



**QUEEN'S
UNIVERSITY
BELFAST**

The impact of excluding or including Death Certificate Initiated (DCI) cases on estimated cancer survival: A simulation study

Andersson, T. M.-L., Myklebust, T. Å., Rutherford, M. J., Møller, B., Soerjomataram, I., Arnold, M., Bray, F., Parkin, D. M., Sasieni, P., Bucher, O., De, P., Engholm, G., Gavin, A., Little, A., Porter, G., Ramanakumar, A. V., Saint-Jacques, N., Walsh, P. M., Woods, R. R., & Lambert, P. C. (2021). The impact of excluding or including Death Certificate Initiated (DCI) cases on estimated cancer survival: A simulation study. *Cancer Epidemiology*, 71(Pt A), 101881. Advance online publication. <https://doi.org/10.1016/j.canep.2020.101881>

Published in:
Cancer Epidemiology

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

Copyright 2021 Elsevier.

This manuscript is distributed under a Creative Commons Attribution-NonCommercial-NoDerivs License

(<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access

This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

1 The impact of excluding or including Death Certificate Initiated (DCI) cases on estimated cancer
2 survival: a simulation study

3
4 Therese M-L Andersson*¹, Tor Åge Myklebust^{2,3}, Mark J Rutherford^{4,5}, Bjørn Møller², Isabelle
5 Soerjomataram⁵, Melina Arnold⁵, Freddie Bray⁵, D Maxwell Parkin ^{5,6}, Peter Sasieni⁷, Oliver
6 Bucher⁸, Prithwish De⁹, Gerda Engholm¹⁰, Anna Gavin¹¹, Alana Little¹², Geoff Porter¹³,
7 Agnihotram V Ramanakumar¹⁴, Nathalie Saint-Jacques¹⁵, Paul M. Walsh¹⁶, Ryan R Woods¹⁷,
8 Paul C Lambert^{1,4}

9
10 1. Department of Medical epidemiology and Biostatistics, Karolinska Institutet, Stockholm
11 Sweden.

12 2. Cancer Registry of Norway, Institute of Population-based Cancer Research, Oslo, Norway

13 3. Department of Research and Innovation, Møre and Romsdal Hospital Trust, Ålesund, Norway

14 4. Department of Health Sciences, University of Leicester, Leicester, United Kingdom.

15 5. Cancer Surveillance Section, International Agency for Research on Cancer (IARC/WHO),
16 Lyon, France.

17 6. Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

18 7. Faculty of Life Sciences & Medicine, School of Cancer & Pharmaceutical Sciences, Guys
19 Cancer Centre, Guys Hospital, King's College London, London, United Kingdom.

20 8. Department of Epidemiology and Cancer Registry, CancerCare Manitoba, Winnipeg, MB,
21 Canada

22 9. Analytics and Informatics, Ontario Health (Cancer Care Ontario), Toronto, Ontario, Canada

23 10. Surveillance and Pharmacoepidemiology, Danish Cancer Society Research Center,
24 Copenhagen, Denmark

- 25 11. Northern Ireland Cancer Registry, Queen’s University Belfast, Northern Ireland, United
26 Kingdom.
- 27 12. Cancer Institute NSW, Alexandria NSW, Australia
- 28 13. Canadian Partnership Against Cancer, Toronto, Ontario, Canada.
- 29 14. Research-Institute, McGill University Health Center, Montreal, Quebec, Canada
- 30 15. Nova Scotia Health Authority Cancer Care Program, Registry & Analytics, Halifax, Nova
31 Scotia, Canada
- 32 16. National Cancer Registry Ireland, Cork, Ireland
- 33 17. Cancer Control Research, BC Cancer, Vancouver, British Columbia, Canada

34

35 Corresponding author: Therese M-L Andersson. E-mail: therese.m-l.andersson@ki.se

36

37 Key words: Cancer registry; Death certificate initiated cases; Survival; Simulation study

38

39 Word count: Abstract 248, Main text 3308, Highlights 65.

40

41 Abbreviations: DCN – death certificate notified; DCI – death certificate initiated; DCO – death
42 certificate only; ICBP – International Cancer Benchmarking Partnership; RR – relative risk; RSR
43 – relative survival ratio

44

Highlights:

- This simulation study shows that including cases initiated through death certificates in the survival analysis of population-based registry data will downwardly bias relative survival estimates.
- Excluding cases initiated through death certificates will in most situations overestimate survival.
- The extent of the bias depends on how missed cases differ from those registered through other routine sources.
- Registries should report the DCI proportion alongside the DCO proportion.

