

1 **UKCTOCS Update: Applying insights of delayed effects in cancer screening**  
2 **trials to the long-term follow-up mortality analysis**

3

4 Matthew Burnell *PhD*<sup>1</sup>, Aleksandra Gentry-Maharaj *PhD*<sup>1</sup>, Steven J Skates *PhD*<sup>2</sup>,  
5 Andy Ryan *PhD*<sup>1</sup>, Chloe Karpinskyj *MSc*<sup>1</sup>, Jatinderpal Kalsi *PhD*<sup>3</sup>, Sophia  
6 Apostolidou *PhD*<sup>1</sup>, Naveena Singh *FRCPath*<sup>4</sup>, Anne Dawnay *PhD*<sup>5</sup>, Robert Woolas  
7 *FRCOG*<sup>6</sup>, Lesley Fallowfield *DPhil*<sup>7</sup>, Stuart Campbell *DSc*<sup>8</sup>, Alistair McGuire *PhD*<sup>9</sup>,  
8 Ian J Jacobs *FRCOG*<sup>3,10</sup>, Mahesh Parmar *DPhil*<sup>1</sup>, Usha Menon *FRCOG*<sup>1</sup>

9

10 <sup>1</sup>*MRC CTU at UCL, Institute of Clinical Trials and Methodology, University College*  
11 *London, 90 High Holborn, 2<sup>nd</sup> Floor, London, WC1V 6LJ, UK;* <sup>2</sup>*MGH Biostatistics,*  
12 *Massachusetts General Hospital and Harvard Medical School, 55 Fruit Street,*  
13 *Boston, MA 02114, US;* <sup>3</sup>*Department of Women's Cancer, Institute for Women's*  
14 *Health, University College London, 84-86 Chenies Mews, London WC1E 6HU, UK;*  
15 <sup>4</sup>*Department of Pathology, Barts Health National Health Service Trust, The Royal*  
16 *Hospital, Whitechapel Rd, London E1 1BB, UK;* <sup>5</sup>*Department of Clinical*  
17 *Biochemistry, Barts Health National Health Service Trust, Clinical Biochemistry,*  
18 *Barts Health, 4th floor, Pathology and Pharmacy, 80 Newark St, London E1 2ES,*  
19 *UK;* <sup>6</sup>*Department of Gynaecological Oncology, Queen Alexandra Hospital, Cosham,*  
20 *Portsmouth PO6 3LY, Hampshire, UK;* <sup>7</sup>*Sussex Health Outcomes Research and*  
21 *Education in Cancer, Brighton and Sussex Medical School, University of Sussex,*  
22 *Science Park Road, Falmer, Brighton, BN1 9RX, UK;* <sup>8</sup>*Create Health, 150*  
23 *Cheapside, London EC2V 6ET, UK;* <sup>9</sup>*Department of Social Policy, London School of*  
24 *Economics, Houghton Street, London WC2A 2AE, UK;* <sup>10</sup>*University of New South*  
25 *Wales, UNSW Sydney, NSW 2052, Australia.*

26

27 **Corresponding Author**

28 Professor Usha Menon

29 MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology

30 University College London

31 90 High Holborn, 2nd Floor, London WC1V 6LJ

32 +44 (0)20 7670 4649 [u.menon@ucl.ac.uk](mailto:u.menon@ucl.ac.uk)

33

34 **Abstract**

35

36 **Background**

37 During trials that span decades, new evidence including progress in statistical  
38 methodology, may require revision of original assumptions. An example is the  
39 continued use of a constant-effect approach to analyse the mortality reduction which  
40 is often delayed in cancer-screening trials. The latter led us to re-examine our  
41 approach for the upcoming primary mortality analysis(2020) of long-term follow-up of  
42 the United Kingdom Collaborative Trial of Ovarian Cancer Screening (LTFU  
43 UKCTOCS), having initially(2014) used the proportional hazards(PH) Cox-model.

44 **Methods**

45 We wrote to 12 experts in statistics/epidemiology/screening-trials, setting out current  
46 evidence, importance of pre-specification, previous mortality analysis (2014) and  
47 three possible choices for the follow-up analysis (2020) of the mortality outcome -  
48 (A)all data(2001-2020) using the Cox-model(2014) (B)new data(2015-2020) only  
49 (C)all data(2001-2020) using a test that allows for delayed effects.

50 **Results**

51 Of 11 respondents, eight supported changing the 2014-approach to allow for a  
52 potential delayed effect (optionC), suggesting various tests while three favoured  
53 retaining the Cox-model (optionA). Consequently, we opted for the Versatile test  
54 introduced in 2016 which maintains good power for early, constant or delayed  
55 effects. We retained the Royston-Parmar model to estimate absolute differences in  
56 disease-specific mortality at 5,10,15 and 18 years.

57 **Conclusions**

58 The decision to alter the follow-up analysis for the primary outcome on the basis of  
59 new evidence and using new statistical methodology for long-term follow-up is novel  
60 and has implications beyond UKCTOCS. There is an urgent need for consensus  
61 building on how best to design, test, estimate and report mortality outcomes from  
62 long-term randomised cancer screening trials.

63

64 Trial registration: (ISRCTN22488978, Registration date: 6/4/2000)

65

66 **Key words**

67 UKCTOCS, follow-up, mortality analysis, ovarian cancer, cancer screening, delayed  
68 effect  
69

70 **BACKGROUND**

71 Randomised controlled trials (RCT) are the cornerstone of the evidence base for  
72 clinical management of millions of patients across the world. RCTs evaluating the  
73 mortality impact of cancer screening typically involve large numbers of participants  
74 followed up over many years, sometimes decades. The general rule in clinical trials  
75 is strict adherence to the statistical analysis plan specified prior to unblinding and  
76 analysis of outcome data. Sometimes, during continued long-term follow-up of these  
77 trials, new understanding based on evidence from other trials and new analytical  
78 methods, may require re-evaluation of the analysis plan.

79

80 One important example is the accumulating evidence in cancer-screening trials of a  
81 delay of several years before a mortality reduction is observed between the screen  
82 and control arms[1-3]. Almost all the cancer-screening trials, breast[4-14], prostate,  
83 colorectal and lung[15-31] in their graphic representation of disease-specific mortality  
84 over time have reported a delayed difference (if present) between screen and control  
85 arms(Table 1). Most have an initial time window in the first several years after start of  
86 screening during which there is little or no mortality reduction, followed by one in  
87 which the reduction becomes evident[2]. However, almost none of these cancer-  
88 screening trials have used analytical methods which formally allow for a non-  
89 constant effect (non-proportional hazards). All have described the screening effect  
90 using relatively simple methods, usually a single Poisson-based rate ratio (RR)[4, 12,  
91 24, 30, 32, 33] or Cox model with a single hazard ratio (HR) estimate[18, 22]. A  
92 single HR is only appropriate if the reduction in hazard rates is relatively immediate  
93 and constant over time. In screening trials, such estimates cannot reliably describe  
94 the changing effects of screening on mortality over time.

