Nationale Krebsregistrierungsstelle
Organe national d'enregistrement du cancer
Servizio nazionale di registrazione dei tumori
National Agency for Cancer Registration

Foundation
National Institute
for Cancer Epidemiology
and Registration

Methods for the estimation and projection of cancer prevalence

Contents

A. Introduction

Cancer prevalence is defined as the number of persons alive at the date of reference (the index date) and who have previously been diagnosed with cancer. We updated and extended our previously published prevalence estimates [Lorez et al. 2014] for index dates ranging from 31.12.2010 to 31.12.2025.
The cancer prevalence proportion is the number of prevalent persons divided by the population at risk at index date, usually expressed as fraction of 100'000 or as percentage. Persons are stratified by sex and age attained at the index date.

Limited-duration prevalence represents the number of prevalent persons who had a diagnosis of cancer within the past x years (e.g. x = 1, 2, 5 or 10 years). For example, the ten-year limited-duration prevalence for 31.12.2005 comprises all persons alive at that date with a diagnosis within 1.1.1996 to 31.12.2005. Limited-duration prevalence is accessible by a simple counting method and is described below (chapter B).

Complete or total prevalence on the other hand, is an estimate of the number of prevalent persons who had been diagnosed with cancer, no matter how long ago that diagnosis was. Complete prevalence estimation typically requires statistical modelling to estimate the number of survivors diagnosed before the starting date of cancer registration.

The one- and two-year limited-duration prevalence is relevant for assessing current demands on public health services regarding cancer diagnosis, staging, primary treatment and supportive care for recovery from its effects. Prevalence at two to five years after diagnosis likely includes patients under close clinical assessment for recurrence. On the other hand, the five to ten-year limited duration prevalence, or higher, includes many persons who no longer receive cancer related treatment but may be utilizing services for late and long-term effects of their cancer diagnosis and treatment.

B. Counting prevalent persons for index dates ≤ 31.12.2016.

B.1. Data sources.

Pseudonymized individual cancer diagnoses were taken from the National Cancer Dataset managed by the Foundation National Institute for Cancer Epidemiology and Registration (NICER) for the purpose of national cancer monitoring in Switzerland. Data from 11 Swiss cantons registering cancer cases with diagnosis year at least 1999 up to 2016, and providing vital status follow-up until 2016, were eligible for this study.

The predominantly German-speaking part of Switzerland (SA) was represented by six cantons: Appenzell Innerrhoden (AI), Appenzell Ausserrhoden (AR), Glarus (GL), Grison (GR), St. Gallen (SG), and Zurich (ZH). The predominantly French/Italian-speaking part of Switzerland (SR) was represented by five cantons: Jura (JU), Neuchâtel (NE), Ticino (TI), Valais (VS), and Vaud (VD). Thus, the SA region of Switzerland was covered by about 40%, and the SR region by about 70%.

[Note that canton JU as part of cancer registry of NE/JU started registration only in 2005. JU contributes about 2% of Swiss cases. The 10-ye prevalence estimates will thus be slightly underestimated for index dates < 2014.]

B.2. Completeness of case ascertainment.

Completeness of case ascertainment had been previously assessed by the Flow method, and the method of comparing mortality to incidence ratio with survival proportions, without detecting signs of overt under-registration [Lorez et al. 2017].

B.3. Extrapolation to whole Switzerland.

Prevalence counts specific for cancer type, attained age, sex and language-region (since we cannot exclude regional differences in cancer risk and survival), were extrapolated from observed partially covered language regions to full language regions. Whole Switzerland prevalence counts were estimated as the sum of extrapolated SA and SR language regions.
Prevalence proportions for whole Switzerland were estimated specific for cancer type, attained age, and sex as the weighted sum of language-region specific proportions. As weight, the population ratio of language region with whole Switzerland was used (about 0.7 for SA, and 0.3 for SR, depending on age and sex).

B.4. Case inclusion/exclusion criteria.

Diagnoses had been coded originally according to the "International Classification of Disease for Oncology" (ICD-O Third Edition) [IARC 2004], checked for validity and consistency using the JRC-ENCR quality check software v1.8.1, developed by the European Commission Joint Research Centre (http://www.encr.eu/index.php/downloads/jrc-encr-qcs). Whether a diagnosis was primary was decided according to international rules issued by IARC, IACR and ENCR (http://www.iacr.com.fr/MPrules_july2004.pdf). Conversion from ICD-O-3 into ICD-10 ("International Statistical Classification of Diseases and Related Health Problems", 10th Revision) was done using IARCcrgTools version 2.13 (http://www.iacr.com.fr/index.php?option=com_content&view=category&layout=blog&id=68&Itemid=445).