46 **Abstract**

47 Background: Population-based cancer registries strive to cover all cancer cases diagnosed within
48 the population, but some cases will always be missed and no register is 100% complete. Many
49 cancer registries use death certificates to identify additional cases not captured through other
50 routine sources, to hopefully add a large proportion of the missed cases. Cases notified through
51 this route, who would not have been captured without death certificate information, are referred
52 to as death certificate initiated (DCI) cases. Inclusion of DCI cases in cancer registries increases
53 completeness and is important for estimating cancer incidence. However, inclusion of DCI cases
54 will generally lead to biased estimates of cancer survival, but the same is often also true if
55 excluding DCI cases. Missed cases are probably not a random sample of all cancer cases, but
56 rather cases with poor prognosis. Further, DCI cases have poorer prognosis than missed cases in
57 general, since they have all died with cancer mentioned on the death certificates.

58 Methods: We performed a simulation study to estimate the impact of including or excluding DCI
59 cases on cancer survival estimates, under different scenarios.

60 Results: We demonstrated that including DCI cases underestimates survival. The exclusion of
61 DCI cases gives unbiased survival estimates if missed cases are a random sample of all cancer
62 cases, while survival is overestimated if these have poorer prognosis.

63 Conclusion: In our most extreme scenarios, with 25% of cases initially missed, the usual practice
64 of including DCI cases underestimated 5-year survival by at most 3 percentage points.

65 **1. Introduction**

66
67 Cancer survival, when estimated from population-based cancer registry data, is an important
68 measure of the overall effectiveness of health systems given it estimates the average prognosis of
69 cancer patients in the entire population. When comparing population-based cancer survival
70 estimates between countries or jurisdictions, there has been some debate on how differences in
71 registration processes and practices affect the observed survival differences (1). Previous studies
72 have investigated different aspects, including the impact of: i) a failure to link cancer cases to
73 their death information; ii) missing long-term survivors; iii) cancer cases notified from death
74 certificates and iv) finding a date of recurrence instead of a date of diagnosis (2-5).

75
76 In this paper we focus on the impact on estimated survival of including or excluding cases
77 notified through death certificates. Many cancer registries periodically receive notifications of
78 cancer diagnoses based on death certificates, usually denoted as death certificate notified (DCN)
79 cases (6, 7). For a majority of these DCN cases, the registry will also receive a notification from
80 another source (e.g. pathology or hospital records). Yet for some cases, no additional
81 notifications will be received, indicating these cases would not have been known to the registry
82 were it not for the use of death certificate information. These cases are therefore not reported to
83 the cancer registry when diagnosed.

84
85 For the DCN cases with no other notification to the registry, trace-back is often performed to
86 actively ascertain when the cancer was first diagnosed and to verify that the case was a reportable
87 cancer. The subset of DCN cases deemed reportable are referred to as DCI (death certificate
88 initiated) cases (6, 7). DCI cases are therefore cases that are included in the cancer register solely

89 due to the use of death certificate notification, and would not have been reported from another
90 source. DCI cases can be further subdivided into cases where trace-back was successful in
91 finding a date of diagnosis and cases where the trace-back did not yield any additional
92 information. The latter cases are commonly referred to as death certificate only (DCO) cases, and
93 they are a subset of the DCI cases (6, 7). Some registries receive death certificate information
94 more rapidly than notifications through other routine sources and therefore have a large group of
95 cases initially notified from death certificates. However, these cases should not be referred to as
96 DCI cases since they are reported to the registry through independent routine sources although at
97 a later time. Only cases that would not have been known to the registry, if it was not for the death
98 certificate, are DCI cases.

