

1 **TITLE**

2 Pitfalls in Time-to-Event Analysis of Registry Data: A Tutorial based on Simulated and Real
3 Cases.

4

5 **AUTHORS**

6 Mickaël ALLIGON¹, BSc, Nizar MAHLAOU^{1,2}, MD, MPH, PhD, Olivier BOUAZIZ³, PhD

7

8 **AFFILIATIONS**

9 1. French National Reference Center for Primary Immune Deficiencies (CEREDIH),
10 Necker Enfants Malades University Hospital, Assistance Publique-Hôpitaux de Paris
11 (APHP), Paris, France.

12 2. Immuno-Haematology and Rheumatology Unit, Necker Enfants Malades University
13 Hospital, Assistance Publique-Hôpitaux de Paris (APHP), Paris, France.

14 3. Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

15

16 **CORRESPONDING AUTHOR**

17 Olivier Bouaziz, Université Paris Cité, CNRS, F-75006 Paris, France.

18 E-mail: olivier.bouaziz@parisdescartes.fr

19 Tel: +33-176-530-379

20

21 **Standfirst**

22 Based on both the use of real-world data and a simulation approach, this tutorial
23 article looks at how overlooking right censoring, left truncation, competing events and
24 recurrent event methods in analyses of time-to-event data can lead to suboptimal or biased
25 estimations. The simulation approach enables a comparison with the truth and thus
26 concrete estimations of bias for methods that (i) do not consider right censoring (leading to
27 underestimation of survival), (ii) do not consider left truncation (leading to overestimation of
28 survival), (iii) treat competing risks as right-censoring (leading to overestimation of survival),
29 and (iv) consider only the first of a series of recurrent events for a given individual.

30

31 **Contributors and sources**

32 In this tutorial article, we look at how overlooking right censoring, left truncation,
33 competing events and recurrent event methods in analyses of time-to-event data can lead to
34 suboptimal or biased estimations.

35 The present work is novel because it uses both real-world data and a simulation
36 approach. This enables comparisons with the truth and thus the provision of concrete
37 estimations of the biases associated with various methods.

38 Mickaël ALLIGON is a junior statistician at the French National Registry for Primary
39 Immunodeficiencies (CEREDIH). He trained under the supervision of CEREDIH director Nizar
40 MAHLAOUI MD, PhD (a clinician and epidemiologist with 18 years of experience) and Olivier
41 BOUAZIZ PhD (a senior statistician in the Laboratory of Applied Mathematics in Health
42 Sciences, with great expertise in survival analyses). All the authors contributed equally to
43 analyzing the data and drafting the manuscript. MA and OB performed the statistical
44 analyses.

45 OB is the guarantor of the article. The CEREDIH registry was the source of the real-
46 world data. This research did not receive any specific funding from agencies or organizations
47 in the public, commercial, or not-for-profit sectors .

48

49

50

51 **KEYWORDS**

52 registries; survival analysis; left truncation; competing risks; recurrent events; time-
53 to-event analysis; right censoring; primary immunodeficiencies; rare diseases.

54

55 **ABBREVIATIONS:**

56 CID: combined immunodeficiency

57 CIF: cumulative incidence function

58 CVID: common variable immunodeficiency

59 HSCT: hematopoietic stem cell transplantation

60 MSM: multistate model

61 PID: primary immunodeficiency

62 SCID: severe combined immunodeficiency

63

64

65 **Key messages box**

66 When comparing naïve approaches and the proper methodology, we show that:

- 67 • Not considering right censoring leads to underestimation of survival
- 68 • Not considering left truncation leads to overestimation of survival
- 69 • Treating competing risks as right-censoring leads to overestimation of survival
- 70 • Appropriate recurrent event methods allow to study all events for each patient and
- 71 not only account for the first event.

72

73 **Competing interests declaration**

74 The authors declare that they have no conflicts of interest with regard to the present manu-

75 script.

76

77 **ABSTRACT**

78 Survival analysis (also referred to as time-to-event analysis) is the study of the time
79 elapsed from a starting date to some event of interest. In practice, these analyses can be
80 challenging and, if methodological errors are to be avoided, require the application of
81 appropriate techniques. By using simulations and real-life data based on the French national
82 registry of patients with primary immunodeficiencies (CEREDIH), we sought to highlight the
83 basic elements that need to be handled correctly when performing the initial steps in a
84 survival analysis. We focused on right censoring, left truncation, competing risks, and
85 recurrent events. Our simulations show that ignoring these aspects induces a bias in the
86 results; we then explain how to analyze the data correctly in these situations. Rare disease
87 registries are extremely valuable in medical research. We discuss the application of
88 appropriate methods for the analysis of time-to-event from the CEREDIH registry. The
89 objective of this tutorial article is to provide clinicians and healthcare professionals with
90 better knowledge of the issues facing them when analyzing time-to-event data.