Persons with multiple primary malignant cancer diagnoses in different cancer reporting groups were included separately in each cancer group. For persons with multiple primary malignant diagnoses in a single cancer group, we included only the first diagnosis. The prevalence estimate for all cancer types combined excluded diagnoses for non-melanotic skin cancer (ICD-10 code C44) and counted the first primary malignant diagnosis in the respective diagnosis interval (1, 2, 5, 10 years).

For DCO (death-certificate-only) cases, the diagnosis date is unknown. Such cases are infrequent in Swiss cancer registration (<5%) for the majority of sites, with the exception of hepatic or pancreatic cancer [Lorez et al. 2017]. We excluded DCO cases for prevalence estimation.

B.5. Correction for vital status lost to follow-up or missing active follow-up.

For person's who are lost to follow-up or are without active follow-up, the vital status at certain index dates is unknown. The probability P_Δ of each patient still being alive at the index date, conditional on the length of observed survival $P(fu)$, was estimated as

$$P_\Delta = P(fu + \Delta)/P(fu) \quad (1)$$

with Δ being the time from last follow-up to the index date. These survival probabilities were estimated by flexible parametric models of observed survival using the *stpm2* command written for the Stata® language [Royston and Lambert 2011]. In brief, flexible parametric models are modified Weibull models where the time after diagnosis is incorporated as restricted cubic spline function. Cancer registry, sex and age at diagnosis were used as covariates. We assumed proportional hazards (i.e. the relative effects of covariates were constant in time after diagnosis). Non-parametric Kaplan-Meier estimation was used to visually validate the survival functions derived with flexible parametric modelling.

Complete type survival analysis was performed for each index date by selecting a cohort of patient with diagnosis within ten years before that date.
Cases lost to follow-up were then selected according to the probability to be alive at index date P_Δ. The expected survivors were added to those cases observed alive at index date.

B.6. Estimating confidence intervals (CIs) for prevalence counts and proportions.

Clegg et al. (2002) and Gigli et al. (2006) approximated the sum of patients observed alive and those expected alive among those lost to follow-up (see B.5.) as Poisson distributed.
Exact Poisson 95% CIs are given for the sum of observed and expected prevalence counts. The variance of extrapolated prevalence counts for Swiss language regions, and whole Switzerland, (see B.3), was derived with error propagation rules for uncorrelated errors. 95% CIs were calculated based on Wald limits of log-transformations.
The variance of prevalence proportion in language-regions was estimated by

$$Var(P(x)) = \frac{C_1(x) + C_2(x)}{N(x)^2} \quad (2)$$

where $P(x)$ is the prevalence proportion in age-class x, $C_1(x)$ are the number of cases in age class x observed alive, $C_2(x)$ are the number of cases in age class x expected alive and $N(x)$ is the general population in age class x. CIs were derived by relying on the relationship between Poisson and Chi-Square distribution (Johnson and Kotz, 1969). Lower (P_L) and upper (P_U) CIs at levels 1-α are derived as

$$P_L = \chi^{-2}_{2(c_1+c_2),\alpha/2} \big/ 2N \quad (3)$$

$$P_U = \chi^{-2}_{2(c_1+c_2+1),1-\alpha/2} \big/ 2N \quad (4)$$

where c_1+c_2 is the observed value of C_1+C_2 and $\chi^{-2}_{f,p}$ is the pth quantile of the Chi-Square distribution with f d.f.

The variance of derived prevalence proportions for whole Switzerland (see B.3) was derived with error propagation rules for uncorrelated errors. 95% CIs were calculated based on Wald limits of log-transformations.

B.7. Quality of passive and active vital status follow-up.

All registries performed passive follow-up via annual linkage to the official vital statistics. Active follow-up encompasses regular assessment of the vital status of each registered person. Completeness of active follow-up as of 31.12.2016 was different between registries. Active follow-up for all cases was provided by cantons AI, AR, GE, SG, and TI. The most recent available follow-up date was sometime before 31.6.2016 in GL and GR (17% of cases), JU and NE (6%), VD (32%), VS (35%), and ZH (20%). Such cases without recently updated vital status were treated as lost to follow-up. To cases without any active follow-up information, a survival time of 1-day was assigned after which the case was considered lost to follow-up.

B.8. Population at risk.