99

100 While it is important for cancer registries to include DCI cases to increase the completeness of
101 cancer incidence statistics, including DCI cases when estimating survival will generally lead to
102 biased results. The existence of DCI cases indicates that there are cases in the population who are
103 not notified to the registry through the course of their disease and who are either alive, or have
104 died without cancer mentioned as a cause of death. This is illustrated in Figure 1, the interest is in
105 the survival of all cancer cases, i.e. the yellow box. However, some cancer cases are not
106 registered through routine sources, and missed by the registry at diagnosis, represented by the
107 grey solid box in Figure 1. A cancer registry that does not perform trace-back only includes the
108 cases in the green solid box, those that are registered through routine sources. Some of these
109 individuals will be alive at the time the cancer registry performs the survival analysis, some will
110 have died with cancer mentioned on the death certificate and some will have died due to other
111 causes, but all these cases are included. In the unlikely situation that these cases are a random
112 sample of those in the yellow box this should yield unbiased estimates of survival. When a cancer

113 registry receives DCN cases, performs trace-back and then include the DCI cases, a subset of the
114 missed cases are also included (the box with light green borders), the subset who died due to
115 cancer (or where cancer is mentioned on the death certificate). The cases missed (not notified
116 through routine sources, the solid grey box) that are still alive or died without cancer mentioned
117 on the death certificate will not be retrieved, and continue to be missed by the registry. Since the
118 DCI cases are not a random sample of the cases missed (solid grey box) by the registry, the
119 inclusion of DCI cases when estimating survival can give biased results, even if the whole group
120 of missed cases are a random sample of all cancer cases. The problem can be illustrated in a
121 simple way by considering all cause survival among 1000 individuals. If the survival probability
122 at 5 years is 0.8 and there is no censoring, one would expect there to be (800 people alive at 5-
123 years ($800/1000=0.8$). If 20% of cases were initially missed (at random) then there would be 800
124 individuals initially with $800*0.8=640$ alive at 5 years ($640/800=0.8$). Of those missed, one
125 would expect $200*0.2=40$ to die. Including these in the analysis leads to a 5-year survival of
126 ($640/(800+40) =76.2%$, i.e. an underestimate as we have only added individuals to the
127 denominator. There is often concern with respect to the validity of data from those registries
128 unable to use death certificates to find additional cases, since it is known that excluding DCI
129 cases will usually overestimate survival. However, the converse - that including DCI cases almost
130 always underestimates survival is often not recognised.

131
132 The International Cancer Benchmarking Partnership (ICBP) SURVMARK-2 study aims to
133 quantify disparities in cancer survival across high-income countries and identify possible reasons
134 for them. As part of this international partnership, we performed a simulation study using a range
135 of scenarios to quantify the impact on estimated cancer survival of including or excluding DCI

136 cases. The overarching aim was to comprehensively understand the potential impact of this bias
137 on benchmarking cancer survival across populations.

138

139 **2. Methods**

140 To investigate the impact of including or excluding DCI cases on survival estimates, we
141 simulated cohorts of 5000 cancer patients. For each cancer patient, a time of death due to cancer
142 and a time of death due to other causes was simulated (8), and for each individual, their cause of
143 death was determined by the event that occurred first: either death due to cancer, or death due to
144 other causes. All survival times were censored at 10 years. We used three separate Weibull
145 distributions for simulating time to death, representing a cancer site associated with low (Weibull
146 parameter $\lambda=0.61$ and $\gamma=0.63$), medium ($\lambda=0.4$ and $\gamma=0.6$) and high ($\lambda=0.12$ and $\gamma=0.64$) cancer-
147 specific survival, since the bias we wish to investigate can depend on the underlying cancer
148 survival. We also used two levels (high and low, roughly corresponding to the survival of a 65
149 and an 80 year old in UK) of other cause (expected) survival, also with Weibull distributions
150 ($\lambda=0.034$; $\gamma=1.25$ and $\lambda=0.13$; $\gamma=1.19$, respectively), since this can have an additional impact on
151 the bias. The survival and hazard functions for both cancer-specific and other cause survival are
152 shown in the Appendix Figure A1.