91

92 **1.INTRODUCTION**

93 The collection of patient data in academic- and/or industry-led registries is one of the
94 key elements of medical and translational research. The advent of many disease registries
95 (including registries for rare diseases) in the early 2000s helped to improve our knowledge of
96 disease occurrence (incidence and prevalence being the key epidemiological factors most
97 frequently assessed), the natural history of those diseases, and the effectiveness and safety
98 of various procedures and therapies (such as stem cell therapy) at the national and
99 international levels (1,2). Furthermore, registry data is of value in (i) designing national or
100 international orphan drug trials, (ii) standardizing patient management, and thus (iii)
101 improving the patients' health-related outcomes and quality of life (3).

102 In France, the creation of a series of five-year national rare disease plans and national
103 reference centers for rare diseases in 2004 prompted the creation of registries for single
104 diseases or groups of diseases. The CEREDIH French national reference center for children
105 and adult patients with primary immunodeficiencies (PIDs) created France's first national
106 registry for these conditions. The registry complied with the official criteria: the continuous,
107 exhaustive registration of cases (defined as a condition, disease, health issue or healthcare
108 procedure such as surgery, hematopoietic stem cell transplantation [HSCT], etc.) in a defined
109 geographical area by a team of trained professionals (4,5).

110 PIDs constitute a large, heterogeneous group of more than 500 mostly inherited
111 diseases that expose patients to a greater risk of infections, severe allergies,
112 autoimmune/inflammatory manifestations, and/or malignancies (6,7). Furthermore, PIDs
113 can be classified as deficiencies of the adaptive immune system (subdivided into T-cell
114 deficiencies and B-cell deficiencies) and deficiencies of the innate immune system. The T-cell

115 deficiency group includes severe combined immunodeficiencies (SCIDs, also known as “boy-
116 in-a-bubble diseases”) and other combined immunodeficiencies (CID). The B-cell deficiency
117 group can be subdivided into common variable immunodeficiencies (CVIDs) and
118 hypogammaglobulinemias (also referred to as “non-CVIDs”, which also include
119 agammaglobulinemia).

120 Since the CEREDIH registry’s inception in 2005, we have sought to include all patients
121 diagnosed with a PID in France (8,9). As of June 2nd, 2022, more than 8,500 patients had
122 been registered. 1,563 of these patients are now deceased.

123 CEREDIH uses the European Society for ImmunoDeficiencies platform to enter data.
124 All European Society for ImmunoDeficiencies registry documenting centers share a common
125 dataset, and CEREDIH has a complementary, specific dataset. Overall, the collected data
126 encompass several medical variables recorded at one or more timepoints in the patient’s
127 life: the symptoms that led to the diagnosis of PID, the main PID-related clinical
128 manifestations (malignancies, autoimmune/inflammatory manifestations, allergies, and
129 infections), the main PID-related therapies (mainstay therapies like immunoglobulin
130 replacement therapy and curative therapies like HSCT, thymus transplant, and gene
131 therapy), and the cause of death. After inclusion, all the patient files are updated every two
132 years or more frequently. Since the dates of these main events are recorded, it is possible to
133 construct time-to-event variables for a given event.

134 Alongside data completeness, data quality is essential at all stages: at registration and
135 through follow-up documentations as well as through the implementation of relevant and
136 efficient quality control and data management procedures. Furthermore, the entry of
137 multiple time points per patient ensures that information is as up to date as possible and (ii)

138 the quality of the indicators produced by the statistical analyses (especially survival data
139 indicators) is as high.

140 Underfunding is a concern because it leads to issues of registry sustainability, which
141 include (but are not limited to) understaffing, impairing collaboration with statisticians who
142 have relevant expertise in this field. As a result, some studies may include a suboptimal or
143 even biased statistical methodology, which in turn can lead to incorrect results. One of the
144 primary roles of a registry is to highlight overall trends and relationships in data (10).
145 Research groups can then use specific methods to validate or reject medical hypotheses (e.g.
146 with regard to disease mechanisms, survival, covariates leading to one or more
147 comorbidities of interest, etc.). Therefore, the use of incorrect statistical methods that do
148 not consider potential bias in the data might lead to unreliable estimations and harmful
149 medical decisions. Improving patient management is one of the main goals of patient
150 registries and involves the analysis of time-related data. In the field of health, survival
151 analyses are among those that suffer the most from statistical bias; this is primarily due to
152 the use of inappropriate approaches that do not consider censoring.