End- and mid-year populations at completed age (i.e. age at last birthday) for 1981-2016 as well as predictions for future end-year populations 2017-2025, stratified by canton, age and sex were provided by the Swiss Federal Statistical Office (SFSO) at STAT-TAB: "Die interaktive Statistikdatenbank".
http://www.bfs.admin.ch/bfs/portal/de/index/infothek/onlinedb/stattab.html . For predictions, we employed the "reference" scenario for future growth (reference code: px-x-0104020000_101; accessed 20.3.2020).
Since we presented prevalence proportions for end-year index-dates, end-year populations were used as denominators. For predictions of event rates 2017 to 2025, mid-year populations were employed, derived as the mean of successive end-year populations (see C.1.).

C. Medium-term projection of prevalence index dates 31.12.2017 – 31.12.2025.

We projected data observed until 2016 for 9 years to 2025. Firstly, we estimated the future expected incidence and future expected survival and secondly, we combined both estimates to derive the expected prevalence as suggested in Pisani et al. [2002].

C.1. Incidence projection.

We employed the conventional simplified approach to predict future event rates as depending predominantly on age at diagnosis and calendar period of diagnosis, assuming that rates are similar if different birth cohorts were compared. The relatively short period of projection justified the simplicity of the approach. Stata® commands written by Hakulinen and Dyba [1994] were used for age-period (AP) modelling. In short, the model assumed that incident cases are Poisson distributed and that rates can be extrapolated as simple log-linear or even linear trends (the latter is recommended for increasing trends in order to avoid an explosion of prediction produced by exponential models). For stable or decreasing incidence trends, projections were derived with

$$\ln(c_{it} / n_{it}) = \alpha_i + \beta_i * t \qquad (5)$$

and for increasing incidence trends with

$$c_{it} / n_{it} = \alpha_i * (1 + \beta * t) \qquad (6)$$

where c_{it} is the number of cases in age group i = 1, 2,…, 10 for age-groups (0-9), (10-19),…, (80-89), (90+) and time t for the calendar years, n_{it} is the number of person-years for age-group i and year t, α_i is the age effect, and β_i is the calendar-period effect for age-group i. β is an overall calendar-period effect. The projections take the official predicted population growth in 2017 to 2025 into account (see B.8.). The projection is based on observed trends within the period 2007-2016.

The prediction model for increasing trends (equation 6) does not assume the same constant proportional change in incidence over time in each age group, which is unrealistic. It is allows steeper changes at higher baseline rates, creating a special case of age-specific trend [Dyba et al. 1997].
95% CIs of the prediction interval, which accounts for uncertainty of the slope as well as the random error in future observations, assumed asymptotic normality.

Restricted cubic spline regression was applied to observed and projected incidences in 10-year age-groups for calendar years 2010 to 2025 in order to interpolate them for single years of age, a prerequisite for the method of Pisani et al. [2002] (see C.3.). Knots were placed empirically at ages 5, 15, 35, 45, 65, 75, 85, and 95.

C.2. Survival projection.

The future survival may be estimated by applying the period approach to patients with very recent dates of follow-up [Brenner and Gefeller 1996; Talbäck et al. 2004]. In short, period analysis selects cases by follow-up dates. This is

achieved by left truncation of person-times at risk at the beginning of the specified follow-up period in addition to right censoring at its end. The specified follow-up period encompassed the latest 5 available years in the National Cancer Dataset (2012-2016). The survival functions were estimated with flexible parametric models [Royston and Lambert 2009]. Separate models were fitted for cancer sites and each combination of sex and language area. Models included the covariate age at diagnosis as linear term. Proportional hazard was always assumed. Survival probabilities at 0.5, 1.5, …, 9.5 years after diagnosis for single years of age from age 0 to age 95 were predicted for the Pisani et al. [2002] method (see C.3.).

C.3. Combining incidence and survival.

If future expected incidence counts and survival probabilities have been estimated for single years of age and for each calendar year, the future prevalence counts can be derived as proposed by Pisani et al. [2002] as

$$P_k(n-year) = \sum_{i=1}^{n} \underbrace{I_{k+1-i} * S_{k+1-i}(i-0.5)}_{P_{i,j} \text{ in eq. 8}} \qquad (7)$$

where k is the age-class, n is either the 1-, 2-, 5- or 10-years limited duration period, I_k is the annual incidence for age-class k and S_k is the survival for age-class k. The simplifying assumption is that all incidences occur at mid-year. For example, the 2-year prevalence for age-class k at the end of 2015 would be estimated as expected incidence in age-class k for 2015 times the survival probability *S(0.5 years)* for age-class k plus the expected incidence in age-class k-1 for 2014 times the survival probability *S(1.5 years)* for age-class k-1.

