Approaches to handle missing follow-up time: A comparative analysis of contralateral breast cancer incidence

Sarah R Haile (1) (0000-0002-4704-6570); Miriam Wanner (2) (0000-0003-0888-0690); Dimitri Korol (2) (0000-0002-9217-8038); Sabine Rohrmann (1, 2) (0000-0002-2215-1200)

1. Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Zurich, Switzerland
2. Cancer Registry of the Cantons of Zurich, Zug, Schaffhausen and Schwyz, Switzerland

Corresponding author: Sarah Haile (sarah.haile@uzh.ch)

Abstract

Purpose: We aimed to compare various approaches for handling missing vital status or follow-up time. As a case study for application of these methods, we estimated incidence of metachronous contralateral breast cancer (CBC).

Methods: For 1980-2016, incidence of metachronous CBC with follow-up through 2024 was estimated using Poisson regression with overdispersion, by age at incidence, year of diagnosis, histology and follow-up period. Missing follow-up time was ignored in the naive approach, simulated once using the average hazard derived from published Swiss cancer registry data, or multiply imputed using 3 different imputation models.

Results: 24,612 women aged 20-84 had unilateral breast cancer between 1980 and 2016 in the Swiss cantons of Zurich and Zug. Of those, 5% (n=1264) were lost to follow-up. Over 291463 person-years, 1145 contralateral breast malignancies were diagnosed, corresponding to 393 per 100000 person-years (95% CI 353 to 438). Incidence rates have been decreasing over time to 238 (171 to 333) for the incidence period 2010-2016. The same overall pattern was observed regardless of how we handled missing follow-up times. However, using a single imputation generally produced lower incidence rates compared to the naive approach, with multiple imputation giving higher estimates. The most complex multiple imputation model gave incidence estimates that were very similar to those from the naive approach.

Conclusion: CBC incidence has declined in recent decades, likely due to increased use of hormone therapies. Different methods to handle missing follow-up times yielded similar results. Multiple imputation is likely an appropriate method to handle missing follow-up data.

Keywords: missing data, multiple imputation, vital status, cancer registry

Introduction

Data collected by cancer registries is necessary to understand which cancers are diagnosed at what frequency, and what happens with people with cancer after the initial diagnosis. Such data aids in answering questions such as how effective cancer screening programs are, and which individuals are likely to be diagnosed with a second primary tumor. To accomplish these goals, cancer registries must ensure the data they collect is accurate and up-to-date. Well-established guidelines on data quality in cancer registries recommend assessing four key indicators: comparability (does data collection follow international guidelines, such as those outlined by the European Network of Cancer Registries (ENCR) (European Network of Cancer Registries 2025)?), validity (Are most diagnoses morphologically verified?), timeliness (in what time frame can cancer diagnoses expect to be available to the registry, and later available in the registry's database?), and completeness (what proportion of all incident cancers in the corresponding population are reported to the registry?) (Bray and Parkin 2009; Parkin and Bray 2009). In Switzerland, the Cancer Registry Zurich and Zug has previously reported on these indicators (Wanner et al. 2018).

One aspect of comparability is certainly having accurate and up-to-date follow-up times and if the patient has died, a date of death (Bray and Parkin 2009, Table 1). This information is especially important for estimating not only absolute and relative overall survival (Brenner and Hakulinen 2009), but also incidence of e.g. second primary cancers. Obtaining correct vital status and date of death varies greatly according to the legal and administrative situation, as it often requires linkage to vital statistics offices. If a unique identifier is not used, such linkage may depend on probability matching. Sometimes death of registered patients has been missed due to incorrect linkage or emigration out of the registry's geographic area, or missing ability of the registry to conduct a (full) linkage, leading to "immortals", that is, individuals that appear to be alive forever in the registry as death information can no longer be obtained (Andersson et al. 2022).

While the situation in Switzerland with respect to such immortal participants has improved for newly registered patients due to the 2020 Cancer Registration Act (Krebsregistrierungsgesetz (KRG) 2023), some older cases remain immortal in the registry due to lack of linkage possibilities and lack of digitized data in earlier years. It is not always clear how to analyze such data. A variety of *ad hoc* methods have been used in analyses of cancer registry data to handle individuals with missing follow-up time, missing vital status or potentially immortal status. While some publications that have likely presented such data do not mention handling of missing data, others use a so-called simulation approach, creating a single imputation based on hazard rates from cancer registry data. Previously, we performed a simulation study assessing how 3 approaches (ignoring missing follow-up times and vital status, single imputation or multi-

ple imputation (Sterne et al. 2011)) worked in typical analysis tasks of a cancer registry where the true value of the outcome was known, and came to the conclusion that multiple imputation was likely the least biased approach (Richter et al. 2024).