95

96 Alongside, new analytical methods have been developed for trials lacking treatment  
97 proportionality. Tests that combine evidence from more than one aspect of the data  
98 have gained popularity as a way to mitigate the effects of potential but unknown non-  
99 proportionality of hazards, although some may work best in a specific scenario. The  
100 'joint test' appears in simulations to be preferentially beneficial under late effects[34,  
101 35] whilst the 'combined test' appears to be preferentially beneficial under early  
102 effects[36, 37]. Another recent addition is the Versatile test[38], which seeks to cover  
103 all bases by combining three (weighted) log-rank tests giving good power for the test

104 under early effects, proportional hazards(PH) and late effects, respectively. These  
105 tests are likely better suited than the Cox model for analysis of outcomes which are  
106 non-proportional across the duration of a trial.

107

108 In the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)  
109 too, the initial mortality analysis in 2014 used a PH Cox model and reported an  
110 average mortality reduction estimate. However, given the growing external evidence,  
111 there have been extensive discussions within the UKCTOCS trial committees to  
112 ensure the outcome data is analysed appropriately. We believe that this issue will be  
113 important for any long-term cancer screening trial. The Cox model, while valid, could  
114 be viewed as restrictive and failing to utilise the most appropriate analytical  
115 approach, given the delayed mortality reductions seen in many screening trials  
116 across a range of cancers (Table1)[14, 17, 24, 31]. Furthermore, retention of the Cox  
117 model based on pre-specification may result in suboptimal interpretation of  
118 UKCTOCS data and therefore an abrogation of our responsibility to the huge  
119 collective investment by the trial volunteers, the funding agencies, charities, the  
120 National Health Service (NHS), researchers and most importantly women who  
121 develop ovarian cancer in the future. This is balanced by a concern that changes to  
122 the 2014 analysis plan could be controversial and lead to criticism of cherry-picking  
123 methodology that gives the 'best' test result.

124

125 Many trialists may face similar dilemmas, when new evidence suggests that trial  
126 design, conduct or analysis may need to be amended. Decisions are often made by  
127 the Trial Management Committee (TMC) with input from independent oversight  
128 bodies such as a Trial Steering (TSC) or Scientific Advisory (SAC) Committees. We  
129 report on the process we undertook in UKCTOCS to re-examine our approach for  
130 the upcoming analysis (2020) of the primary mortality outcome at the end of  
131 extended follow-up and how we addressed the issue of delayed effects.

132

## 133 **METHODS**

134 Between 2001 and 2005, 202,638 postmenopausal women aged 50-74 were  
135 recruited to UKCTOCS. They were randomised to screening using a longitudinal  
136 serum CA125 algorithm (multimodal group, MMS, 50,640), transvaginal ultrasound  
137 (ultrasound group,USS,50,639) or no screening (control group,C,101,279) as

138 described previously[39-41]. Women in the screen groups underwent screening until  
139 the end of 2011 and received a median of nine annual screens. At median follow-up  
140 of 11.1 years (administrative censorship 31 Dec 2014), a higher proportion of women  
141 were diagnosed with low-volume (stage I, II, and IIIa) tubo-ovarian cancer in the  
142 MMS(40%; $p<0.0001$ ) compared to C(26%) group. The Cox-model indicated a trend  
143 to mortality reduction in favour of MMS (HR 0.85;95%CI:0.70-1.03, $p=0.10$ ) and USS  
144 (HR 0.89;95% CI:0.73-1.07, $p=0.21$ ), which was not statistically significant at the 5%  
145 level. A Royston-Parmar (RP) flexible parametric model showed that HR varied over  
146 time. In the MMS group, it was 0.92(95% CI:0.69-1.20) in years 0-7 and 0.77(95%  
147 CI:0.54-0.99) in years 7-14. In the USS group, it was 0.98(95% CI:0.74-1.27) in  
148 years 0-7 and 0.79(95%:CI 0.58-1.02) in years 7-14[39]. Follow-up was extended to  
149 30 June 2020 to assess the long-term mortality impact (LTFU UKCTOCS)[39, 42].  
150 Final receipt of death data from the registries is anticipated by the end of September  
151 2020, with unblinding and analysis planned for November 2020.

152

153 To ensure independent input into our statistical conundrum, the TMC proposed  
154 seeking the views of a broad panel of international experts with statistical and  
155 screening trial expertise who had not been involved in any aspect of UKCTOCS. The  
156 process was developed through detailed discussions with the independent members  
157 of the TSC. In September 2019, 12 experts (Table 2) were approached by the Trial  
158 Statistician for advice. They were sent a letter briefly describing UKCTOCS together  
159 with a summary of the current evidence from other cancer-screening trials,  
160 importance of pre-specification and our 2014 mortality analysis results. Three  
161 potential options for the primary analysis of the extended follow-up data developed  
162 with the TSC were described sequentially, each including possible pros and cons, in  
163 a neutral manner. These were:

164 A) analyse all outcome data (2001-2020) using the PH Cox-model of the original  
165 UKCTOCS analysis, representing the pre-specification viewpoint

166 B) analyse only the outcomes that occurred since the original censorship (31  
167 December 2014), either assuming PH or not, to address the view that data should  
168 not be re-used, without formal statistical accommodation for multiple analyses.

169 C) model all outcome data using a method of analysis and model that allows for a  
170 late effect of screening on mortality and reflects current understanding of cancer-  
171 screening trials - a pragmatic evidential approach. The specific model suggested for

172 C) was the RP model[43] as it had been used as a secondary analysis method for  
173 the 2014 analysis[39].

174

175 Experts were asked to critique and state a preference or suggest another option  
176 (Supplementary Materials 1). Results were collated and summarised based on 1)  
177 indicated choice of A, B, C or other and 2) pertinent comments provided.

178

## 179 **RESULTS**

180 In total 12 individuals were contacted from the UK (5), USA (5), Canada (1) and  
181 Belgium (1) and 11 responded (see acknowledgement). Their anonymised  
182 responses can be found in Table 2 and Supplementary Table 1.