C.4. Estimating the variance and CIs for projected prevalences.

The variance of the incidence counts in 2010 to 2016 was derived from the Poisson assumption, while the variance for the forecasted incidence counts in 2017 to 2025 was derived from the Stata® command written by Hakulinen and Dyba (1994). The variance of survival probabilities was derived from the flexible parametric models of Royston and Lambert (2009). The variance of the prevalence derived with the method of Pisani et al. (2002) was approximated by applying error propagation rules for sums and products of random, uncorrelated and correlated variables as

$$Var[P_k(n-year)] = \sum_{i=1}^{n} (Var[I_{k+1-i}] * S_{k+1-i}(i-0.5)^2 + Var[S_{k+1-i}(i-0.5)] * I^2{}_{k+1-i} + Var[S_{k+1-i}(i-0.5)] * Var[I_{k+1-i}])$$

$$+ 2 * \sum Cov[P_i, P_j] \quad , \text{with } i<j \qquad (8)$$

using the same notation as in eq. 7. We acknowledged that the n summands labelled $P_{i,j}$ in eq. 7 are highly correlated, thus the covariance between these terms was taken into account.

D. Possible sources of errors.

Caution should be applied when using these data for health service planning purposes at a national level or for international comparisons. Prevalence estimates are susceptible to the same biases that affect the estimation of incidence and survival, particularly with respect to those cancers that are not uniformly and rapidly fatal. The observed trends in prevalence estimates can be influenced by changes in data quality and coding conventions, e.g. differences in levels of vital-status loss to follow-up, the proportion of cases known only from death certificates, the migration of cancer patients in and out of the registered population, the diagnosis of multiple cancers in the same person and, perhaps most importantly, the achieved completeness of case ascertainment by cancer registration.

E. References.

Brenner H and Gefeller O (1996). An alternative approach to monitoring cancer patient survival. *Cancer* **78**(9), 2004-2010.

Clegg L, Gail M and Feuer E (2002). Estimating the Variance of Disease-Prevalence Estimates from Population-Based Registries. *Biometrics* **58**, 684-688.

Dyba T, Hakulinen T and Päivärinta L (1997). A simple non-linear model in incidence prediction. *Statistics in Medicine* **16**, 2297-309.

Gigli A, Mariotto A, Clegg LX, Tavilla A, Corazziari I, Capocaccia R, Hachey M, and Scoppa S (2006). Estimating the variance of cancer prevalence from population-based registries. *Statistical Methods in Medical Research*; **15**: 235–253.

Hakulinen T and Dyba T (1994). Precision of incidence predictions based on Poisson distributed observations. *Stat Med* **13**, 1513-23.
Software download available here: http://www.cancer.fi/syoparekisteri/en/general/links/.

IARC (2004). International Rules for Multiple Primaries (ICD-O Third Edition). IARC, IACR & ENCR. Internal Report No. 2004/02. IARC, Lyon.

Johnson NL and Kotz S (1969). Discrete distributions. Houghton-Mifflin, Boston/ J Wiley & Sons, New York.

Lorez M, Heusser R and Arndt V (2014). Prevalence of Cancer Survivors in Switzerland. *Swiss Cancer Bulletin* **4**, 285-289.

Lorez M, Bordoni A, Bouchardy C, Bulliard JL, Camey B, Dehler S, Frick H, Konzelmann I, Maspoli M, Mousavi SM, Rohrmann S, and Arndt V (2017). Evaluation of completeness of case ascertainment in Swiss cancer registration. *European Journal of Cancer Prevention* **26**, 139–146.

Pisani P, Bray F and Parkin M (2002). Estimates of the World-wide prevalence of cancer for 25 sites in the adult population. *Int. J. Cancer* **97**, 72-81.

Royston P and Lambert P (2009). Further development of flexible parametic models for survival analysis. *The Stata Journal*, **9**(2), 265-290.

Royston P and Lambert P (2011). Flexible Parametric Survival Analysis using STATA: Beyond the Cox Model. Stata Press, ISBN-10: 1-59718-079-3.

Talbäck M, Rosen M, Stenbeck M and Dickman PW (2004). Cancer patient survival in Sweden at the beginning of the third millennium – predictions using period analysis. *Cancer Causes and Control* **15**, 967-976.

F. Contact

Matthias Lorez, PhD MAS
National Agency for Cancer Registration (NACR)
National Institute for Cancer Epidemiology and Registration (NICER)
Hirschengraben 82
CH-8001 Zürich
ml@nicer.org
nacr@nicer.og

Zürich, 23.3.2020