GLOBOCAN 2022 annexes

Annexes A-E (additional supporting information to be available online)

Annex A. Cancer incidence and mortality data: sources and methods by country.
Annex B. List of cancer types included in GLOBOCAN 2022 and criteria for including and allocating certain malignancies.
Annex C. Modelling of incidence and mortality.
Annex D. Computation of the standard error by method of estimation.
Annex E. Methods of estimation and penalties used to correct the standard error.

Annex A. Cancer incidence and mortality data: sources and methods by country

Estimates of cancer incidence by country

- For 55 countries with 6-10 years of recent national cancer incidence data available, corresponding rates for 2022 were predicted using short-term prediction models (*method 1*) [1]. Cancer and sex-specific prediction models were fitted only when at least 50 cancer-specific cases for all ages were recorded per year. For the cancer and sex combinations where these criteria were not satisfied, the rates for 2022 were derived from the annual average rates recorded in the most recent period (of at least three years).

- For 39 countries where no retrospective national cancer incidence data existed (*method 1* above), or where no national mortality data were available (*method 3* below), the most recent cancer incidence rates from one subnational cancer registry (*method 2a*, 22 countries) or from multiple subnational registries (*method 2b,* 17 countries) within the country were used as proxy for 2022.

- Where registries were subnational and where national mortality data were available, national incidence was estimated from national mortality using statistical models, with the fitted mortality to incidence (M:I) ratios (*method 3*, 52 countries). Cancer and sex-specific M:I ratios were derived from recorded data from one or more cancer registries within the country (*method 3a*) or from recorded data from neighbouring countries, with M:I ratios between countries scaled according to levels of Human Development Index (HDI) (*method 3b*, see Annex C) or from survival estimates (method 3c)[1][2].

- When neither national or subnational registries, nor national mortality data were available, and the within-country source information was considered to lack the necessary level of accuracy, a set of age- and sex-specific national incidence rates for all cancers combined were obtained averaging overall rates from neighbouring countries. These rates were then partitioned to obtain the national incidence for specific sites using available cancer-specific relative frequency data (*method 4*, 1 countries).

- When neither national or subnational registries, nor national mortality data were available, and the within-country source information was either unavailable or compatible, average incidence rates from neighbouring countries in the same region were used to derive national incidence within the country (*method 9*, 38 countries).

Estimates of cancer mortality by country

To maximize comparability across countries, deaths coded to ill-defined categories (ICD-10, chapter XVIII) were redistributed pro rata across cancers (malignant neoplasms ICD-10 'C' category) and all other causes excluding injuries, by year and sex. The corrected "cancer deaths" 'C' category was then partitioned into cancer-specific categories using proportions from the uncorrected data. Vital registration is also known to be incomplete during the period under study for some countries and the source data were therefore corrected using the estimated completeness as reported by the

[1] Arnold M, Rutherford M, Lam F, Bray F, Ervik M, Soerjomataram I (2019). ICBP SURVMARK-2 online tool: International Cancer Survival Benchmarking. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/survival/survmark/, accessed [12/12/2023].

[2] Soerjomataram I, Cabasag C, Bardot A, Fidler-Benaoudia MM, Miranda-Filho A, Ferlay J, et al.; on behalf of the SURVCAN-3 collaborators (2023). Cancer survival in Africa, Central and South America, and Asia (SURVCAN-3): a population-based benchmarking study in 32 countries. Lancet Oncol. https://doi.org/10.1016/S1470-2045(22)00704-5. Available from: https://gco.iarc.who.int/survival/survcan/, accessed [12/12/2023].

WHO, where necessary[3]. Depending on the coverage, completeness, and degree of detail of the mortality data available, four methods were utilised:

- For 90 countries where national mortality data were available historically and a sufficient number of recorded cancer deaths were available (at least 150 annually), mortality rates were, as for incidence, projected to 2022 using the short-term prediction models and applied to the 2022 national population estimate (*method 1*)

- When recent mortality data were available from national or subnational sources, the most recent mortality rates from one source within the country (*method 2a*, 3 countries), were used as proxy for 2022.

