stretch the empirical minima and maxima.[21] We selected the minima and maxima of the scales by examining the relationships each of the inputs had with life expectancy at birth and under-5 mortality and by identifying points of limiting returns at both high and low values if they occurred before theoretical limits (eg, a TFU25 of 0) were reached.

Thus, for each covariate input, an index score of 0 represents the minimum level of each covariate input past which selected health outcomes can get no worse, and an index score of 1 represents the maximum level of each covariate input past which selected health outcomes cease to improve. As a composite, a location with an SDI of 0 would have a theoretical minimum level of sociodemographic development relevant to these health outcomes, and a location with an SDI of 1 (before multiplying by 100 for reporting) would have a theoretical maximum level of sociodemographic development relevant to these health outcomes.

We computed the index scores underlying SDI as follows:

$$I_{Cly} = \max\left(\frac{C_{ly} - C_{low}}{C_{high} - C_{low}}, 0.005\right)$$

Where:

I_{cly} is the index for covariate *C*, location *l*, and year *y* and is equal to the difference between the value of that covariate in that location-year and the lower bound of the covariate divided by the difference between the upper and lower bounds for that covariate

If the values of input covariates fell outside the upper or lower bounds, they were mapped to the respective upper or lower bounds. We also note that the index value for TFU25 was computed as $1 - I_{TFU25ly}$ because lower TFU25s correspond to higher levels of development and thus higher index scores. For GBD 2021, we expanded the computation of SDI to 1075 national and subnational locations spanning the time period 1950–2021.

The composite SDI is the geometric mean of these three indices for a given location-year. The cut-off values used to determine quintiles for analysis were then computed by using country-level estimates of SDI for the year 2019, excluding countries with populations less than 1 million.

For GBD 2021, final SDI values were multiplied by 100 for reporting, in order to improve understanding of and broader engagement with the values. As such, GBD 2021 SDI is calculated as it was in 2019, but multiplied by 100 at the end (see example calculation below). Final reporting values are on a 0 to 100 scale.

Example calculation

We present the equation used to calculate SDI for a hypothetical country in the year 2010:

$$TFU25 = 1.09;\ Mean\ educ\ yrs\ pc = 8.23;\ lnLDI = 9.60$$

$$I_{TFU25} = 1 - \frac{1.09 - 0}{3 - 0} = 0.637$$

$$I_{Educ} = \frac{8.23 - 0}{17 - 0} = 0.484$$

$$I_{\ln LDI} = \frac{9.60 - 5.52}{11.00 - 5.52} = 0.744$$

$$SDI = \sqrt[3]{I_{TFU25} \cdot I_{Educ} \cdot I_{\ln LDI}} = \sqrt[3]{.637 \cdot .484 \cdot .744} = 0.611$$

$$I_{lnLDI} = \frac{9.58 - 5.52}{11.00 - 5.52} = 0.741$$

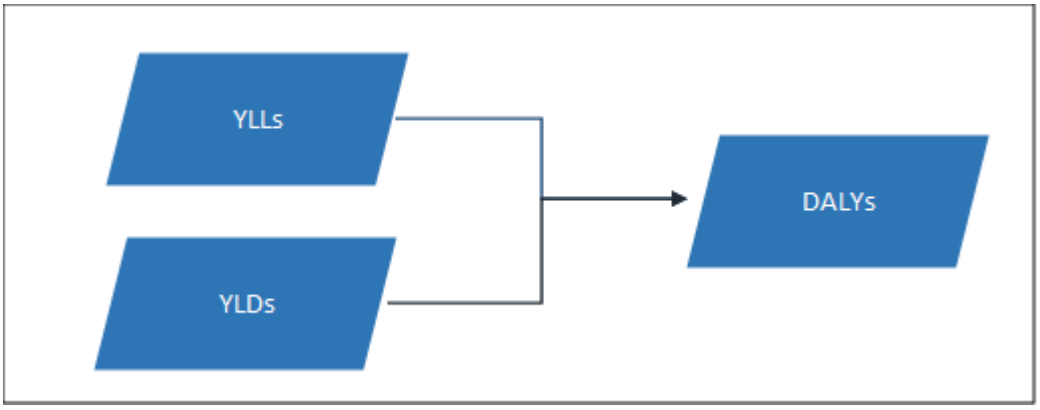
$$SDI = \sqrt[3]{I_{TFU25} \cdot I_{Educ} \cdot I_{\ln LDI}} = \sqrt[3]{.855 \cdot .543 \cdot .741} = 0.701$$

$$\text{GBD 2019 reporting } SDI = 0.701 * 100 = 70.1$$

Section 4: Estimation process for DALYs

To estimate DALYs for GBD 2021, we started by estimating cause-specific mortality and non-fatal health loss. For each year for which YLDs have been estimated, we computed DALYs by adding YLLs and YLDs for each age-sex-location (Figure A). Uncertainty in YLLs was assumed to be independent of uncertainty in YLDs. We calculated 500 draws for DALYs by summing the first draw of the 500 draws for YLLs and YLDs and then repeating for each subsequent draw. 95% UIs were computed by using the 25th and 975th ordered draw of the DALY uncertainty distribution. We calculated DALYs as the sum of YLLs and YLDs for each cause, location, age group, sex, and year.

Figure 9. DALY burden estimation for GBD 2021



Section 5: HALE

The first step to calculating healthy life expectancy for a population (defined by sex, country, and year) was to compute average health of individuals for every age group in that population. We combined information about prevalences for all sequelae and their associated disability weights, and accounted for comorbidity with a Monte Carlo simulation approach. We made the assumption that comorbidities were independent within each age group. We created simulations where individuals were exposed to each sequela with a probability equal to the estimated prevalence of that sequela in each age group. This created a simulated population where the frequencies of many possible multi-morbidities were consistent with the underlying estimates of prevalence. We define 1 minus the disability weight as the positive health associated with each sequela. The combined health for a simulated individual was the product of these positive health values for all relevant sequelae in the presence of multiple sequelae.

Average health values are computed as 1 minute the YLD per person in a population, which are then used to compute health adjusted person years. We incorporated average health values into the life table using Sullivan's method. First, we multiplied values in the nLx (average person-years lived within an age interval starting at age x) column of the life table by the corresponding average health value in that interval. We recalculated the rest of the life table using the adjusted nLx values. Sullivan's method began with an adjusted estimate of health adjusted life years within the terminal age interval (equal to nLx multiplied by the average health value for the terminal age group) and subsequent calculations we produced estimates by iterating through younger age intervals, summing the health-adjusted person-years with all age intervals above the current age interval to generate health adjusted person years lived above a certain age (adjusted Tx) for each age group. After calculating adjusted Tx for all age groups, HALE was calculated by dividing the adjusted Tx for each age group by the proportion of hypothetical birth cohort still alive at age x (al K. H., 2018).