In this analysis, we aimed to examine the question of best statistical approach to handle individuals with a cancer registry with a common cancer registry analysis where the true value is not known. Therefore we aimed to update an 2016 analysis of metachronous contralateral breast cancer incidence (Prater et al. 2016), while comparing the results after adjusting for missing vital status or follow-up time in different ways.

Methods

The epidemiological Cancer Registry of the Cantons of Zurich, Zug, Schaffhausen and Schwyz has been been collecting data on cancer cases in the Swiss canton of Zurich since 1980, in the canton of Zug since 2011 and in the cantons of Schaffhausen and Schwyz since 2020. Included in this analysis are the two cantons, Zurich and Zug, that provide at least 5 years of follow-up. Together, Zurich and Zug have a population of around 1.7 million inhabitants and make up about 20% of the total population of Switzerland (Bundesamt für Statistik 2024).

This analysis included all female patients with a unilateral breast cancer from the cantons of Zurich or Zug in the time period from 1980 to 2016, with follow-up until at least 2021. Of the 28,288 patients identified in the registry, 7 were excluded as either younger than 20 or older than 85 years. A further 1135 were excluded because of other malignancies prior to the first breast cancer incidence. To distinguish between synchronous and metachronous cancers, we excluded 1933 participants with follow-up for 90 days or shorter. Finally, 324 subjects were excluded because of missing laterality or a first bilateral breast cancer.

The person-time at risk for the calculation of incidence of contralateral breast cancer (CBC) was counted from 90 days after diagnosis of the initial primary breast cancer until occurrence of CBC or censored at second non-breast cancer, death or end of follow-up period. Quasi-poisson regression to account for overdispersion was performed to estimate incidence rates and incidence rate ratios of metachronous CBC adjusting for age at diagnosis (5-year categories) and year of diagnosis of unilateral breast cancer (UBC, in 5-year periods), histology of UBC (ductal, lobular, other) and follow-up period (3-year intervals). Results were presented as number of cases per 100,000 person-years, and corresponding 95% confidence intervals (95% CI) were obtained.

Several different approaches were considered to fill in the missing follow-up times. In the naive approach (NI), the person-years were left as they were in the original data. In the single im-

putation approach (SI), person-years were simulated once with Swiss breast cancer survival data (publicly available (National Agency for Cancer Registration 2024)). A constant hazard rate of 0.0591 per year for survival after UBC was assumed for the whole period from 1980 to 2021 to compute random survival time, similar to the 0.0591 used in a previous analysis of CBC in this population (Prater et al. 2016). Additionally 3 multiple imputation (MI) approaches were considered. MI model A adjusted for age group, incidence period, morphology and the Nelson-Aalen estimator of the baseline hazard (White and Royston 2009). Considering age and incidence period as continuous variables, MI model B included age, age^2, incidence year, incidence year^2, morphology and baseline hazard. The simplest model, MI model C, included age group, incidence period and morphology only. The 3 MI approaches created 25 imputations each and incidence estimates were pooled using Rubin's rules (Rubin 1987).

Statistical analyses were performed in R (version 4.5.0 (R Core Team 2024)). Multiple imputation was performed using the `mice` package (van Buuren and Groothuis-Oudshoorn 2011).

Results

A total of 24612 women with UBC aged 20-84 years met the inclusion criteria for the present analysis. Patient characteristics are summarized in Table 1. Of the included women, 5.1% (n = 1253) were defined by the cancer registry as being lost to follow-up or had last vital status of alive prior to 2021. Missingness rates by key variables are found in Tables S1 and S2.

Over 291463 person-years, 1145 contralateral breast malignancies were diagnosed, corresponding to an overall incidence rate of 393 per 100,000 person years (95% CI 353 to 437) (Table 2, MI (B) approach). Incidence rates decrease somewhat with age, ranging from 494 (333 to 732) for those diagnosed in their 30s down to 277 (152 to 503) for those diagnoses after the age of 80. Recent decades have also seen a steady decrease in CBC incidence, estimated to be 238 per 100,000 (171 to 333) for those diagnosed between 2010 and 2016. Examining the incidence rates by histology, those with initial ductal histology had the lowest incidence of CBC (376, 329 to 429). Incidence rates were fairly constant over the follow-up period (Figure S1 and Table S3).