183

184 Eight (73%) of the 11 experts recommended changing the pre-specified analysis to  
185 one that more appropriately allows for a delayed effect (Table 2). *EX4* was not  
186 troubled by the shift from a pre-hoc to post-hoc decision - “reason” should have a  
187 role in science. Similarly, *EX8* argued “a conclusion should be reached based on a  
188 proper consideration of the full evidence” and use scientific principles – “full  
189 information from data should be extracted”. Indeed, rather than viewing it as “data-  
190 dredging” or “changing the endpoint”, *EX8* described this approach as just “using  
191 common sense”. *EX9* felt the lack of (complete) pre-specification a weakness, but  
192 not “a violation of good scientific principles”. For “a major and definitive screening  
193 trial ..... such regulatory constraints should not be the primary consideration” but  
194 instead “approximating the truth as well as possible”. *EX11* was not persuaded by  
195 the pre-specification argument, and claimed keeping a plan that is less preferable  
196 “turns research rules into an irrational, mindless, and restricting obsession with  
197 methodological procedure”; “rules have a purpose, but when the higher priority is  
198 understanding phenomena in a reasoned disciplined way... then a compelling  
199 argument can be made to deviate from them”. *EX11* stated that no screening trial  
200 has shown an immediate effect and appealed to the common sense of the scientific  
201 audience; “we can discern the difference in attempts by a study team to game the  
202 analysis to gain statistical significance, from a good faith effort to apply a statistical  
203 technique that is more appropriate for the data”. Different screening trials will have  
204 different results and delayed effects, all dependent on differing facets of trial design

205 and the cancer itself, the effects of which are largely unknown until we do the study.  
206 “Point is, we are still learning how to design and analyse RCT screening trial data.”

207 Three of the eleven (*EX2*, *EX3*, *EX1*) believed that we should retain the initial  
208 analysis approach (option A). This was based on the pre-specification argument -  
209 “avoids the appearance of trying to get a significant result by changing the  
210 test”(EX2), “maintains credibility in the scientific community”(EX3), “most likely to be  
211 accepted as valid by the cancer research and policy community”(EX1). However,  
212 *EX1* did suggest modifying the pre-specified plan to limit analysis to only cancers  
213 diagnosed within the screening period.

214

215 Of the eight who suggested changing the pre-specified analysis, five (*EX7*, *EX8*,  
216 *EX9*, *EX10* and *EX11*) explicitly selected approach C (using all acquired outcome  
217 data and a model that allows for delayed effects). While there were positive  
218 comments about the suggested RP model (credibility due to pre-specification *EX7*,  
219 informative of the screening effect over time *EX9*), none gave a clear endorsement  
220 of this approach. The main reason was interpretability (*EX7*, *EX9*, *EX4*, *EX6*). *EX10*  
221 noted that power was little studied under various “flavours” of non-PHs, and  
222 suggested separating testing from estimation, opting for a versatile weighted log-  
223 rank test for the former. *EX4* and *EX6* formally indicated an alternative option. *EX6*’s  
224 preference was for dividing the data into yearly bins and estimating the HR in each,  
225 possibly with some smoothing. *EX6* argued extensively we should avoid a single HR  
226 estimate, which will provide “a very blurred, incomplete and misleading picture of  
227 how much/little good screening did for the 100,000 participants screened, or of how  
228 much future women might expect from a screening regimen based on these  
229 screening tools.” *EX4* stated that the number needed to screen was the most  
230 suitable measure for a screening study. *EX5* recommended a test based on the  
231 difference of restricted mean survival times (RMST) which “does not need any  
232 modelling and the results can be interpreted easily clinically”.

233

234 None of the 11 responders chose Approach B. This was mainly because it did not  
235 use the full dataset. In addition, there were concerns that it could lead to  
236 ‘unfavourable early results (important data) being censored(*EX11*) and a  
237 “disconnected” HR(*EX6*).



238

239 Based on the feedback, we decided to change the primary analysis test for LTFU  
240 UKCTOCS. Table 3 summarises the major pros and cons of available approaches to  
241 dealing with non-PH in terms of tests. We used two main criteria to choose the  
242 specific test - (1) minimal *a priori* specification on the specific form of the mortality  
243 difference over time (2) able to accommodate delayed effects while maintaining good  
244 power in a variety of potential scenarios. Based on these criteria, we opted for the  
245 Versatile test[16], suggested by EX10. The RP model was retained to estimate  
246 absolute differences in disease-specific mortality at 5, 10, 15 and 18 (our estimate of  
247 the upper limit of reliable follow-up given administrative censorship on 30 June 2020)  
248 years. Options A and B were included as secondary analyses of the primary  
249 mortality outcome. These amendments were incorporated into the statistical analysis  
250 plan (20 February 2020), which was endorsed by the independent TSC.

251

## 252 **DISCUSSION**

253 Given the now large body of evidence of a delay in mortality reduction in long-term  
254 cancer-screening randomised trials, and the majority view of independent statistical,  
255 epidemiological and screening trial experts, we altered the approach for our primary  
256 mortality analysis for the LTFU from that used for our 2014 analysis. The new  
257 approach allows for a delayed effect in contrast to our previous analysis which  
258 assumed a constant screening effect. There were a variety of opinions on the  
259 specific test which suggests an urgent need for consensus building on how best to  
260 design, analyse and report mortality outcomes in cancer-screening trials.

261

262 Our decision to change the statistical analysis plan for extended follow-up is a  
263 significant decision. The large majority of the published cancer-screening trials[17,  
264 25, 26, 31, 32, 44] have retained the same primary mortality analysis methodology  
265 for both their initial and extended follow-up analysis (Table 1). The only exceptions  
266 we found were the Two County trial which used negative binomial regression[14] for  
267 follow-up analysis in place of Mantel-Haenszel stratified risk-ratios[12] and the  
268 Norwegian Colorectal Cancer Prevention Trial (NORCCAP) which changed the  
269 primary analysis from overall population to subgroups based on gender[21]. In the  
270 Two Country trial, whilst no explanation was given, the change was not substantive;  
271 both initial and follow-up methods estimated risk ratios. For NORCCAP, “because

272 substantial heterogeneity existed between women and men, the steering committee  
273 decided to present results for women and men separately”, which may be argued as  
274 a significant post-hoc data-driven amendment. None of the trials as far as we are  
275 aware sought independent expert opinion. In contrast, we undertook an external  
276 consultation. Although the independent expert panel was not unanimous, the  
277 majority concluded that a rational argument for revision outweighs that of procedure  
278 and pre-specification, and recommended choosing the most appropriate test that  
279 allows for a delayed effect. We accepted the view of *EX7* that one should “do what  
280 you yourselves think is the most effective and secure analysis of all your data,  
281 bearing in mind the current state of information about the field.” There will be debate  
282 about our decision, which we welcome, given the broader implications.