153 Famous examples of incorrect statistical analysis often involve immortal bias. One
154 study (11) found that Academy-Award-winning actors and actresses lived almost 4
155 years longer than their less successful peers. However, a subsequent reanalysis of the study
156 data failed to find a significant difference in survival between the winners and non-winners
157 and showed that the first analysis suffered from immortality bias: the *“winners had to
158 survive long enough to win”, while “performers who did not win had no minimum survival
159 requirement, and some died before some winners had won, that is, before some “longevity
160 contests” could begin.*” Queen Elisabeth II even joked about immortality bias during her 80th

161 birthday celebration in 2006: *“As Groucho Marx once said, ‘Getting older is no problem. You*
162 *just have to live long enough.’ ”*. However, the issue can be more serious when it affects
163 medical research. Using skin cancer as a marker of sun exposure, researchers had concluded
164 that *“having a diagnosis of skin cancer was associated with less myocardial infarction, less*
165 *hip fracture in those below age 90 years and less death from any cause.”* (12) Following this
166 study, two other researchers – both specialists in the analysis of time-to-event data –
167 pointed out the presence of immortality bias in the first analysis: *“in order to get a skin*
168 *cancer diagnosis, and thus become a member of the skin cancer group, it is at least necessary*
169 *to survive until age of diagnosis. For those in the skin cancer group it is impossible to die until*
170 *the age of diagnosis of the cancer, the so-called immortal person-time.”* Another pitfall
171 pertaining to the study of time-to-event data involves competing risks: if the event of
172 interest is non-lethal (such as disease relapse, an infection, or the occurrence of cancer) and
173 death can also occur, the latter must be treated as a competing event (i.e. an event that
174 precludes the occurrence of the event of interest). A common mistake then consists in
175 treating death as censoring, which amounts to assuming that deceased patients are still at
176 risk of experiencing the event of interest. For example, researchers have compared the risk
177 of relapse among HSCT recipients, using the European Group for Blood and Marrow
178 Transplantation (EBMT) dataset (13). They reported that treating death as a censoring
179 variable resulted in a significant overestimation of the probability of relapse: the estimated
180 5-year probability of relapse was 0.515 in the flawed analysis and 0.316 when death was
181 correctly taken into account as a competing risk.

182 Here, we describe the classical methods used to deal with right-censoring, left
183 truncation, competing events, and recurrent events. We first apply a simulation-based
184 approach and then refer to CEREDIH registry data. Our objective is to make clinicians and

185 healthcare professionals more aware of the issues facing them in analyses of time-to-event
186 data.

187 All analyses were conducted with R software and its *{survival}* library. All the codes
188 and a randomized version of the CEREDIH dataset are available on GitHub
189 (<https://github.com/Malligon/Pitfalls-in-Time-to-Event-Analysis-for-Registry-Data>).

190

191 **2. HOW TO PLAN YOUR SURVIVAL ANALYSIS CAREFULLY**

192

193 Survival analysis (also called time-to-event analysis) is the study of the time elapsed
194 from a starting date to an event of interest.

195 Firstly, it is important to precisely define the event of interest, the time scale, the
196 study entry point, and the risk set. The event of interest can be death, recovery, occurrence
197 of a disease, relapse or any medically relevant event. The time scale refers to the time unit
198 used (usually years or months). Study entry is the starting point of the study (birth,
199 treatment initiation, enrolment, etc.). If, for example, a study is designed to analyze survival
200 (in days) after treatment, study entry will be the time at which the patient took his/her
201 treatment, and the time scale will be days. Lastly, the risk set is defined as the pool of
202 patients at risk of experiencing the event of interest. A patient is included in the risk set at a
203 specific time if he/she can experience the event of interest at that time; this means
204 particularly that a patient can enter and leave the risk set at any time.

205 A classical phenomenon in time-to-event analysis is the presence of incomplete data.
206 This can be caused by right-censoring, left truncation, or both. These data might also include
207 recurrent and/or competing events. Failure to take these concepts into account may lead to
208 incorrect estimations and misleading conclusions. Below, we present these four statistical
209 concepts and we explain how they can be handled by properly adjusting the risk set in each
210 case.

211 For some individuals, the exact time of occurrence of the event of interest is not
212 known; instead, an earlier time is observed, and it is only known that the event of interest
213 will occur after this observed time. This is **right-censoring**, which is classically taken into
214 account with the Kaplan-Meier estimator.

215

216 **Left truncation** is a phenomenon that often occurs in time-to-event analysis in which
217 individuals are followed up only from a time after study entry (called the *truncation time*)
218 and not from study entry onwards. In such a case, individuals are observed conditionally on
219 having not experienced the event of interest before the truncation time. In order to avoid
220 biased estimates, those data need to be appropriately taken into account by modifying the
221 risk set in the Kaplan-Meier estimator.