Comparing the results by method used to handle missing follow-up times, we can examine both the magnitude of the incidence estimates, as well as the width of the corresponding 95% CIs. The naive approach tended to be 5 to 10 cases per 100,000 person years higher than the other approaches (Table 2 and Figure S2). SI produced the lowest estimates of CBC incidence in all subgroups. The MI models A and C tended to have higher incidence estimates than model B, which appeared closest to the average subgroup incidence. MI model B was in many

subgroups (e.g. by age category) similar to that of the naive approach. Regarding CI width, SI produced narrower CIs in all subgroups, while MI models A and C generally produced the widest CIs (Table 2 and Figure S3).

Discussion

Incidence of metachronous CBC has decreased in recent years, following trends described by e.g. Prater et al. (2016). This decrease in incidence appears to be mostly driven by lower CBC incidence for those women with first BC incidence in recent years, a pattern likely due to the increased use of hormone therapies such as tamoxifen (Early Breast Cancer Trialists' Collaborative Group (EBCTCG) 2011; Nordenskjöld et al. 2023). Younger patients continue to have a somewhat higher risk of CBC than older patients, and those with ductal or lobular histology have lower risk than those with other histologies.

With respect to the choice of approach for handling missing follow-up or vital status, the various approaches yielded by and large similar results. CBC incidence estimates from SI models were generally smaller than those of other methods and had narrower confidence intervals, an expected phenomenon of SI as it produces standard errors which are too small (Sterne et al. 2011). MI (regardless of choice of imputation model) produced estimates of CBC incidence that were generally within $\pm$ 10 units of the naive approach, but with slightly wider confidence intervals to account for the missing follow-up times. While patients with missing follow-up times could be excluded from the analysis when using the naive approach, MI enables them to be included, and permits analysis with other missing covariates as well.

Missing covariate information can lead to exclusion of patients from the statistical analysis, leading to biased estimates of the outcome of interest without further adjustment such as MI (Pitcha et al. 2023). Using MI to account for missing covariates is not new in analysis of cancer registry, however it is often used to fill in missing baseline characteristics such as tumor stage or hormone receptor status (Nur et al. 2010; Eisemann et al. 2011; Howlader et al. 2012; Luo et al. 2017; Parry et al. 2019). It is well-known that missing vital status and missing event times can be imputed (White and Royston 2009; Carpenter et al. 2023). However, we are unaware of assessments of the use of MI in cases where vital status or follow-up date are missing, with the exception of our previous simulation study (Richter et al. 2024). In that simulation study we concluded that MI produced least biased results in a number of typical tasks of cancer registries, such as overall survival, relative survival and standardized incidence ratio. Incorporating MI into otherwise straightforward analysis tasks gives us reduced bias in the estimated outcomes in exchange for somewhat more complex statistical programming. In Richter et al. (2024), we outlined R and Stata code to perform survival analysis after imputation

of missing follow-up times. The *online resource* here outlines R and Stata code to estimate incidence rates in a similar setting.

The main strengths of this analysis are the large dataset from a cancer registry with 35 years of follow-up, and the prespecified analysis plan with many potential approaches to handle missing follow-up data. We were able to replicate CBC incidence rates from a previous CBC analysis (Prater et al. 2016), without having the previous version of the data. A key limitation is that the true CBC incidence rate is not known, so we could not compare the results of the missing data approaches to a true value. Additionally, we did not consider other potential predictors of CBC, such as hormone therapy, type of surgery, use of radiation, lymph node involvement, time from diagnosis to treatment, or genetic predisposition (Akdeniz et al. 2019; Syleouni et al. 2023).

Conclusions

Incidence rates of CBC have been in steady decline in Switzerland since the 1980s, a trend seen in other studies and likely due to increased use of hormone therapies. Multiple imputation can be used to fill in missing follow-up times or vital status, leading to improved data quality in analyses of cancer registry data.