283

284 A number of factors contribute to delayed mortality effect. In the early trial-years, the  
285 absolute death rates are low as a result of eligibility criteria which exclude women  
286 with cancer diagnosis. The time interval for an individual to be diagnosed with cancer  
287 after joining the trial and then dying of the disease also contributes to the delay in  
288 separation of the mortality curves. Additionally, the impact of screening on cancers  
289 detected at the initial prevalence screen is reduced, as these are necessarily more  
290 advanced when screen-detected compared to screen-detected cancers in later  
291 years. The performance of most screening strategies improve over time as the  
292 number of screens accumulate and the teams involved get more experienced. This  
293 is magnified when longitudinal biomarker algorithms are used as they are based on  
294 detecting change from baseline. Finally, the length of follow-up after end of  
295 screening impacts on the specific form of the mortality difference over time as the  
296 longer the interval, the greater the dilution of screen-detected cancers by cancers  
297 that develop after the end of screening[32].

298

299 The PLCO colorectal[29] and ovarian[19] trials used a test that has better power for  
300 the delayed effect described above. Both used the weighted log-rank test, which is  
301 perhaps the best known method for improving power in such situations. However, it  
302 requires correctly anticipating the specific form of the mortality difference over time,  
303 which will depend on the natural history of the cancer, screening strategy, number  
304 and frequency of screens and years of follow-up. We have chosen the Versatile  
305 test[38], introduced in 2016, which does not require pre-specification of the mortality

306 difference over time. It combines three (weighted) log-rank tests appropriate for  
307 capturing early effects, PH and delayed effects, respectively. It is therefore versatile  
308 enough to maintain good power in all potential scenarios, rather than optimal in any  
309 given scenario.

310

311 Unlike other trials, including the PLCO colorectal[29] and ovarian[19] trials, who  
312 measured the screening effect using a single ‘averaged’ rate-ratio, we will use a  
313 flexible parametric model to estimate absolute differences in disease-specific  
314 mortality at 5,10,15 and 18 years. This is in keeping with the growing view that to  
315 adequately describe what might be achieved with a particular cancer screening  
316 strategy, a more comprehensive set of time-specific measures needs to be reported.  
317 Hanley *et al* has extensively re-analysed cancer screening trial data and shown that  
318 a one-number summary measure systematically dilutes the estimate of mortality  
319 reduction that results from screening[2]. In the most recent re-analysis involving  
320 breast cancer screening data from Funen, Denmark, the average mortality reduction  
321 was 18% using a PH model and ranged from 0 to 30% when a non-PH model was  
322 used that considered the impact at different points over time. The reductions were  
323 largest for periods where sufficient time had elapsed for the impact to manifest[45].

324

325 The key strength of our approach is the independent and transparent process we  
326 have adopted to address a challenging issue and the criteria we used to choose a  
327 new specific approach. This involved accommodating delayed effects while  
328 maintaining good power in a variety of potential scenarios and requiring minimal a  
329 *priori* speculation on the specific form of the mortality difference over time. A  
330 limitation is that given the orthodoxy surrounding pre-specification for analysis of  
331 trials, we have retained the original Cox model with an averaged HR over time as an  
332 estimate for our secondary analysis.

333

334 The screening community is only beginning to understand the challenges posed by  
335 long-term cancer-screening trials. Mortality reductions may have been  
336 underestimated across cancer types by not considering their timing. Given the  
337 importance of early detection in many national cancer strategies, we hope our report  
338 will accelerate much needed consensus building on how best to design, analyse and  
339 report trials testing cancer screening strategies – as it is clear our currently accepted

340 and widely used methods are insufficient. We also hope it will encourage debate  
341 and transparency on how advances in understanding and new analytical methods  
342 can be evaluated and incorporated into long-term trials.

343

344 **List of abbreviations**

345 United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)

346 Long-term follow-up of the United Kingdom Collaborative Trial of Ovarian Cancer  
347 Screening (LTFU UKCTOCS)

348 Randomised controlled trial (RCT)

349 Rate ratio (RR)

350 Hazard ratio (HR)

351 Confidence interval (CI)

352 Proportional hazards (PH)

353 Trial Management Committee (TMC)

354 Trial Steering Committee (TSC)

355 Scientific Advisory Committee (SAC)

356 Multimodal group (MMS)

357 Ultrasound group (USS)

358 Royston-Parmar model (RP)

359 Norwegian Colorectal Cancer Prevention Trial (NORCCAP)

360 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)

361 **Declarations**

362 **Ethics approval and consent to participate**

363 The initial study was approved by the UK North West Multicentre Research Ethics  
364 Committees (North West MREC 00/8/34) on 21 June 2000 with site-specific approval  
365 from the local regional ethics committees and the Caldicott guardians (data  
366 controllers) of the primary care trusts. The long-term follow-up amendment was  
367 approved on 24 January 2017 and the amended protocol including the new statistical  
368 plan was approved on 12 May 2020. All trial participants provided written informed  
369 consent.

370

371 **Consent for publication**

372 All authors have seen the final version of the manuscript and give their consent for  
373 publication.

374

375 **Availability of data and materials**

376 Tables 2 and Supplementary Table 1 contain the exact comments provided by the  
377 experts.

378

379 **Competing interests**

380 UM has stocks in Abcodia Ltd. awarded to her by UCL. SJS and IJJ are co-inventors  
381 of the Risk of Ovarian Cancer Algorithm (ROCA) that has been licensed to Abcodia  
382 Ltd by Massachusetts General Hospital (MGH) and Queen Mary University of  
383 London (QMUL). IJJ has a financial interest in Abcodia. Ltd as a shareholder and  
384 director. IJJ and SJS are entitled to royalty payments via MGH and QMUL from any  
385 commercial use of the ROCA. All other authors declare no competing interests.

386

387 **Funding**

388 The LTFU UKCTOCS is supported by National Institute for Health Research (NIHR  
389 HTA grant 16/46/01), Cancer Research UK (CRUK) and The Eve Appeal.  
390 UKCTOCS was funded by Medical Research Council (G9901012 and G0801228),  
391 CRUK (C1479/A2884), and the Department of Health, with additional support from  
392 The Eve Appeal. Researchers at UCL are supported by the NIHR University College  
393 London Hospitals (UCLH) Biomedical Research Centre and MRC CTU at UCL core  
394 funding (MR\_UU\_12023).

395

396 **Disclaimer:** The views expressed are those of the authors and not necessarily those  
397 of the NHS, the NIHR or the Department of Health and Social Care.