- For 1 country where recent mortality data were not available, national mortality was estimated from national incidence using statistical models, with the fitted incidence to mortality (I:M) ratios derived from recorded data from cancer registries, with I:M ratios between countries scaled according to levels of HDI (*method 3b*).

- For 91 countries where recent mortality data were not available, national mortality was estimated from national incidence using statistical models, with the fitted incidence to mortality (I:M) ratios derived from survival estimation (method 3c).

The source of information (incidence and mortality) used to estimate the burden of cancer in each country is given in the file GLOBOCAN2022_Annex_A.xlsx file available at https://gco.iarc.who.int/

[3] WHO. World health statistics 2018: monitoring health for the SDGs, sustainable development goals. Available from: https://www.who.int/publications/i/item/9789241565585, accessed [12/12/2023].

Annex B. List of cancer sites included in GLOBOCAN 2022 and criteria for including and allocating certain malignancies.

The 38 cancer sites estimated in GLOBOCAN 2022 and listed below include malignant neoplasms only. An exception is bladder cancer incidence which may include *in situ* carcinomas or tumours of uncertain or unknown behaviour (ICD-10 categories D09.0 and D41.4 respectively) depending on registry practice. The categories “Kaposi sarcoma”, “non-Hodgkin lymphoma” and “All cancers” include some disease entities that have been coded in mortality (but not incidence) statistics to the ICD-10 category B21 (HIV disease resulting in neoplasms). The category “Non-melanoma skin cancer (C44)” (NMSC) excludes basal cell carcinomas (BCC) in incidence, while mortality includes deaths from all types of NMSCs.

1. Lip, oral cavity (C00-06)
2. Salivary glands (C07-08)
3. Oropharynx (C09-10)
4. Nasopharynx (C11)
5. Hypopharynx (C12-13)
6. Oesophagus (C15)
7. Stomach (C16)
8. Colon (C18)
9. Rectum (C19-20)
10. Anus (C21)
11. Liver (including intrahepatic bile ducts C22)
12. Gallbladder (C23)
13. Pancreas (C25)
14. Larynx (C32)
15. Lung (including trachea, C33-34)
16. Melanoma of skin (C43)
17. Non-melanoma skin cancer (C44)
18. Mesothelioma (C45)
19. Kaposi sarcoma (C46)
20. Female breast (C50)
21. Vulva (C51)
22. Vagina (C52)
23. Cervix uteri (C53)
24. Corpus uteri (C54)
25. Ovary (C56)
26. Penis (C60)
27. Prostate (C61)
28. Testis (C62)
29. Kidney (C64)
30. Bladder (C67)
31. Brain, central nervous system (C70-72)
32. Thyroid (C73)
33. Hodgkin lymphoma (C81)
34. Non-Hodgkin lymphoma (C82-86, C88)
35. Multiple myeloma (C90)
36. Leukaemia (C91-95)
37. Other specified cancers (C17, C24, C30-31, C37-38, C40-41, C47-49, C57-58, C63, C65-66, C68-69, C74-75)
38. Unspecified sites (C76-80, C96-97)