References

Akdeniz D, Schmidt MK, Seynaeve CM, et al (2019) Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. The Breast 44:1–14. https://doi.org/10.1016/j.breast.2018.11.005

Andersson TM-L, Myklbust TÅ, Rutherford MJ, et al (2022) Five ways to improve international comparisons of cancer survival: Lessons learned from ICBP SURVMARK-2. Br J Cancer 126:1224–1228. https://doi.org/10.1038/s41416-022-01701-0

Bray F, Parkin DM (2009) Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. Eur J Cancer 45:747–755. https://doi.org/10.1016/j.ejca.2008.11.032

Brenner H, Hakulinen T (2009) Implications of incomplete registration of deaths on long-term survival estimates from population-based cancer registries. Int J Cancer 125:432–437. https://doi.org/10.1002/ijc.24344

Bundesamt für Statistik (2024) Räumliche Verteilung der ständigen Wohnbevölkerungnach institutionellen Gliederungen

Carpenter JR, Kenward MG, Bartlett JW, et al (2023) Multiple imputation and its application, 2nd edition. Wiley

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: Patient-level meta-analysis of randomised trials. The Lancet 378:771–784. https://doi.org/10.1016/S0140-6736(11)60993-8

Eisemann N, Waldmann A, Katalinic A (2011) Imputation of missing values of tumour stage in population-based cancer registration. BMC Med Res Methodol 11: https://doi.org/10.1186/1471-2288-11-129

European Network of Cancer Registries (2025) ENCR recommendations

Howlader N, Noone A-M, Yu M, Cronin KA (2012) Use of imputed population-based cancer registry data as a method of accounting for missing information: Application to estrogen receptor status for breast cancer. Am J Epidemiol 176:347–56. https://doi.org/10.1093/aje/

kwr512

Krebsregistrierungsgesetz (KRG) (2023) Cancer Registration Act (Bundesgesetz über die Registrierung von Krebserkrankungen)

Luo Q, Egger S, Yu XQ, et al (2017) Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data. PLOS ONE 12:1–16. https://doi.org/10.1371/journal.pone.0180033

National Agency for Cancer Registration (2024) Cancer survival in Switzerland 1982 - 2021

Nordenskjöld A, Fohlin H, Rosell J, et al (2023) Breast cancer survival and incidence of second primary cancers after 30 years in a randomized study of two versus five years of adjuvant tamoxifen therapy. The Breast 71:63–68. https://doi.org/10.1016/j.breast.2023.07.010

Nur U, Shack LG, Rachet B, et al (2010) Modelling relative survival in the presence of incomplete data: A tutorial. Int J Epidemiol 39:118–128. https://doi.org/10.1093/ije/dyp309

Parkin DM, Bray F (2009) Evaluation of data quality in the cancer registry: Principles and methods part II. completeness. Eur J Cancer 45:756–764. https://doi.org/10.1016/j.ejca.2008.11.033

Parry MG, Sujenthiran A, Cowlin TE, et al (2019) Imputation of missing prostate cancer stage in English cancer registry data based on clinical assumptions. Cancer Epidemiol 58:44–51. https://doi.org/10.1016/j.canep.2018.11.003

Pitcha JK, Rushing CN, Lewis HC, et al (2023) Implications of missing data on reported breast cancer mortality. Epidemiology 197:177–187. https://doi.org/10.1007/s10549-022-06764-4

Prater J, Valeri F, Korol D, et al (2016) Incidence of metachronous contralateral breast cancer in the canton of Zurich: A population-based study of the cancer registry. J Cancer Res Clin Oncol 142:365–371. https://doi.org/10.1007/s00432-015-2031-1

R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Richter C, Wildisen L, Rohrmann S, Haile SR (2024) Better tools for better estimates: Improving approaches to handling missing data in Swiss cancer registries. Eur J Cancer Prev 33:400–

406. https://doi.org/10.1097/CEJ.0000000000000881

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley

Sterne JAC, White IR, Carlin JB, et al (2011) Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ 338:b2393. https://doi.org/10.1136/bmj.b2393

Syleouni M-E, Karavasiloglou N, Manduchi L, et al (2023) Predicting second breast cancer among women with primary breast cancer using machine learning algorithms, a population-based observational study. Int J Cancer 153:932–941. https://doi.org/10.1002/ijc.34568

van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45:1–67. https://doi.org/10.18637/jss.v045.i03

Wanner M, Matthes KL, Korol D, et al (2018) Indicators of Data Quality at the Cancer Registry Zurich and Zug in Switzerland. BioMed Research International 7656197. https://doi.org/10.1155/2018/7656197

White IR, Royston P (2009) Imputing missing covariate values for the Cox model. Stat Med 28:1982–1998. https://doi.org/10.1002/sim.3618

Statements & Declarations

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Author contributions Study concept and design was developed by SH and SR. Data collection was performed by MW and DK. Statistical analysis was performed by SH. The first draft of the manuscript was written by SH, and all authors provided feedback on the manuscript prior to submission. All authors read and approved the final manuscript.