153

154 2.1 Simulating randomly missed cases

155 We simulated the proportion of the cancer cases who were missed, i.e. not notified to the registry,
156 except possibly from death certificates, first assuming that these were a random sample of all
157 cases. Three levels of missingness were investigated: 5%, 15% and 25%. This gave a total of 18
158 simulated scenarios: 3 levels of cause-specific survival, 2 levels of other cause survival and 3
159 levels for proportions of cases not reported to the registry, as listed in Table 1. Within each

160 simulated scenario, cases who were simulated to be missed by the registry and who died due to
161 cancer within 10 years from diagnosis were classified as DCI cases. For simplicity, we assumed
162 that the trace-back procedure found the correct date of diagnosis for all DCI cases, and hence
163 there were no DCO cases. In actual registry data, DCO cases will exist, and they are usually
164 excluded from survival analysis since their survival time is not known. This might have
165 implications for the extent of bias in our simulations, however the direction of the bias is not
166 altered.

167

168 2.2 Simulating non-randomly missed cases

169 We added another layer to the 18 base scenarios to investigate the impact of including a
170 prognostic factor for death that is related to the extent of missingness. This prognostic factor was
171 represented by a binary variable X (e.g. advanced stage), that affected the time to death due to
172 cancer. The effect of Factor X was assigned a hazard ratio (HR) of 4, meaning that patients with
173 the prognostic factor had a four times higher cancer-specific mortality rate than patients who did
174 not have Factor X. Assuming 25% of the patients had this prognostic factor, we then simulated
175 the 18 base scenarios as described above, where the probability of being missed differed by
176 Factor X, while keeping the same overall probabilities of being missing. For each of the 18 main
177 scenarios, 4 sub-scenarios (a, b, c and d) were simulated where the probability of being missed
178 differed between those with and without Factor X with a relative risk (RR) of 1.5, 2, 3 and 5. For
179 example, a RR of 1.5 means that those with Factor X were 50% more likely to be missed as those
180 without the factor. The probability of being missed with and without Factor X, as represented in
181 each scenario, is presented in Table 2. When simulating the time to death due to cancer in all
182 these scenarios, the value of factor X for each individual was replaced by the value minus 0.25,
183 so that the average hazard rate follows the Weibull distributions described above.

184

185 2.3 Estimating bias in cancer survival estimates

186 . We estimated relative survival ratios (RSR) at 1 and 5 years after diagnosis as measures of
187 cancer survival under two situations: (1) all missed cases were excluded from the analysis, thus
188 representing a situation where DCI cases are not included and (2) DCI cases are included. The
189 relative survival was estimated using flexible parametric models (9-11) with 4 degrees of
190 freedom, without inclusion of any covariates, and using the rate as specified from the Weibull
191 distribution used in simulation of time from death due to other causes for the expected mortality.
192 To calculate the bias in the RSR estimates introduced by excluding or including DCI cases, the
193 RSR estimates for situations (1) and (2) were compared with the true cancer specific survival
194 based on the Weibull distributions used for the simulations.. Both the absolute (as percentage
195 points) and relative (percentage) differences were calculated. The proportion of DCI cases was
196 also estimated as the difference in the number of cases included for the two situations, divided by
197 the number of cases included for situation (2). All results presented are averages based on 1000
198 simulations for each scenario.

199

200 2.4 Sensitivity analysis

201 Scenarios with HRs for Factor X of 1.5 and 2 were also simulated, and results from those
202 simulations are provided in the Appendix.

203

204 **3. Results**

205 3.1 Randomly missed cases

206 When cases who are missed by the registry were a random sample of all cancer cases occurring in
207 the population, unbiased estimates for the RSR were obtained when DCI cases were excluded

208 (Figure 2). Including DCI cases however underestimated survival, since the DCI cases are a
209 selection of those missed who have a poorer prognosis. The size of the bias introduced differed
210 across the 18 simulated scenarios, with the most important factor being the proportion of cases
211 missed. When 5% of cases were missed (scenarios 1-6), the bias was small, less than 0.5
212 percentage points for 1-year survival and 0.6 percentage points for 5-year survival. When 15% of
213 cases were missed (scenarios 7-12) the bias in 1-year survival was still lower than 1.5 percentage
214 point, and just above 1.5 percentage points for 5-year survival. The largest bias – 2.5 percentage
215 points for 1-year and 2.8 for 5-year survival – occurred when 25% of cases were not notified
216 (scenarios 13-18).