39. All cancers (C00-97)

Ill-defined codes

Wherever national or subnational data were available for the following ICD-10 unspecified cancer groupings they were redistributed to specific categories by year, sex and age: C14.0 (pharynx, unspecified), C14.8 (overlapping lesion of lip, oral cavity and pharynx), C26.0 (intestinal tract, part unspecified), C26.8-9 (ill-defined sites within the digestive system), C39 (other and ill-defined sites in the respiratory system), C55 (uterus, unspecified), C57.8-9 (female genital organ, unspecified), C63.8-9 (male genital organ, unspecified), C68.8-9 (urinary organs, unspecified) and C75.8-9 (endocrine glands, unspecified); the three-digit C14 (other and ill-defined sites in the lip, oral cavity and pharynx) and C26 (other and ill-defined digestive organs) were redistributed when the 4th digit categories were not available. Given the large variations in the accuracy of death certificates related to cancer of the uterus, with many deaths recorded as "uterus cancer, not otherwise specified" (ICD-10 C55), these proportions were relocated to specified sites. By default, the number of cancer deaths coded as "uterus unspecified" was reallocated to either cervix (C53) or corpus (C54) uterine cancer according to age-specific proportions in the same population when the all-age proportion of uterine cancer deaths coded to the unspecified category was considered to be low (<25% of the total). For the other countries for which country-specific incidence and survival data were available, mortality for cervix uteri (C53) and corpus uteri (C54) cancers was estimated from incidence and 5-year relative survival probabilities. The total number of cancer deaths from uterine cancers (ICD-10 C53-55) estimated in 2022 were then partitioned into cervix and corpus uteri cancers using the proportions obtained from the survival analysis, and then further stratified by age using age-specific death counts of the two sites (C53 and C54) extracted from the WHO mortality database. No attempt was made to reallocate the 'unspecified cancers' group (ICD-10 categories C76-80+C97) into some specific categories: should a simple reallocation by site, sex and age be performed, it could amplify the over representation of some cancer sites such as screen-detectable cancers in incidence or cancers with possible inclusion of metastatic cancers along with primary neoplasms in mortality.

Redistribution of ill-defined codes (ICD-10 4 digit):

- C14.0 => C09-13
- C14.8 => C00-C13 + C14.2
- C23.6 => C17-21
- C26.8-9 => C15-25 + C26.1
- C39.0, C39.8-9 => C30-34, C37-38
- C57.8-9 => C51-54, C56, C57.0-7
- C63.8-9 => C60-62, C63.0-7
- C68.8-9 => C64-67, C68.0-1
- C75.8-9 => C73-74, C75.0-5

Redistribution of ill-defined codes (ICD-10 3 digit):

- C14 => C00-13
- C26 => C15-25
- C39 => C30-38
- C55 => C53-54

Annex C. Modelling of incidence and mortality

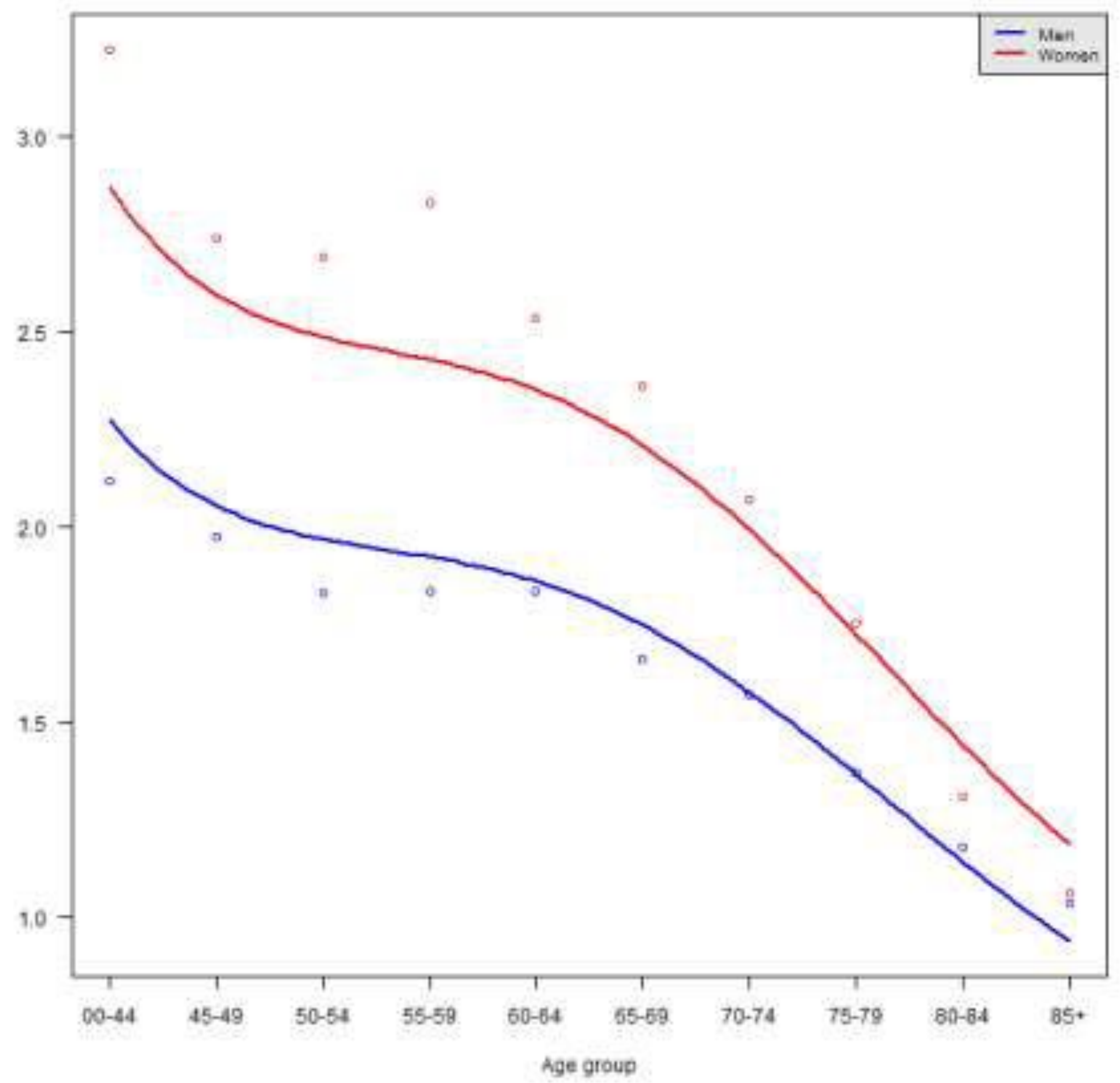
When only incidence of mortality were available, mortality to incidence ratios were used.