398

#### 399 **Author contributions**

400 The process was conceived following many discussions within the TMC involving all  
401 authors. MP and UM supervised the study. MB performed the literature search. MB,  
402 SJS, AMcG, and MP proposed the statistical analysis options with further input from  
403 JC (TSC). The survey was drafted by MB, AGM, MP and UM with input from IJJ,  
404 AMcG, and SJS. AGM, AR and MB collated the results and MB undertook analysis.  
405 All contributed to data interpretation. MB prepared the tables. MB, AGM and UM  
406 drafted the manuscript. AMcG, LF, SA, JK, RW, IJJ, MP and SJS helped revise the  
407 draft. All authors critically reviewed the manuscript and approved the report before  
408 submission.

409

#### 410 **Acknowledgements**

411 We are hugely grateful to the international panel of experts (Professor Marc Buyse,  
412 Professor David Cox, Professor Stephen Duffy, Professor Mitch Gail, Professor Jim  
413 Hanley, Professor David Harrington, Professor Patrick Royston, Professor David  
414 Schoenfeld, Professor Robert Smith, Professor David Spiegelhalter, Professor LJ  
415 Wei) who have contributed their time and expertise. We are also indebted to the  
416 insights and support provided by the members of the Trial Steering Committee -  
417 Professor Henry Kitchener (Chair), Professor Julietta Patnick, Professor Jack Cuzick  
418 and Ms Annwen Jones. We thank all 202,638 volunteers without whom the trial  
419 would not have been possible and all the staff involved in this trial for their hard work  
420 and dedication.

421

422

423

424

425 **References**

- 426 1. Etzioni RD, Thompson IM. What do the screening trials really tell us and  
427 where do we go from here? *Urol Clin North Am* 2014;41(2):223-8.
- 428 2. Hanley JA. Measuring mortality reductions in cancer screening trials.  
429 *Epidemiol Rev* 2011;33:36-45.
- 430 3. Hanley JA, McGregor M, Liu Z, *et al.* Measuring the mortality impact of breast  
431 cancer screening. *Can J Public Health* 2013;104(7):e437-42.
- 432 4. Bjurstam N, Bjorneld L, Duffy SW, *et al.* The Gothenburg breast screening  
433 trial: first results on mortality, incidence, and mode of detection for women ages 39-  
434 49 years at randomization. *Cancer* 1997;80(11):2091-9.
- 435 5. Bjurstam N, Bjorneld L, Warwick J, *et al.* The Gothenburg Breast Screening  
436 Trial. *Cancer* 2003;97(10):2387-96.
- 437 6. Frisell J, Eklund G, Hellstrom L, *et al.* The Stockholm breast cancer screening  
438 trial--5-year results and stage at discovery. *Breast Cancer Res Treat* 1989;13(1):79-  
439 87.
- 440 7. Frisell J, Eklund G, Hellstrom L, *et al.* Randomized study of mammography  
441 screening--preliminary report on mortality in the Stockholm trial. *Breast Cancer Res*  
442 *Treat* 1991;18(1):49-56.
- 443 8. Frisell J, Lidbrink E, Hellstrom L, *et al.* Followup after 11 years--update of  
444 mortality results in the Stockholm mammographic screening trial. *Breast Cancer Res*  
445 *Treat* 1997;45(3):263-70.
- 446 9. Miller AB, To T, Baines CJ, *et al.* The Canadian National Breast Screening  
447 Study: update on breast cancer mortality. *J Natl Cancer Inst Monogr* 1997;  
448 10.1093/jncimono/1997.22.37(22):37-41.
- 449 10. Miller AB, To T, Baines CJ, *et al.* Canadian National Breast Screening Study-  
450 2: 13-year results of a randomized trial in women aged 50-59 years. *J Natl Cancer*  
451 *Inst* 2000;92(18):1490-9.
- 452 11. Miller AB, Wall C, Baines CJ, *et al.* Twenty five year follow-up for breast  
453 cancer incidence and mortality of the Canadian National Breast Screening Study:  
454 randomised screening trial. *BMJ* 2014;348:g366.
- 455 12. Tabar L, Fagerberg CJ, Gad A, *et al.* Reduction in mortality from breast  
456 cancer after mass screening with mammography. Randomised trial from the Breast  
457 Cancer Screening Working Group of the Swedish National Board of Health and  
458 Welfare. *Lancet* 1985;1(8433):829-32.

- 459 13. Tabar L, Vitak B, Chen HH, *et al.* The Swedish Two-County Trial twenty years  
460 later. Updated mortality results and new insights from long-term follow-up. *Radiol*  
461 *Clin North Am* 2000;38(4):625-51.
- 462 14. Tabar L, Vitak B, Chen TH, *et al.* Swedish two-county trial: impact of  
463 mammographic screening on breast cancer mortality during 3 decades. *Radiology*  
464 2011;260(3):658-63.
- 465 15. Andriole GL, Crawford ED, Grubb RL, 3rd, *et al.* Mortality results from a  
466 randomized prostate-cancer screening trial. *N Engl J Med* 2009;360(13):1310-9.
- 467 16. Andriole GL, Crawford ED, Grubb RL, 3rd, *et al.* Prostate cancer screening in  
468 the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial:  
469 mortality results after 13 years of follow-up. *J Natl Cancer Inst* 2012;104(2):125-32.
- 470 17. Atkin W, Wooldrage K, Parkin DM, *et al.* Long term effects of once-only  
471 flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible  
472 Sigmoidoscopy Screening randomised controlled trial. *Lancet*  
473 2017;389(10076):1299-1311.
- 474 18. Atkin WS, Edwards R, Kralj-Hans I, *et al.* Once-only flexible sigmoidoscopy  
475 screening in prevention of colorectal cancer: a multicentre randomised controlled  
476 trial. *Lancet* 2010;375(9726):1624-33.
- 477 19. Buys SS, Partridge E, Black A, *et al.* Effect of screening on ovarian cancer  
478 mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening  
479 Randomized Controlled Trial. *JAMA* 2011;305(22):2295-303.
- 480 20. Hocking WG, Hu P, Oken MM, *et al.* Lung cancer screening in the  
481 randomized Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial.  
482 *J Natl Cancer Inst* 2010;102(10):722-31.
- 483 21. Holme O, Loberg M, Kalager M, *et al.* Long-Term Effectiveness of  
484 Sigmoidoscopy Screening on Colorectal Cancer Incidence and Mortality in Women  
485 and Men: A Randomized Trial. *Ann Intern Med* 2018;168(11):775-782.
- 486 22. Holme O, Loberg M, Kalager M, *et al.* Effect of flexible sigmoidoscopy  
487 screening on colorectal cancer incidence and mortality: a randomized clinical trial.  
488 *JAMA* 2014;312(6):606-15.
- 489 23. Holme O, Loberg M, Kalager M, *et al.* Long-Term Effectiveness of  
490 Sigmoidoscopy Screening in Women and Men. *Ann Intern Med* 2018;169(9):663-  
491 664.