222 **Competing risks** methods are involved in a situation that occurs when another event
223 may preclude the observation of the event of interest. This is typically the case when the
224 competing risk is death and the event of interest is the occurrence of a disease, remission,
225 the onset of cancer, etc... While censored data indicates that the true event of interest will
226 occur after the censoring time, the true event of interest can no longer occur after a
227 competing risk. A common error consists in treating competing events as censored data in
228 the calculation of the survival function of the event of interest. This leads to overestimation
229 of the distribution of the event time. The correct approach consists in estimating the
230 cumulative incidence function (CIF), using specific methods.

231 **Recurrent event** data occur when an individual can experience the event of interest
232 several times during his/her lifetime. This can happen for the study of recurrent infections,
233 hospital admissions, cancer relapses, etc... Furthermore, recurrent event data often include
234 a competing event (referred to as the terminal event), which is typically death. For these
235 data, a different quantity may be of interest, such as the expected cumulative number of
236 recurrent events experienced by a patient up to a given time point.

237

238

239

240

TOOLBOX 1: How to plan your survival analysis carefully.

241

- Focus on data completeness during a chosen time period, rather than focusing on its length.
- Always start by defining the event of interest, the time scale, study entry, and the risk set.
- Assess right-censoring, left truncation, and competing risks in advance and use dedicated methods to analyze those data.
- Consider a recurrent event analysis if patients can encounter the event of interest more than once during the study.
- **In general: always consider what might occur, rather than what has been observed.**

242

243

244

245

246 **3. A SIMULATION-BASED APPROACH**

247

248 **A-Right-censoring**

249 As mentioned above, a time-to-event analysis will usually include right-censored
250 data. Right-censoring can mainly occur for two reasons: (i) the patient has not yet
251 experienced the event of interest by the time the study ends, or (ii) the patient is lost to
252 follow-up during the study period (i.e. drop-out). Because such censored times are smaller
253 than the true event times, treating the censored observations as completely observed data
254 will underestimate the distribution of the true event times. In contrast, larger and smaller
255 times of interests will respectively tend to be more or less subject to right-censoring. As a
256 result, keeping only the uncensored observations will result in underestimation of the true
257 event time distribution. These two naïve approaches (treating censored data as completely
258 observed, or removing censored data) show that dedicated methods (namely the Kaplan-
259 Meier estimator, in the case of right-censoring) are needed in this context.

260 When dealing with time-to-event data with right-censoring, the observations for an
261 individual consist of two variables: the observed time and the censoring status (or censoring
262 indicator). The latter variable is binary and indicates whether the observed time is the time
263 of interest or the censored time - a time that is known to be smaller than the time of
264 interest. In order to simulate these types of data for each individual, one needs to: (i)
265 simulate the time of interest; (ii) simulate a censoring time; (iii) measure the minimum
266 between the two times, which is the observed time; and (iv) determine whether the time of
267 interest is smaller than the censoring; if so, the censoring status is equal to 1; if not, the
268 censoring status is equal to 0.

269 In the **Appendix**, we present a simple code that simulates these data as described above. We
270 chose a Weibull distribution for the true event time (with a shape of 2 and a scale of 30) and
271 a uniform distribution over the interval [0;85] for the censoring variable. On average, this
272 choice of parameters will result in 31.24% of censored data. A sample of size 1000 is
273 simulated. It should be noted that a seed was arbitrarily chosen, so that the data can be
274 reproduced easily (see the **Appendix** for more details). **Table 1** shows the first 10 individuals
275 simulated using this code.

276 **In the absence of covariates, the standard quantity of interest in survival analysis** is
277 the **survival function** which represents the probability that the event of interest has not yet
278 occurred at any given time point. The survival function is classically estimated using the
279 Kaplan-Meier estimator (14). Starting from the grid of time points t_1, \dots, t_k consisting of all
280 the uncensored times, the estimator is computed recursively. For a given t_k , the Kaplan-
281 Meier estimator is equal to its value at the previous time point $t_{(k-1)}$ times $(1-d_k/R_k)$, where d_k
282 is the number of uncensored events that occurred at time t_k , and R_k is the number of “at
283 risk” individuals at time t_k (defined as the number of individuals that have not yet
284 experienced the event of interest or who have not yet been censored). In this formula, the
285 estimator is initialized at the value 1 for time equal to 0. In survival analysis, the notion of
286 individuals being “at risk” at a given time is essential. It is only through this risk set that
287 censoring is accounted for, and it is crucial that this set does not include periods of time
288 during which the event of interest cannot occur. The d_k/R_k ratio is called the hazard rate or
289 hazard risk estimator and represents the estimation for the risk of experiencing the event of
290 interest at time t_k , given that this event has not yet occurred. By using the *survfit* function in
291 the R package ‘survival’, one can compute the Kaplan-Meier estimator for the previously
292 generated dataset. This estimator can be compared with the naïve approach described in