Data availability Data analyzed in the current study are not publicly available. A request for anonymized data can be submitted to the National Agency for Cancer Registration. The request procedure is described at https://nacr.ch/en/information-for-researchers/request-for-anonymised-data.

Ethics approval This is a data quality study of previously collected cancer registry data. No ethical approval is required.

Consent to participate Patients with new cancer diagnoses in Switzerland have the right to information about cancer registration, have the right to view their data, and have the right to object to registration (https://nacr.ch/en/information-for-patients). Such an objection can be submitted or withdrawn at any time, and is valid throughout the patient's life and across Switzerland. Personal details, such as name, address, date of birth, and social insurance number are not forwarded to the NACR.

Consent to publish The manuscript contains no individual patient data.

Table 1: Characteristics of included and excluded participants in the cancer registries from the Swiss cantons of Zurich and Zug

Characteristic	**included** N = 24612	**excluded** N = 3676
age at incidence	61 (20-85)	67 (2-85)
year of incidence	2002 (1980-2016)	1995 (1980-2016)
BC diagnosis	24612 (100%)	2534 (69%)
laterality		
unilateral	24612 (100%)	2370 (94%)
bilateral	0 (0%)	164 (6.5%)
Unknown	0	1142
histology		
ductal	17357 (71%)	1404 (38%)
lobular	2808 (11%)	245 (6.7%)
other	4447 (18%)	2027 (55%)
grade		
I	2141 (8.7%)	270 (7.3%)
II	10347 (42%)	1015 (28%)
III	8300 (34%)	910 (25%)
IV	26 (0.1%)	9 (0.2%)
T-cells	0 (0%)	2 (<0.1%)
B-cells	0 (0%)	52 (1.4%)
unknown or unclear	3798 (15%)	1418 (39%)
number of CBC cases	1145	671
total person-years	291463	17825

[1] Median (Min-Max); n (%); Sum

Table 2: estimated CBC incidence per 100,000 person years, by incidence period, age category, and histology as well as by approach to handle missing follow-up time

stratum	naive	SI	MI (A)	MI (B)	MI (C)
Overall					
overall	393 [353 to 437]	373 [335 to 416]	396 [354 to 442]	393 [353 to 438]	396 [355 to 442]
Incidence period					
1980-1984	612 [456 to 822]	558 [412 to 757]	587 [435 to 792]	589 [438 to 792]	588 [437 to 792]
1985-1989	561 [411 to 765]	529 [382 to 732]	560 [407 to 771]	560 [408 to 768]	560 [408 to 769]
1990-1994	433 [332 to 564]	417 [319 to 545]	441 [337 to 575]	438 [336 to 570]	439 [334 to 578]
1995-1999	424 [327 to 549]	397 [305 to 517]	410 [315 to 535]	409 [315 to 532]	410 [314 to 534]
2000-2004	359 [274 to 471]	354 [269 to 465]	374 [283 to 493]	371 [283 to 486]	375 [286 to 492]
2005-2009	303 [228 to 402]	289 [218 to 384]	317 [236 to 426]	311 [234 to 413]	317 [235 to 426]
2010-2016	229 [163 to 321]	222 [157 to 313]	242 [170 to 343]	238 [171 to 333]	241 [171 to 341]
Age at incidence					
20-29	498 [114 to 2184]	486 [109 to 2171]	526 [122 to 2266]	527 [120 to 2314]	524 [122 to 2243]
30-39	477 [323 to 706]	457 [307 to 680]	497 [333 to 742]	494 [333 to 732]	495 [334 to 735]
40-49	430 [342 to 542]	407 [321 to 516]	434 [343 to 550]	429 [340 to 542]	428 [338 to 543]
50-59	417 [338 to 514]	397 [321 to 490]	419 [337 to 520]	417 [338 to 514]	424 [343 to 523]
60-69	360 [293 to 442]	346 [280 to 427]	360 [292 to 444]	358 [290 to 441]	360 [291 to 444]
70-79	333 [244 to 454]	313 [227 to 433]	332 [242 to 456]	332 [242 to 455]	333 [242 to 458]
80-84	278 [153 to 506]	256 [138 to 475]	282 [155 to 512]	277 [152 to 503]	283 [156 to 512]
Histology					
ductal	372 [326 to 424]	356 [312 to 407]	379 [331 to 433]	376 [329 to 429]	379 [331 to 433]
lobular	412 [306 to 554]	391 [288 to 532]	409 [302 to 555]	406 [299 to 550]	408 [301 to 554]
other	462 [364 to 587]	428 [336 to 545]	455 [358 to 577]	454 [358 to 576]	456 [360 to 577]