mortality to incidence ratios was extrapolate from either:

- method 3a : country-specific mortality to incidence ratios

Country-specific mortality to incidence ratios were used (Bosnia Herzegovina, France (metropolitan), Italy, Germany, Poland, Portugal, Spain and Switzerland).

Example:



- method 3b : regional modelling of mortality to incidence ratios

Regional models established, based upon the incidence and mortality data from population-based cancer registries which supplied data to *Cancer Incidence in Five Continents* Vol. XI.

For each sex and cancer site combination, the mortality to incidence and incidence to mortality ratios
by age groups were scaled using the ratio of the HDI (Human Development Report 2019 (http://hdr.undp.org/) of the country to the mean HDI of the countries in the model.

$I_{National} = M_{National} * (I_{Regional}/M_{Regional}) * (HDI_{National}/HDI_{Regional})$ (incidence method 3b)
$M_{National} = I_{National} * (M_{Regional}/I_{Regional}) * (HDI_{Regional}/HDI_{National})$ (mortality method 3)

When the resulting ratios by age were lower than 1 (incidence method 3b) or greater than 1 (mortality method 3) these were set to 1. The ratios were then fitted using Poisson regression models as in incidence method 3a.

- method 3c : Survival-proxy as mortality to incidence ratios

A proxy of the mortality to incidence ratios were calculated using this formula:

MI ratio = 1 – (5-year relative survival)

Annex D. Computation of the standard error by method of estimation

Uncertainty intervals (95% UI) of the estimated sex- and site-specific number of new cancer cases and cancer deaths for all ages have been computed using the standard error *se* of the crude incidence or mortality rate used in the estimation. The *se* have been calculated on the log scale and the UI back on the arithmetic scale using the following formulae:

$$\mathrm{CR2022}_{pcs}) - 1.96 * se) * \mathrm{P2022}_{ps} / 100{,}000$$
$$\mathrm{UI}_{upper} = \exp\ (\log(\mathrm{CR2022}_{pcs}) + 1.96 * se) * \mathrm{P2022}_{ps} / 100{,}000$$

Where CR2022$_{pcs}$ is the estimated crude incidence/mortality rate per 100,000 in 2022 for country p, cancer c and sex s; P2022$_{ps}$ is the population of country p and sex s in 2022, and *se* the standard error. The standard error *se* should be corrected for three major causes of bias:

1. *Coverage*
2. The *lag time*
3. The *quality*

For sake of simplicity, the three biases have been considered to have the same importance, and a correction based on three categorical variables having the same range of values from 0 (high) to 10 (low) has been introduced:

$$SE = se * 100/(100\text{-}c) * 100/(100\text{-}t) * 100/(100\text{-}q)$$

Where SE is the combined standard error, *se* is the standard error of the crude incidence or mortality rate used in the estimation calculated on a log scale; the categorical variables c describes the coverage of the dataset; t the time lag expressed in year and q describes the quality of the dataset. For each country, these three categorical variables can be sex- and cancer-specific, depending upon the amount and the quality of available data.