293 the second section, which consisting in removing the censored observations and computing
294 1 minus the empirical distribution function for this subsample (**Figure 1**; see the **Appendix**
295 for more details). Clearly, the Kaplan-Meier estimator produces a very accurate estimation,
296 whereas the naïve approach gives a biased estimation. As expected, the survival function is
297 underestimated when censoring is ignored. For example, the true quantiles (Q_1 , Q_2 and Q_3)
298 for the variable of interest are 16.14, 25 and 35.37, respectively. The three quantiles
299 estimated from the Kaplan-Meier estimator are 16.12, 25.09 and 35.95, respectively, while
300 those given by the naïve estimator are 13.95, 22.06 and 30.42, respectively. By way of an
301 example, one can consider a study in which the event of interest is death: including only
302 observed deaths in the analysis (and thus ignoring censored data due to the end of study or
303 drop-out) will result in underestimation of the survival function. This is because at all time
304 points, the risk set needs to include the censored observations that have not yet occurred.
305 Whereas the number of observed events of interest is the same (d_k in the Kaplan-Meier
306 estimator), the number of individuals at risk (R_k in the Kaplan-Meier estimator) should be
307 increased.

308

309

310

TOOLBOX 2: Right censoring

- Right censoring is extremely common in time-to-event analyses.
- The Kaplan-Meier method is the standard approach for estimating the survival function for right censored data.
- Ignoring right censoring leads to underestimation of the survival function.
- When observations are censored, the event of interest will happen at a later, non-observed time.

311 **B. Left truncation**

312 Another frequent phenomenon in time-to-event analysis is left truncation (delayed
313 entry), when individuals are followed from a later time (the truncation time) and not from
314 the starting point. In such a case, individuals are observed conditionally on having not yet
315 experienced the event of interest before the truncation time. Again, in order to avoid biased
316 estimates, those data need to be taken into account appropriately by modifying the risk set
317 in the Kaplan-Meier estimator. While right-censoring is often correctly taken into account in
318 the analysis of time-to-event data, left truncation is more difficult to apprehend and is
319 therefore sometimes overlooked. Not taking into account left truncation results in an
320 immortality bias because individuals are considered to be at risk before the truncation time
321 but cannot die – if death is the event of interest – before the truncation time. This is typically
322 the case when the time scale is age since, very often, patients cannot be followed up from
323 birth. In that case, it is important to take into account the data observation scheme:
324 depending on the study, an individual will start to be followed up at the time of diagnosis, at
325 the date when the treatment started, or at some other time. Individuals having experienced
326 the event before they started to be followed up will never be observed. If the time scale is
327 age and the patient enters the study at a specific time, then he/she should not be part of the
328 risk calculation for earlier times.

329 Left truncation can be easily taken into account by modifying the risk set in the
330 Kaplan-Meier estimator. At a given time point, an individual should be in the risk set if
331 he/she (i) has not yet experienced the event, (ii) has not yet been censored, and (iii) the
332 truncation time occurred earlier than the time point.

333 It is important to stress that ignoring left truncation would result in overestimation of
334 the survival function because the risk set would be too large at time points where all
335 patients have not yet entered the study. Since the Kaplan-Meier estimator is computed in a
336 recursive way, this bias for initial time points will have an impact on all later times, and this
337 incorrect survival function will be overestimated.

338 Using the same simulation scheme as before, we generated a truncation variable
339 with a uniform distribution over the interval [0;50]. As a result, 42.4% of the observations
340 are not observed because the event of interest occurred before the truncation time. We
341 then estimated the survival function by applying two approaches based on the Kaplan-Meier
342 estimator: the correct one that modifies the risk set according to the truncation variable,
343 and a naïve approach in which left truncation is ignored. Modifying the risk set in the Kaplan-
344 Meier method is easily achieved in the survival library by using the start and stop variables
345 instead of the usual observed time variable. The start and stop variables correspond
346 respectively to the truncation time and the observed time (***Figure 2***; see the ***Appendix*** for
347 more details).

348 One can see clearly that the naïve approach overestimates the survival function. For
349 example, the true quantiles of order 0.25, 0.5, 0.75 for the variable of interest are equal to
350 16.14, 25, and 35.37, respectively. The estimated quantiles from the Kaplan-Meier
351 estimator when left truncation is taken into account are equal to 15.14, 24.44, and 35.43,
352 respectively, while the estimated quantiles from the naïve estimator are equal to 22.58,
353 31.25 and 42.14, respectively.