- 492 24. National Lung Screening Trial Research T, Aberle DR, Adams AM, *et al.*  
493 Reduced lung-cancer mortality with low-dose computed tomographic screening. *N*  
494 *Engl J Med* 2011;365(5):395-409.
- 495 25. Pinsky PF, Miller E, Prorok P, *et al.* Extended follow-up for prostate cancer  
496 incidence and mortality among participants in the Prostate, Lung, Colorectal and  
497 Ovarian randomized cancer screening trial. *BJU Int* 2019;123(5):854-860.
- 498 26. Pinsky PF, Prorok PC, Yu K, *et al.* Extended mortality results for prostate  
499 cancer screening in the PLCO trial with median follow-up of 15 years. *Cancer*  
500 2017;123(4):592-599.
- 501 27. Sandblom G, Varenhorst E, Lofman O, *et al.* Clinical consequences of  
502 screening for prostate cancer: 15 years follow-up of a randomised controlled trial in  
503 Sweden. *Eur Urol* 2004;46(6):717-23; discussion 724.
- 504 28. Sandblom G, Varenhorst E, Rosell J, *et al.* Randomised prostate cancer  
505 screening trial: 20 year follow-up. *BMJ* 2011;342:d1539.
- 506 29. Schoen RE, Pinsky PF, Weissfeld JL, *et al.* Colorectal-cancer incidence and  
507 mortality with screening flexible sigmoidoscopy. *N Engl J Med* 2012;366(25):2345-  
508 57.
- 509 30. Schroder FH, Hugosson J, Roobol MJ, *et al.* Screening and prostate-cancer  
510 mortality in a randomized European study. *N Engl J Med* 2009;360(13):1320-8.
- 511 31. Schroder FH, Hugosson J, Roobol MJ, *et al.* Screening and prostate cancer  
512 mortality: results of the European Randomised Study of Screening for Prostate  
513 Cancer (ERSPC) at 13 years of follow-up. *Lancet* 2014;384(9959):2027-35.
- 514 32. Moss SM, Wale C, Smith R, *et al.* Effect of mammographic screening from  
515 age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a  
516 randomised controlled trial. *Lancet Oncol* 2015;16(9):1123-1132.
- 517 33. Segnan N, Armaroli P, Bonelli L, *et al.* Once-only sigmoidoscopy in colorectal  
518 cancer screening: follow-up findings of the Italian Randomized Controlled Trial--  
519 SCORE. *J Natl Cancer Inst* 2011;103(17):1310-22.
- 520 34. Royston P, Parmar MK. An approach to trial design and analysis in the era of  
521 non-proportional hazards of the treatment effect. *Trials* 2014;15:314.
- 522 35. Royston P, Parmar MK. Augmenting the logrank test in the design of clinical  
523 trials in which non-proportional hazards of the treatment effect may be anticipated.  
524 *BMC Med Res Methodol* 2016;16:16.

- 525 36. Royston P. Power and sample-size analysis for the Royston–Parmar  
526 combined test in clinical trials with a time-to-event outcome. *The Stata Journal*  
527 2018;18(1):3-21.
- 528 37. Royston P, Choodari-Oskooei B, Parmar MKB, *et al.* Combined test versus  
529 logrank/Cox test in 50 randomised trials. *Trials* 2019;20(1):172.
- 530 38. Karrison TG. Versatile Tests for Comparing Survival Curves Based on  
531 Weighted Log-rank Statistics. *The Stata Journal* 2016;16(3):678-690.
- 532 39. Jacobs IJ, Menon U, Ryan A, *et al.* Ovarian cancer screening and mortality in  
533 the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised  
534 controlled trial. *Lancet* 2016;387(10022):945-956.
- 535 40. Menon U, Gentry-Maharaj A, Ryan A, *et al.* Recruitment to multicentre trials--  
536 lessons from UKCTOCS: descriptive study. *BMJ* 2008;337:a2079.
- 537 41. Jacobs I, Gentry-Maharaj A, Burnell M, *et al.* Sensitivity of transvaginal  
538 ultrasound screening for endometrial cancer in postmenopausal women: a case-  
539 control study within the UKCTOCS cohort. *Lancet Oncol* 2011;12(1):38-48.
- 540 42. UKCTOCS\_Group. *Long term impact of screening on ovarian cancer mortality*  
541 *in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS).*  
542 <http://ukctocs.mrcctu.ucl.ac.uk/long-term-impact/>.
- 543 43. Royston P, Parmar MK. Flexible parametric proportional-hazards and  
544 proportional-odds models for censored survival data, with application to prognostic  
545 modelling and estimation of treatment effects. *Stat Med* 2002;21(15):2175-97.
- 546 44. Alexander FE, Anderson TJ, Brown HK, *et al.* 14 years of follow-up from the  
547 Edinburgh randomised trial of breast-cancer screening. *Lancet*  
548 1999;353(9168):1903-8.
- 549 45. Hanley JA, Njor SH. Disaggregating the mortality reductions due to cancer  
550 screening: model-based estimates from population-based data. *Eur J Epidemiol*  
551 2018;33(5):465-472.
- 552 46. Andersson I, Aspegren K, Janzon L, *et al.* Mammographic screening and  
553 mortality from breast cancer: the Malmo mammographic screening trial. *BMJ*  
554 1988;297(6654):943-8.
- 555 47. Miller EA, Pinsky PF, Schoen RE, *et al.* Effect of flexible sigmoidoscopy  
556 screening on colorectal cancer incidence and mortality: long-term follow-up of the  
557 randomised US PLCO cancer screening trial. *Lancet Gastroenterol Hepatol*  
558 2019;4(2):101-110.

559 48. Pinsky PF, Yu K, Kramer BS, *et al.* Extended mortality results for ovarian  
560 cancer screening in the PLCO trial with median 15years follow-up. *Gynecol Oncol*  
561 2016;143(2):270-275.