217
218 There was no clear trend in the extent of bias in terms of the prognosis of the cancer (low,
219 medium or high survival), or the level of other cause survival. Rather it was the combination of
220 cancer and other cause survival which was important, since the extent of bias depends on the
221 proportion of the missed cases who were added when the DCI cases were included in the
222 analysis. As the bias will also depend on the true RSR, the relative bias is also presented in
223 Figure 2.

224
225 3.2 Non-randomly missed cases
226 For the next set of results (Figure 3) we assumed that cases with a poorer prognosis were more
227 likely to be missed. In this analysis, the exclusion of DCI cases led to an overestimation of
228 survival, and for many scenarios this overestimation was greater than the underestimation
229 introduced when DCI cases were included. The bias introduced by either including or excluding
230 DCI cases was largest for the scenarios where 25% of cases were missed by the registry,
231 suggesting that the proportion of cases missed was the most important driver of potential bias.

232 When DCI cases were excluded, the gap between the estimated and true survival widened, with
233 an increasing RR of being missed for those with Factor X. The opposite was true when DCI cases
234 were included. The largest bias observed when including DCI cases was an underestimation of 1-
235 year survival by 2.7 percentage points and 5-year survival by 2.9 percentage points. The largest
236 bias observed when excluding the DCI cases was an overestimation of 1-year survival by 5.9
237 percentage points and 5-year survival by 5.4 percentage points. Again, there was no clear trend in
238 the extent of bias in terms of cancer-specific survival, or other cause survival.

239

240 3.3 Proportion of Death Certificate Initiated cases

241 The proportions of DCI cases for each simulated scenario are presented in Table 3. The
242 proportion of DCI cases depends on the proportion of missed cases, since it can never be higher
243 than the proportion missed. For any given value of the proportion missed, the proportion of DCI
244 cases decreased with increasing cause-specific survival, as there would be a diminishing number
245 of cases who die from cancer. On the other hand, the proportion of DCI cases was higher for
246 higher other cause survival. This is because a larger proportion of cases will die due to cancer if
247 fewer die due to other causes. Finally, the proportion of DCI cases also increased with increasing
248 RR of being missed for those with Factor X compared to those without Factor X.

249

250 3.4 Sensitivity analysis

251 For scenarios where the HR for Factor X was changed to 1.5 or 2, the pattern of the results were
252 similar to the scenarios where the HR was 4, however, the bias introduced by excluding DCI
253 cases was smaller with a lower HR (Appendix Figures A2 and A3). The bias introduced by
254 including DCI cases was less affected by the size of the HR for Factor X.

255

256 **4. Discussion**

257
258 Our simulation study shows that performing trace-back to include DCI cases, does not resolve the
259 problem of missing cases biasing survival estimates, and can in certain circumstances lead to an
260 even larger bias than that resulting from excluding DCI cases from the analyses. The inclusion of
261 DCI cases in cancer registries is a necessary procedure to achieve the highest possible
262 completeness in terms of cancer incidence. When estimating survival, the inclusion of DCI cases
263 will underestimate survival, while their exclusion will overestimate survival. The utilization of
264 death certificates as a source for cancer notifications implies that some cancer cases are not
265 reported to the registry when diagnosed, and even if those missed are a random sample of cases,
266 inclusion of the DCI cases will lead to biased survival estimates. Thus, excluding DCI cases
267 when estimating survival will lead to unbiased survival estimates only if those cases not notified
268 represent a random sample of all cancer cases – which is unlikely in most situations– otherwise,
269 survival will be overestimated if the missed cases have more severe disease.

270
271 In our study we have demonstrated the impact on survival estimates of including and excluding
272 DCI cases. This has consequences for survival benchmarking. For two countries where one
273 includes DCI cases that were successfully traced back and the other does not, both estimates of
274 cancer survival will be biased, but in opposite directions. Even when comparing two populations
275 with the same practice in terms of including or excluding DCI cases, the bias may be of different
276 magnitudes depending on the true proportion missing within each registry, the mechanisms that
277 dictate the degree of missingness and the amount of trace-back. The inclusion of DCI cases could
278 also lead to greater underestimation if the trace-back doesn't find the true date of diagnosis but
279 rather a later date such as that at recurrence, but this was not evaluated in this study. An

280 additional issue when investigating trends over calendar time is that there is less opportunity for
281 cases diagnosed (and missed) more recently, to be obtained from death certificates due to their
282 shorter follow-up.