354 In summary, it is important to check that a survival analysis' risk set is well defined. In
355 other words, the researcher should ask him/herself "Is there a period of time during which

356 the individuals cannot experience the event of interest?”. If so, then the risk set needs to be
357 modified accordingly by using the start and stop variables. Similarly, it is important to choose
358 an appropriate time scale for use in practice. Most of the time, this choice will be based on
359 medical considerations. Does it make more sense to study the risk of death on the age time
360 scale? Or should the scale be the time elapsed since treatment allocation? In the second
361 scenario, a wide age range might make it necessary to also adjust for age. In the first scenar-
362 io, it is very likely that the data will suffer from left truncation. Lastly, it should be noted that
363 left truncation might deteriorate the performance of the Kaplan-Meier estimator when the
364 risk set is too small for short time periods. A small risk set will result in a high hazard rate and
365 a high variance of the hazard rate. Given that errors at early times will have an impact on all
366 future time points, this issue can be problematic. Some other options for managing this
367 problem can be found in the literature (15).

368

TOOLBOX 3 : Left truncation

369

- Left truncation is very common in registry analyses - especially when patients are followed up from birth.
- Left truncation is a specific type of immortal time bias.
- The Kaplan-Meier estimator accommodates with left truncation by adjusting the risk set (adding patients or removing them) at a given time point.
- Ignoring left truncation will lead to overestimation of the survival function.
- Early events may impact and bias the survival function if the risk set is too small at early time points.

370 **C. Competing risks**

371 As mentioned above, competing risks occur when another event may preclude the
372 observation of the event of interest. This is typically the case when the event of interest is
373 not terminal, e.g. the occurrence of an infection or a diagnosis of cancer. Death is then a
374 competing event and if it occurs in the dataset, it must be properly taken into account. A
375 common mistake is to treat death as a censoring variable. The major difference between a
376 competing event and censoring is that the event of interest may occur after the censoring
377 timepoint (even though it is not observed) but will never occur after a competing event. If
378 the competing event is death and the event of interest is cancer, then it is clear that a
379 patient can no longer develop a cancer after he/she had died. The Kaplan-Meier estimator
380 treats censoring as a variable that stops the observation of future events for the patient but
381 includes the information that the event of interest will occur after the censoring variable.
382 Consequently, computing a survival curve using the Kaplan-Meier estimator in a competing
383 risk situation where death is treated as a censoring variable will give a biased estimation.
384 Since dead individuals will remain “at risk” in the computation of the survival function, the
385 estimate will be biased upwards and the survival curve will be overestimated.

386 Competing events are often not correctly analyzed because they can be treated as a
387 censoring variable when estimating the hazard rate. This is a computational trick that works
388 well because the hazard rate is a quantity defined for an infinitesimally short period of time.
389 In other words, studies that use the Cox model to evaluate the effect of one or more
390 covariates on the event of interest might treat the competing event as a censored variable
391 (16). This approach will provide correct estimates of hazard ratios. However, this practice is
392 no longer appropriate for estimating cumulative quantities, such as the cumulative hazard

393 function or the survival function. Since the Cox model is beyond the scope of this paper, we
394 shall not discuss this issue further.

395 It should also be noted that the last example when cancer is the event of interest and
396 death is the competing event is more precisely an illness-death situation. Strictly speaking,
397 competing event situations encompass data for which the events of interest are mutually
398 exclusive (17). A typical example is when different causes of death are recorded and
399 analyzed. Again, one cause of death can only occur if the other cause of death has not yet
400 occurred, and this has to be properly taken into account in both scenarios. In the
401 cancer/death example, the death event might be studied simply by computing the Kaplan-
402 Meier estimator because cancer does not preclude the occurrence of death. We
403 nevertheless chose to simplify the presentation by considering this example with cancer and
404 death, because the illness-death model (a particular example of a multistate model (MSM))
405 is beyond the scope of this article (17). Furthermore, situations in which cancer is of interest
406 and individuals are also at risk of death are frequently encountered in registry data. This will
407 be illustrated below on the CEREDIH dataset.

408 In the presence of competing risks, the quantity of interest is usually the CIF. For the
409 cancer example, the CIF is simply the probability of experiencing a cancer before any time
410 point. For a given time t it is computed by cumulating for all time points t_k occurring before
411 t , the product of the hazard risk for the event of interest (computed as the ratio d_k/R_k at t_k ,
412 where the risk set R_k includes individuals that have not yet experienced any of the different
413 types of events and have not yet been censored) and the probability to have “survived” up
414 until time t_k . This last quantity is basically the Kaplan-Meier estimator for the compound
415 event composed of all the types of event; in other words, it is the Kaplan-Meier estimator

416 where the event of interest is the first event among all competing events). This estimator
417 can be calculated from the survival library by simply considering the status variable as a
418 factor with more than two levels: one level (always the first) for censoring and the other
419 levels for the competing events (18).