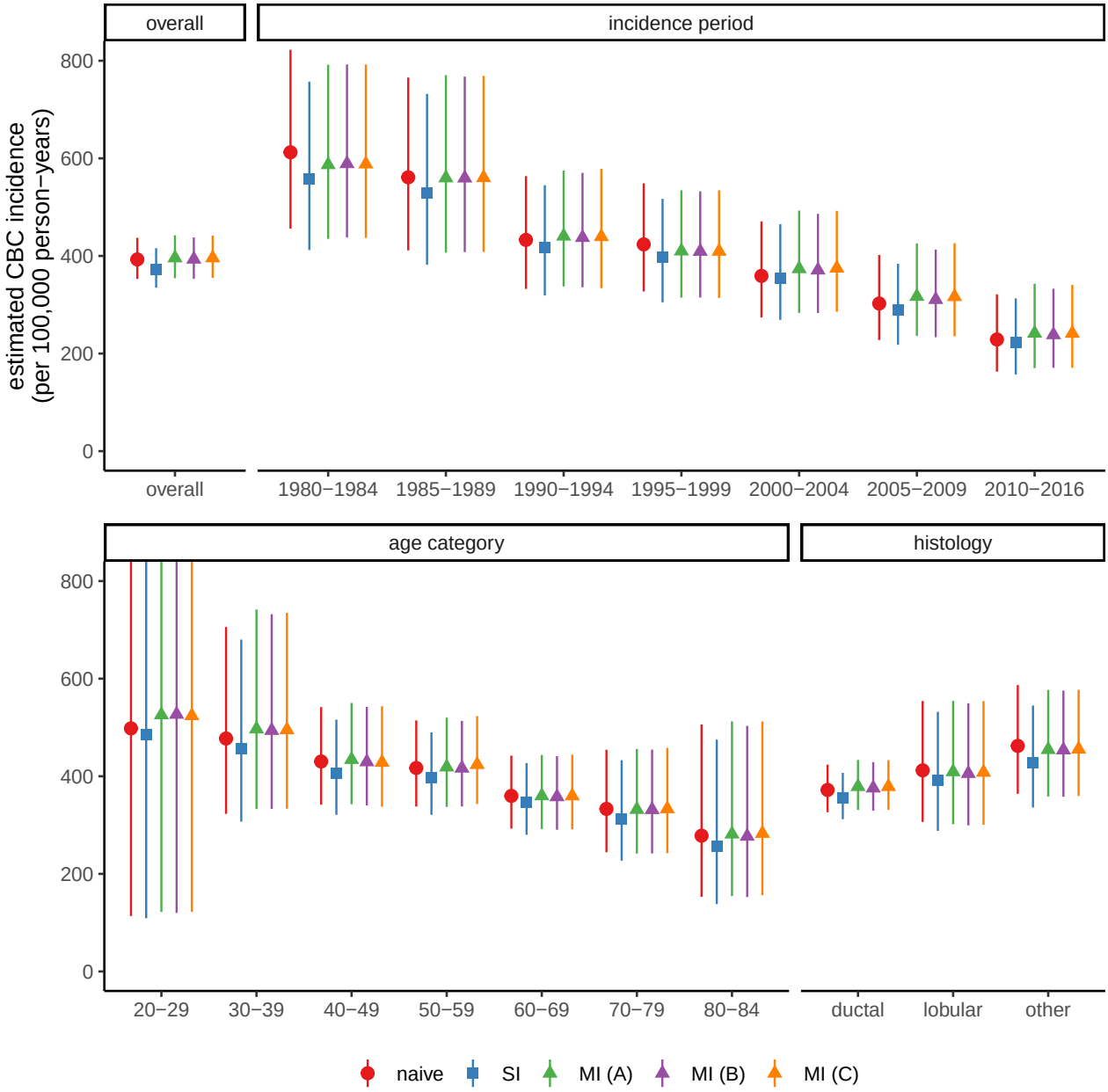

Figure 1: Estimated CBC incidence per 100,000 person-years, overall, and by incidence period, age category and histology. Primarily, MI (B, in purple) estimates are discussed in the text.