283
284 All cancer registries participating in ICBP SURVMARK-2 include DCI cases, although the
285 proportion of DCI cases is often unknown. Unfortunately, most cancer registries are not able to
286 retrospectively identify DCI cases in their data as typically this information is superseded when
287 other information relating to time prior to death is retrieved. However, for registries within ICBP
288 SURVMARK-2 where the proportion of DCI cases is available, a proportion of about 15% can
289 be observed for cancer sites with poor prognosis, indicating that scenarios 7 and 8 are plausible
290 scenarios for a poor prognosis cancer. For cancer sites with better prognosis, a proportion of DCI
291 cases of about 3-4% has been observed in real data, indicating that scenarios 3-6 are plausible.
292 However, given the small number of registries that have information on DCI cases, and the
293 uncertainty in the proportion of missed cases, we explored a wider range of scenarios in this
294 study.

295
296 A few limitations should be noted in relation to our study. We did not simulate an age
297 distribution within the data, but rather investigated two levels of other cause survival. In all
298 simulations we assumed that the prognostic Factor X was only associated with cancer-specific
299 survival, but not other cause survival, which might be violated if the prognostic factor is, for
300 example, the presence of comorbidity. We also assumed that cause of death is recorded
301 accurately for all cases. Another aspect that could be of interest is specification of DCO cases. In
302 our simulations we assumed that the true date of diagnosis is found for all DCI cases, resulting in
303 no DCO cases. We also assumed that all death certificates had been retrieved by the registry by

304 the time the survival analysis was performed, so all cases were correctly classified. Even so, this
305 simulation study showed clearly how the inclusion of DCI cases underestimates survival, and
306 excluding DCI cases instead overestimates survival if cases who were not notified were not a
307 random sample of all cancer patients in the population.

308
309 The extent of bias largely depends on the proportion of cases who are not notified, but the bias
310 also differs depending on the extent to which the missed cases are notified as DCI cases (i.e. the
311 proportion of the missed cases who have died and had cancer mentioned on their death
312 certificates). It is reassuring to see that our scenarios give a bias of at most 3 percentage points in
313 the situation when DCI cases are included. It is by definition impossible to know the true
314 proportion of cases missed by a registry, but the proportion of DCI cases serves as an important
315 indicator in this respect. Registries should therefore report the proportion of DCI cases along with
316 the more commonly reported proportion of DCO cases.

317

318 Conflict of interest

319 The authors declare no competing interests.

320

321 Funding

322 The ICBP is funded by the Canadian Partnership Against Cancer; Cancer Council Victoria;
323 Cancer Institute New South Wales; Cancer Research UK; Danish Cancer Society; National
324 Cancer Registry Ireland; The Cancer Society of New Zealand; NHS England; Norwegian Cancer
325 Society; Public Health Agency Northern Ireland on behalf of the Northern Ireland Cancer
326 Registry; DG Health and Social Care, Scottish Government; Western Australia Department of
327 Health; Public Health Wales NHS Trust.

328

329 Acknowledgements

330 The authors would also like to thank Lucie Hooper, Samantha Harrison, Charles Norell, Shanta
331 Keshwala and Charlotte Lynch of Cancer Research UK for managing the programme. The ICBP
332 Clinical Committees for their advice. The ICBP SurvMark-2 Academic Reference Group for
333 providing independent peer review and advice for the study protocol and analysis plan
334 development. Finally, we are thankful to the ICBP Programme Board for their oversight and
335 direction.

336

337 Author statement

338 Where authors are identified as personnel of the International Agency for Research on
339 Cancer/WHO, the authors alone are responsible for the views expressed in this article and they do
340 not necessarily represent the decisions, policy or views of the International Agency for Research
341 on Cancer/WHO.