420 Lastly, given that the competing event precludes the occurrence of the other event, it
421 is good practice to always display the CIFs of all the competing events as well as the CIF of the
422 quantity of interest (19). This is important because otherwise, the CIF of the quantity of
423 interest might be misleading. A low risk of experiencing an event might simply be due to the
424 fact that the patients are at high risk of experiencing the competing event. Taking again the
425 cancer/death example, individuals might be at a low risk of cancer only because they are at
426 high risk of dying. Another illustrative example (from the CEREDIH dataset) will be given
427 later.

428 We generated two competing events, along with a censoring variable (**Figure 3**; see
429 the **Appendix** for more details). The CIF was calculated in two different ways: the correct
430 way, by considering the other event as a competing risk (as described above), and the naïve
431 approach based on a Kaplan-Meier estimator where the other event is treated as a censoring
432 variable (in the latter case, the curve is obtained by computing one minus the Kaplan-Meier
433 estimator). One can see clearly how important it is to analyze competing risks correctly: the
434 naïve approach clearly overestimates the CIFs. Again, this is because the naïve approach
435 considers individuals to be at risk after they have died, as illustrated by the fact that both
436 curves tend to 1 as time goes to infinity. In contrast, with the correct method, the sum of the
437 two probabilities tends to 1 as time goes to infinity; each individual will experience one (and
438 only one) of the two events with probability one in the future.

439 In the **Appendix**, we present some simple code that simulates a competing risk
440 situation. We chose a Weibull distribution for the true event time of interest (shape: 2; scale:
441 30) and for the competing event (shape: 2; scale: 40). This choice of parameters will result in
442 24.98% of censoring, 48% of observed events of interest and 27.01% of observed competing
443 events on average. The true curve was implemented based on calculations from the
444 Supplementary Data 1.

445 The true quantiles of order 0.1, 0.2, 0.3 for the competing event are equal to 13.69,
446 21.61 and 32.12 respectively. The estimated quantiles from the competing risk method
447 estimator are equal to 14.19, 22.28 and 35.15, respectively, while the estimated quantiles
448 from the naive estimator are equal to 13.57, 19.88 and 25.10, respectively.

TOOLBOX 4: Competing risks

- Competing risks are often present in analyses of a non-terminal event.
- The quantity of interest is usually the cumulative incidence function (CIF).
- Competing risks and right censoring are different: the event of interest cannot occur after the competing event has occurred.
- Treating competing events as right-censored observations leads to overestimation of the CIF.
- Always give the CIF for the competing risks as well as the CIF for the event of interest.
- If the competing events are not of interest, they can be grouped together as a single competing event.

449 **D. Recurrent events**

450 Recurrent events arise when the event of interest can be experienced several times
451 for each individual. In this case, a classical quantity of interest is the average number of
452 recurrent events that a patient will experience up to a given time point, which is usually
453 referred to as the cumulative mean number of recurrent events. Recurrent events occur
454 when the event of interest is (for example) cancer recurrence, an infection, or hospital
455 admission and when the objective is to estimate the average number of such events that a
456 patient will experience up to any time point. Since censoring often occurs in this type of
457 study, dedicated methods again have to be used to estimate such quantity of interest –
458 typically, by appropriately estimating the hazard rate. In particular, ignoring censoring will
459 clearly result in underestimation of the true number of recurrent events since censored
460 patients will be followed-up on a shorter time period as compared to the situation where
461 censoring did not occur. Furthermore, death is often observed as a competing event in
462 medical studies (also called a terminal event in recurrent event studies) and must be
463 accounted for; patients will not experience a recurrent event after death.

464 An estimator for the cumulative mean number of recurrent events was first
465 developed simultaneously by Nelson and Aalen in the context of right-censoring; it is
466 therefore referred to as the Nelson-Aalen estimator (20). The estimator was subsequently
467 extended by Ghosh *et al.* to the case in which a terminal event is also present (21). Ghosh *et*
468 *al.* also developed formulas for confidence intervals. It is important to stress that those
469 formulas are very general and do not make any assumptions about the dependence
470 structure of the recurrent event increments. In particular, they do not assume that recurrent
471 events have independent increments, which is very often not the case in practice. More

472 precisely, the recurrent events that may occur between any two time points are not
473 assumed to be independent. In practice, this means that having already experienced one or
474 more recurrent events may or may not influence the risk of further occurrences. This is a
475 remarkable feature of the confidence interval formula because in practice, patients that
476 have already experienced an event are often more likely to experience future occurrences
477 (22).