562

## Table legends

**Table 1**: Summary of mortality analyses of randomised controlled cancer-screening trials

**Table 2**: Summary of choices and additional suggestions if not in concordance with A, B or C of the experts

**Table 3**: Summary of pros and cons of potential statistical tests that could be used when there is a time varying mortality difference (non-proportional hazards)

## Supplementary material legends

**Supplementary Material 1**: Cover Letter to Independent International Expert panel, Outline of Options, Comment Form

**Supplementary Table 1**: Summary of Responses from Independent International Group

**Table 1: Summary of mortality analyses of randomised controlled cancer-screening trials**

Trial name	Disease area	Country	No. of participants	Recruitment period	Number of screens	Screening period	Censorship date	Median FU from randomisation	Original analysis		LTFU analysis		No of years from randomisation to mortality reduction*
									Statistical analysis methodology	Final mortality reduction (95%CI)	Statistical analysis methodology (if different)	Final mortality reduction (95%CI)	
Two county	Breast	Sweden	162981	1977	4	1977-1984	end 1984	5.93 years (mean) (29 years? LTFU)	"Mantel-Haenszel" techniques - stratified by county and age	<b>RR=0.69 ; p=0.013</b>	Negative binomial regression, robust SEs for cluster randomization	RR=0.69 95% CI: 0.56-0.84; p=0.0001	~4 years (Figure 1)[14]
Malmo	Breast	Sweden	42283	1976-1978	5	1976-1986	end 1987	8.8 years (mean)	Relative risk (RR), test based CI	RR=1.29 95% CI: 0.74-2.25			No screening effect (no figure in analysis time)[46]
Gothenburg	Breast	Sweden	51611	1982-1984	4-5	1982-1991	end 1996	11.8 years (mean) (~14 years LTFU)	RR, poisson regression. Test based on Likelihood ratio	<b>RR=0.56, 95% CI: 0.31-0.99; p=0.046</b>	RR, poisson regression adjusted for birth cohort	RR=0.79 , 95% CI: 0.58-1.08; p=0.14	~0 years (Figure 1)[5]
Edinburgh	Breast	UK	54654	1978-1985	2-4 (depending on cohort)	1978-1988	1992	~9 years? 10 years max (12.8 years (mean) LTFU)	Logistic regression modified for cluster randomisation and stratified by age. ITT	RR= 0.82, 95% CI 0.61-1.11 [RR with LR??]	Same	RR=0.87 (95%CI: 0.70-1.06) [RR with LR?]	~6 years (Figure 2)[44]
UK Age Trial	Breast	UK	160921	1991-1997	7	1991-2004?	end 2004	10.7 years (17.7 years LTFU)	RRs, no detail. ITT	RR=0.83, 95% CI: 0.66-1.04; p=0.11	Poisson regression (presumably as before).	RR=0.88 , 95% CI: 0.74-1.04; no p-value	~3 years (Figure 2)[32]
ERSPC	Prostate	Europe (7 countries)	162 387 (in the core age group)	1991-2003	up to 3?	1991-2003	end 2006	9.0 years (13 years LTFU)	Poisson regression to estimate mortality ratio (RR), stratified by	<b>RR=0.80 (95% CI: 0.65-0.98; P = 0.04).</b>	Same	RR=0.79 (95%CI: 0.69 to 0.91) p=0.001	~7 years (Figure 2)[31]

									centre and age group. ITT				
SCORE	Colorectal	Italy	34292	1995-1999	1	1995-1999?	2006?	11.4 years	RRs based on average mortality rates (poisson distribution). ITT	RR = 0.78; 95% CI = 0.56 to 1.08			~5-6 years (Figure 2c)[33]
NORCCAP	Colorectal	Norway	98792	1999-2001	1	1999-2001	end 2011	10.9 years (14.2 years LTFU (mean))	HRs from Cox model, adjusted for age. ITT	HR= 0.73 [95%CI, 0.56-0.94]; p=0.02	Same, except primary analysis was now separate estimates for men and women	Men HR=0.63 (0.47 to 0.83) Women HR=1.01 (0.77 to 1.33)	~5-9 years (~3 years for men) (Figure 2c)[21]
PLCO	Prostate	USA	76693	1993-2001	4-6	1993-2005?	2008	11.5 years (14.8 years LTFU)	RRs assuming poisson distribution. ITT. No mention of WLR test and no p-value given subsequently	RR=1.13; 95% CI: 0.75 to 1.70	Same	RR=1.04 ; 95% CI: 0.87 to 1.24	no screening effect (Figure 1)[26]
PLCO	Lung	USA	154901	1993-2001	4	1993-2005?	end 2009	11.9 years	RRs assuming poisson distribution. Adjusted p for sequential analyses (interim). No mention of how p calculated	RR=0.99, 95% CI, 0.87-1.22; p=0.48			no screening effect (no figure)[20]
PLCO	Colorectal	USA	154900	1993-2001	2	1993-2004	end 2009	11.9 years (15.8 years)	Weighted (0,1) LR test with RRs assuming poisson distribution.	<b>RR= 0.74; 95% CI: 0.63 to 0.87; P&lt;0.001</b>	Same for RRs though notably no test/p-value	RR= 0.75, 95% CI 0.66–0.85	~3 years (Figure 2a)[47]

									Adjusted p for sequential analyses (interim)				
PLCO	Ovarian	USA	78216	1993-2001	4-6	1993-2005?	28th Feb 2010	12.4 years (14.8 years LTFU)	Weighted (0,1) LR test (one-sided?) with RRs assuming poisson distribution. Adjusted p for sequential analyses (interim)	RR= 1.18; 95% CI, 0.82-1.71 - sequentially adjusted. No p-value reported possibly because test was 1-sided?	Same for RRs though notably no test/p-value (also added a Cox model)	RR=1.04 (95% CI: 0.87–1.24)	no screening effect (Figure 1)[48]
NLST	Lung	USA	53454	2002-2004	3	2002-2007	end 2009	5.4 years (mean)	Rrs assuming poisson distribution. Adjusted p for sequential analyses. Weighted	<b>RR=0.80 (95% CI: 0.73-0.93; P = 0.004).</b>			~1.5 years (Figure 1B)[24]
UK Flexible Sigmoidoscopy Screening Trial (UKFSST)	Colorectal	UK	170□034	1994-1999	1	1994-1999	31st Dec 2014	17.1 years	HRs from Cox model. ITT	<b>HR=0.57 (0.45–0.72); HR=0.56 (0.45–0.69) CRC verified</b>	Same	HR=0.59 (0.49–0.70)	~3 years (Figure 1G)[17]
Canadian National Breast Screening Study (CNBSS)	Breast	Canada	50430	1980-1985	5	1980-1985	end 1991	8.5 years (mean) (25 years LTFU)	T-test on difference of proportions	RR=1.36 (95% CI: 0.84-2.21)	Cox PHs model	HR=0.99 (95% CI 0.88 to 1.12; P=0.87)	no screening effect (Figure 3)[11]