- **For method 1**: for prediction of incidence and mortality rates, *se* is given by the model. The coverage (c) and the lag time (t) = 0 (no penalty) because the variance is composed of the variance of the model and of the variance of the projected crude rate. Quality (q) is based on the quality of the dataset (see below).

- **For methods 2**: most recent rates are used as proxy (incidence and mortality), $se = 1/\sqrt{n}$ where n is the number of cases or deaths (all ages) used to compute the crude rate. Because the crude rate is based on the at least three most recent years (see manuscript), n can be very large (e.g. for the Chinese registries, for example) yielding extremely narrow uncertainty intervals, we used the annual number of cases when it was greater than 20 (per year). The coverage (c), the lag time (t) and the quality (q) are defined below.

- **For incidence method 3a** (modelling of M:I ratios using country-specific incidence and mortality data): *se* is the standard error of the crude incidence rate of the pooled incidence data included in the model. The coverage (c), the lag time (t) and the quality (q) are defined below (similar to method 2).

- **For incidence method 3b, 3c and mortality method 3** (modelling of M:I/I:M ratios using incidence and mortality data from neighbouring countries): the standard error for incidence (*i*) and mortality (*m*) (se_i and se_m respectively) is defined as the standard error of the estimated crude mortality rate (*se* (CR2022$_m$), incidence method 3b), or of the crude incidence rate (*se* (CR2022$_i$), mortality method

3). The lag time (t) is defined below and the coverage (c) and the quality (q) = 10 (no data, maximum penalty).

- **For incidence method 4** ('All sites' incidence rates from neighbouring registries partitioned using frequency data): *se* is computed using the annual number of cases in the frequency dataset, the lag time (t) is defined below, the coverage (c) = 10 (no incidence rates) and quality (q) = 8 (poor quality).

- **For method 9** (average of rates from neighbouring countries, incidence and mortality): *se* is defined as the largest standard error of the crude incidence/mortality rate in the neighbouring countries, the lag time (t) as the population weighted average of neighbouring countries lag times, the coverage (c) and the quality (q) = 10 (no data, maximum penalty).

All sites (by country and sex)

se= 1/√n with n total annual number of cancer cases or cancer deaths by sex.
t = max lag time within the 38 specific cancers
q = c = highest values for quality and coverage within the 38 specific cancers

Area/World by sex and site (including "All sites")
se= 1/√n with n = population weighted average of the annual number of cases or deaths by sex and cancer in the countries/areas.
t = population weighted average of the lag times by sex and cancer in the countries/areas.
q = population weighted average of quality by sex and cancer in the countries/areas.
c = population weighted average of coverage by sex and cancer in the countries/areas.