342

343 **References**

- 344
- 345 1. Beral V, Peto R. UK cancer survival statistics. *Bmj*. 2010;341:c4112.
 - 346 2. Robinson D, Sankila R, Hakulinen T, Moller H. Interpreting international comparisons of cancer
347 survival: the effects of incomplete registration and the presence of death certificate only cases on
348 survival estimates. *Eur J Cancer*. 2007;43(5):909-13.
 - 349 3. Moller H, Richards S, Hanchett N, Riaz SP, Luchtenborg M, Holmberg L, et al. Completeness of
350 case ascertainment and survival time error in English cancer registries: impact on 1-year survival
351 estimates. *Br J Cancer*. 2011;105(1):170-6.
 - 352 4. Woods LM, Coleman MP, Lawrence G, Rashbass J, Berrino F, Rachet B. Evidence against the
353 proposition that "UK cancer survival statistics are misleading": simulation study with National Cancer
354 Registry data. *Bmj*. 2011;342:d3399.
 - 355 5. Rutherford MJ, Moller H, Lambert PC. A comprehensive assessment of the impact of errors in the
356 cancer registration process on 1- and 5-year relative survival estimates. *Br J Cancer*. 2013;108(3):691-8.
 - 357 6. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part
358 II. Completeness. *Eur J Cancer*. 2009;45(5):756-64.
 - 359 7. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part
360 I: comparability, validity and timeliness. *Eur J Cancer*. 2009;45(5):747-55.
 - 361 8. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*.
362 2013;32(23):4118-34.
 - 363 9. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models
364 for censored survival data, with application to prognostic modelling and estimation of treatment effects.
365 *Stat Med*. 2002;21(15):2175-97.
 - 366 10. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with
367 application in coronary heart disease. *Stat Med*. 2007;26(30):5486-98.
 - 368 11. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis.
369 *Stata J*. 2009;9(2):265-90.

370
371

372 Table 1. Combinations of probability of cases being missed in the registry, cancer-specific
 373 survival, and other cause (non-cancer) survival included in the 18 simulated main scenarios.

Scenario	Probability missed	Cancer-specific survival	Other cause survival
1	0.05	Low	Low
2	0.05	Low	High
3	0.05	Medium	Low
4	0.05	Medium	High
5	0.05	High	Low
6	0.05	High	High
7	0.15	Low	Low
8	0.15	Low	High
9	0.15	Medium	Low
10	0.15	Medium	High
11	0.15	High	Low
12	0.15	High	High
13	0.25	Low	Low
14	0.25	Low	High
15	0.25	Medium	Low
16	0.25	Medium	High
17	0.25	High	Low
18	0.25	High	High

374

375 Table 2. Probability of a case with and without prognostic Factor X being missed by the registry,
 376 in four sub-scenarios* for each of the 18 base scenarios.

Scenarios	Sub-scenario	Probability missed among cases without Factor X	Probability missed among cases with Factor X
1-6	a	0.044	0.066
1-6	b	0.040	0.080
1-6	c	0.033	0.100
1-6	d	0.025	0.125
7-12	a	0.133	0.200
7-12	b	0.120	0.240
7-12	c	0.100	0.300
7-12	d	0.075	0.375
13-18	a	0.222	0.333
13-18	b	0.200	0.400
13-18	c	0.166	0.500
13-18	d	0.125	0.625

377 *Sub-scenarios a to d represent relative risk of being missed in the registry of 1.5; 2; 3 and 5,
 378 respectively