\* Estimate of mortality curve separation comes from visual inspection of appropriate published mortality plot if provided. The Figure number and paper reference are given to allow the reader to make their own judgement

Footnote: FU - Follow up; LTFU - long term follow up; RR - rate ratio; HR - hazard ratio; ITT - intention to treat analysis; LR – log-rank

**Table 2: Summary of choices and additional suggestions if not in concordance with A, B or C of the experts**

<b>Expert</b>	<b>Expertise</b>	<b>Choice</b>	<b>Additional suggestions</b>
EX1	Biostatistics, public health	A	Suggests only include cancers diagnosed from period of intervention.
EX2	Biostatistics, clinical trials and cancer research	A	
EX3	Statistics	A	Ticked 'alternative' but suggested hybrid of A for testing and C for estimation – interpreted as A
EX4	Cancer epidemiology, prevention and screening	Change analysis	Suggested 'number needed to screen'.
EX5	Biostatistics, cancer epidemiology	Change analysis	Did not complete form but indicated choice by email, test based on difference of restricted mean survival time (RMST).
EX6	Biostatistics and epidemiology	Change analysis	Suggested splitting data into yearly bins and assess HR in each, possibly with smoothing. Avoid single HR.
EX7	Biostatistics, clinical trials and cancer research	C	Did not complete form but indicated choice by email. Prefers more parsimonious model with interpretable parameters.
EX8	Biostatistics, clinical trials	C	
EX9	Biostatistics, public health	C	Prefers more parsimonious model with interpretable parameters.
EX10	Cancer epidemiology, public health	C	Also suggests 'versatile weighted log-rank test'
EX11	Statistics, public policy	C	
EX12	Biostatistics	-	Did not respond within timeframe



**Table 3: Summary of pros and cons of potential statistical tests that could be used when there is a time varying mortality difference (non-proportional hazards)**

METHOD	PROS	CONS
Weighted log-rank test	<p>Not model-based</p> <p>Known to improve power in situations of non-PH.</p> <p>Most widely used and established test for non-PHs in clinical trials</p>	<p>Need to formally pre-specify the expected mortality differences over time (functional form of the HR) for the test to have statistical validity. This may prove difficult given that differences will depend on the natural history of the cancer, screening strategy, number of screens, years of follow-up etc.</p> <p>There is an associated risk of mis-specifying the form of the HR, and simulations suggest incorrectly assuming a late effect, for example, may incur a greater penalty than assuming PHs under early or late effects [43, 44].</p> <p>Subjects' deaths are given a differential (and arbitrary) weighting which may be hard to justify. A further conceptual problem with weights based on the data is that if a trial subsequently reports again, the weight allocated to each event will change, likely significantly.</p>
Flexible parametric model such as the Royston-Parmar model (cubic splines) or fractional polynomial survival model (joint test of all screen arm related terms)	<p>No need to pre-specify specific functional form mortality effect</p> <p>Can mimic a non-PH function to almost arbitrary degree.</p> <p>Allows one to accurately describe the hazards and their ratio over time.</p>	<p>No precedence for use as primary analysis in RCTs</p> <p>Flexibility make it easy to over fit and include random data artefacts.</p> <p>Power properties not well known. Will lose power with too many model parameters.</p> <p>Need to pre-specify number of knots/degrees of freedom and placement of knots for RP model. FP model requires choice of selection of powers and degree. Can be guided by information criteria but then data dependent, and may reflect artefacts.</p>

	Relatively easy to fit	Test, as proposed, considers if mortality curves are 'different'. Significant result could theoretically result from crossing curves, even curves with no difference in AUC.
Weibull model (with separate shape parameters for group)	Can reflect simple time-varying differences in mortality curves succinctly Easy to fit	Unlikely to capture more complex curves sufficiently. All hazard functions must be monotonic (constant decrease or increase)
Cox model with time varying coefficient (TVC)	Extension of Cox model, so perhaps more readily acceptable given prior use  Able to incorporate non-PHs without specifying differences in mortality curves (functional form). For example, choose linear function of time, then time-varying effect could be linear decreasing or increasing.  No need to consider baseline hazard function	Need to pre-specify function of time that the non-PHs apply to – usually a simple linear or log function of time  Interpretation not straightforward  Awkward and (very) time-consuming to fit (splits data at each failure)  No definite agreement on test of significance. Could be similar to the joint test on 2 degrees of freedom.
Difference in restricted mean survival time	No need to be model-based, can use non-parametric estimation.	Need to pre-specify choice of time restriction, possibly including initial time $t_0$ , as well as final time limit $t_1$ .

(RMST)	<p>Can reflect any time-varying difference in mortality - estimate of RMST difference graphically corresponds to the difference in area between the respective survival curves.</p> <p>Do not need to speculate on particular form of time varying difference in mortality. However choice of time restriction may depend on expectation of difference (HR functional form).</p> <p>Gives a meaningful single summary estimate even with non-PHs</p>	<p>Time consuming to estimate, including standard error.</p> <p>As the test looks for differences in AUC, survival curves that come back together can result in a significant test result.</p>
Combined test (of Cox test with a permutation test based on RSMTs on 2 df)	<p>Simulations suggest power not much lower than Cox alone under PHs and more powerful in more situations than joint test [43, 44].</p> <p>Enhanced power for early effect</p>	<p>Difficult to explain</p> <p>Time-consuming to fit (permutation test).</p> <p>Issues of RMST (see above) – choice of time restriction</p> <p>Simulations suggest not powerful for late effects</p>
Joint test (of Cox proportional screen arm effect + Grambsch-Thurneau non-PH test on 2 df)	<p>Test based on results of the Cox model (screen arm effect and the Schoenfeld residuals), so perhaps more readily acceptable given prior use of the Cox-model</p> <p>Relatively simple test (with degree of intuitiveness), but more powerful than just screen arm effect under non-PHs</p>	<p>Simulations suggest better under late effects but not good power for early effects [43, 44].</p>
Combination tests such as Versatile Test	<p>Not model-based</p>	<p>Appears complicated (need for reference to a correlated multivariate z-distribution for test statistic)</p>

<p>(maximum test statistic of 3 weighted tests-early, PHs, late effects) or “max-combo” (also includes ‘middle’ effects)</p>	<p>Provides good power in all situations, covers bases with small price in efficiency</p> <p>Best choice if one wants to be agnostic of specifying the time varying mortality difference</p>	<p>Not the most powerful test.</p> <p>Can feasibly reject the null hypothesis both in favour of the study arm and of the control arm using the same data.</p>
--	--	---