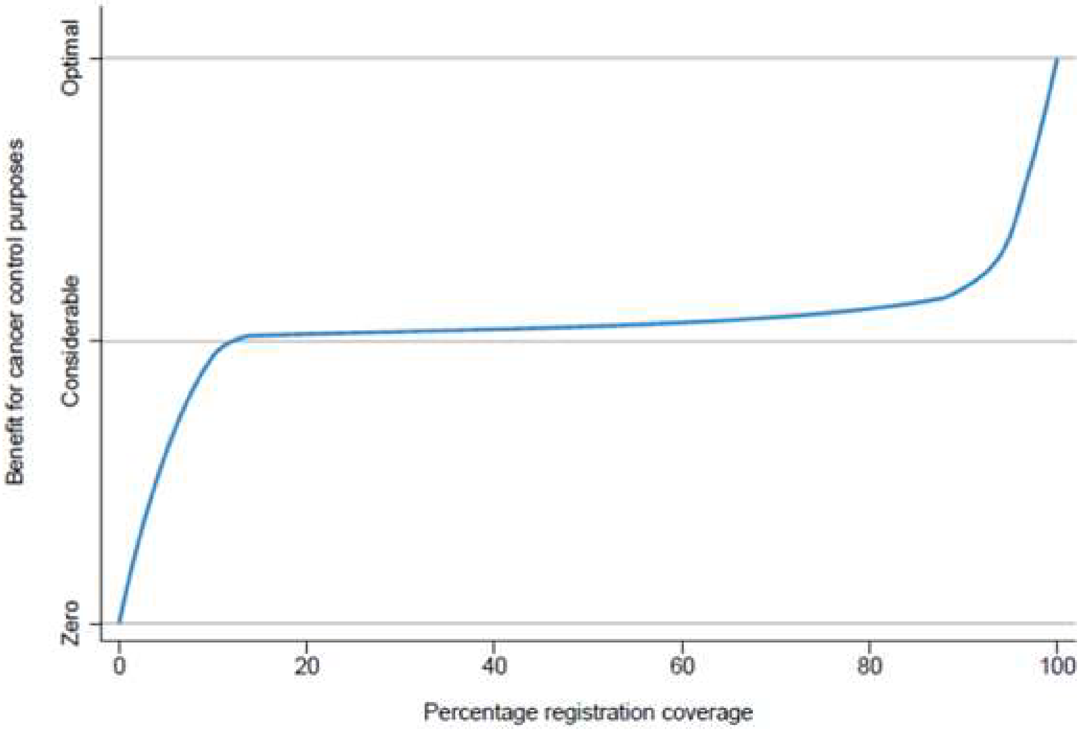
Both sexes combined
$se = \sqrt{se_m^2 + se_f^2}$.
$t = (t_m+t_f)/2$
$q = (q_m+q_f)/2$
$c = (c_m+c_f)/2$

Where *m* is for males and *f* is for females.
Definition of the categorical variables:

Coverage: c is defined as followed:

c=0 if coverage = 100% or if the standard error is obtained from prediction (method 1) because coverage is already taken into consideration in the computation of the variance of the projected crude rate.
c=10 coverage = 0% (no data)
c=9 coverage < 1%
c=8 coverage < 5%
c=7 coverage < 10%
c=1 if coverage > 95%
c=2 coverage > 90%
c=3 coverage > 80%
c=4 coverage > 60%
c=5 coverage > 50%
c=6 otherwise (between 10% and 50%).

Benefits of increasing population coverage by cancer registration.



Source: Planning and Developing Population-Based Cancer Registration in Low- and Middle-Income Settings. IARC Technical Report 43. https://publications.iarc.who.int/Book-And-Report-Series/Iarc-Technical-Publications

Lag time: t is the difference (in year) between the mid-period of the most recent incidence or mortality rates used to compute the sex and site-specific estimates, and the target year (2022). t = 0 when method 1 is used because the lag time is already taken into consideration in the computation of the variance of the projected crude rate.

Quality (incidence): q is based on results from the *Cancer Incidence in Five Continents* Vol. XI (CI5) editorial process.

q=0 if the national or all sub-national registries are included in CI5 without an asterisk.
q=1 if the national or all sub-national registries are included in the last volume of CI5 with an asterisk.
q=2 to 7 are based on the review of the datasets by IARC staff (including Visiting Scientist) during the editorial processes of CI5 and Cancer in Sub Saharan Africa projects.
q=8 if frequency data were used
q=10 if there were no data.

Quality (mortality): q is based on the percentage of garbage codes (GC), as defined in the World health statistics 2017 (monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2017).

q=10 no data
q=0 if GC = 0%

q=1 if GC < 5%
q=2 if GC < 10%
q=3 if GC < 15%
q=4 if GC < 20%
q=5 if GC < 25%
q=6 if GC < 30%
q=7 if GC < 35%
q=8 if GC < 40%
q=9 if GC >= 40%