379 Table 3. Proportion of Death Certificate Initiated (DCI) cases in each simulated scenario.

Scenario	% DCI	Scenario	% DCI	Scenario	% DCI
1	3.7	7	11.3	13	19.4
1a	3.7	7a	11.3	13a	19.4
1b	3.8	7b	11.6	13b	19.8
1c	3.9	7c	12.0	13c	20.5
1d	4.1	7d	12.5	13d	21.3
2	4.4	8	13.4	14	22.6
2a	4.3	8a	13.2	14a	22.3
2b	4.4	8b	13.3	14b	22.5
2c	4.5	8c	13.6	14c	22.9
2d	4.6	8d	13.8	14d	23.3
3	3.0	9	9.3	15	16.2
3a	3.1	9a	9.6	15a	16.7
3b	3.2	9b	10.0	15b	17.3
3c	3.4	9c	10.6	15c	18.2
3d	3.6	9d	11.2	15d	19.3
4	3.9	10	11.9	16	20.3
4a	3.8	10a	11.8	16a	20.2
4b	3.9	10b	12.1	16b	20.6
4c	4.1	10c	12.5	16c	21.2
4d	4.2	10d	13.0	16d	21.9
5	1.4	11	4.4	17	8.0
5a	1.6	11a	5.2	17a	9.4
5b	1.7	11b	5.6	17b	10.1
5c	1.9	11c	6.1	17c	11.0
5d	2.1	11d	6.8	17d	12.2
6	2.2	12	7.0	18	12.4
6a	2.4	12a	7.6	18a	13.5
6b	2.5	12b	8.1	18b	14.2
6c	2.7	12c	8.6	18c	15.2
6d	3.0	12d	9.4	18d	16.4

380

381

382 Figure 1. Illustration of Death Certificate Initiated (DCI) cases as a subset of all cases of cancer
383 arising in the population.

384

385

386 Figure 2. Absolute and relative differences in 1- and 5-year relative survival ratios (RSR) for the
387 18 base scenarios* described in Table 1 (where the missed cases are a random sample of all
388 cases): including or excluding death certificate initiated cases compared to the full cohort.
389 Negative values refer to underestimation of survival, and positive values overestimation of
390 survival.

391

392

393 Note that the absolute and relative differences are shown with different scales

394

395 * 5%, 15%, 25% missing registration for scenarios 1-6, 7-12 and 13-18 respectively with different combinations of
396 low, medium and high cancer specific survival and level of other cause survival.

397

398 Figure 3. Absolute and relative differences in 1- and 5-year relative survival ratios (RSR) for the
399 72 simulation scenarios* described in Table 1 and Table 2: including or excluding death
400 certificate initiated cases compared to the full cohort. For each of the 18 base scenarios, sub-
401 scenario a to d are displayed with varying degrees of transparency, a with least and d with most
402 transparent circles. Negative values refer to underestimation of survival, and positive values
403 overestimation of survival.

404

405

406 Note that the absolute and relative differences are shown with different scales

407

408 *5%, 15%, 25% missing registration for scenarios 1-6, 7-12 and 13-18 respectively with different combinations of
409 low, medium and high cancer specific survival and level of other cause survival. Sub scenarios a-d: Relative risk of
410 being missed for those with Factor X (with higher risk of dying) relative to those without of 1.5, 2, 3 and 5,
411 respectively.

412

413 **Appendix**

414 Figure A1. Cause-specific and other cause survival (a) and hazard (b) functions used in
415 simulations, representing scenarios with low, medium and high cancer-specific survival and high
416 and low other cause survival.

417

418

419 (a)

420

421

422 (b)

423

424 Figure A2: Absolute and relative differences in 1- and 5-year relative survival ratios (RSR) for
425 the 72 simulation scenarios* described in Table 1 and Table 2 using hazard ratio of 1.5 for cases
426 with the prognostic Factor X.

427

428

429 Note that the absolute and relative differences are shown with different scales

430

431 * 5%, 10%, 15% missing registration for scenarios 1-6, 7-12, 13-18 respectively with different combinations of low,
432 medium and high cancer specific survival and level of other cause survival. Sub scenarios a-d: Relative risk of being
433 missed for those with Factor X (with higher risk of dying) relative to those without of 1.5, 2, 3 and 5, respectively.

434

435 Figure A3. Absolute and relative differences in 1- and 5-year relative survival ratios (RSR) for
436 the 72 simulation scenarios* described in Table 1 and Table 2 using hazard ratio of 2 for cases
437 with the prognostic Factor X.

438

439

440 Note that the absolute and relative differences are shown with different scales

441

442 *5%, 15%, 25% missing registration for scenarios 1-6, 7-12, 13-18 respectively with different combinations of low,
443 medium and high cancer specific survival and level of other cause survival. Sub scenarios a-d: Relative risk of being
444 missed for those with Factor X (with higher risk of dying) relative to those without of 1.5, 2, 3 and 5, respectively.

445