478 We have developed code to generate recurrent events and have also implemented
479 Ghosh *et al.*'s formulas for the estimator of the cumulative number of recurrent events in
480 the presence of a terminal event (21). For the code, we refer the reader to the ***Appendix***. In
481 order to implement the estimator, the dataset needs to be arranged in a start, stop structure
482 (also called a counting process data structure). Each patient needs to have one line one line
483 per recurrent event and one line for the censoring or terminal event time. On each lines, the
484 start time is the occurrence of the previous recurrent event and the stop time is the
485 occurrence of the next recurrent event. On the first line, the start time will be equal to the
486 time when the patient enters the risk set (generally 0) and the stop time will be the
487 censoring or terminal event time. This structure can also take into account left truncation: in
488 such a case, the truncation time will be the start time of the first line. An example of data
489 generated using this structure is given in ***Table 2***.

490 Patient #1 developed an event three times (at the age of 13.4, 23 and 24.2 years) and died at
491 the age of 26. Patient #2 did not experience any events and was censored at the age of 6.7
492 years.

493 Of note, the *survSplit* function from the *survival* library can also be used to create the
494 counting process database.

495 The expected number of recurrent events is calculated using Ghosh *et al.*'s formula
496 (21). We compared it with the naïve approach, which ignores censoring and the terminal
497 event (**Figure 4**). The naïve estimator was implemented by simply counting the number of
498 recurrent events that had occurred before a given time point, divided by the sample size.
499 While the correct estimator only includes patients at risk of experiencing a recurrent event
500 in the risk set, the naïve approach uses a fixed risk set that includes all the patients in the
501 study. Since the risk set is too large in the naïve approach, the estimator underestimates the
502 expected number of recurrent events. As recommended in the previous section, we
503 advocate to also display the CIF of the terminal event because individuals at high risk of
504 death will tend to experience fewer recurrent events. We computed it as one minus the
505 Kaplan-Meier estimator (**Figure 4**, right-hand panel). As time goes on, the CIF moves closer
506 to 1. This explains why the frequency of recurrent events appears to decrease slightly (on
507 the left-hand panel); at late time points, the competing event is more likely to have
508 occurred.

509 Recurrent events often occur in medical registry dataset with often long follow-up
510 periods and many repeated measurements of medical outcomes. Researchers are often not
511 aware of the right method for handling recurrent events, and it is customary to analyze only
512 the first event. This can lead to an important loss of information even though a recurrent
513 analysis can be straightforwardly implemented using standard libraries for survival data. In
514 order to compute the confidence intervals under the general dependence structure of the
515 recurrent event increments, we implemented Ghosh *et al.*'s formula (21). Our code is
516 available in the **Appendix** and can be applied to any recurrent event situation. At the time of
517 writing, no packages were publicly available, and so we decided to implement the formula
518 ourselves. Very recently, Klaus Holst and Thomas Scheike implemented a new function in the

519 {*mets*} R package, which computes Ghosh *et al.*'s estimator. Both approaches give the same
520 estimations.

521

TOOLBOX 5: Recurrent events

- Recurrent events occur when patients may experience the same event repeatedly over time.
- The cumulative mean number of recurrent events is an interesting summary measure of the frequency evolution of recurrent events over time.
- Competing risks often occur in recurrent event analyses.
- Ignoring right censoring will result in underestimation of the mean number of recurrent events.
- Treating competing risks as right-censored observations will lead to overestimation of recurrent events.

522 **4. Use of the CEREDIH registry: a real case**

523 PIDs are a very heterogeneous group of rare immune system diseases cause by
524 defects in 485 genes (according to the latest international classification (6)). From a medical
525 point of view, it is not usually relevant to analyze all the patients' data together; usually,
526 appropriate statistical analyses are conducted on subgroups of PID patients. Indeed, the data
527 include patients suffering from very different diseases, such as T-cell deficiencies (mainly
528 SCIDs and CIDs), B-cell deficiencies, and innate immunodeficiencies (***Figure 5***).

529 The data in the CEREDIH registry are typically right-censored because they are
530 collected in real time; as a result, most patients are alive at the time of registration. The data
531 are also retrospective in the sense that deceased patients can also be registered if they were
532 diagnosed with a PID before death. In fact, the diagnosis of PID is a requirement for
533 registration. This means that PID patients who die before being diagnosed are not included
534 in the registry. This is a typical example of left truncation when studying the patient's age at
535 death and must be taken into account appropriately in the statistical analysis.

536 When dealing with time-to-event data, one must define the starting point for the
537 follow-up of each patient. Ideally, the choice of the starting point is guided by medical
538 considerations, i.e. the time that makes the most sense for the patient. For the CEREDIH
539 registry, one possibility is to set the "start of follow-up" at the date of the clinical diagnosis.
540 In that case, the time-to-event variable will be the time elapsed since diagnosis. Another
541 possibility is to set the "start of follow-up" at the date of birth, in which case the time-to-
542 event variable of interest will be age (i.e. age when the event of interest occurred). The
543 latter option makes more sense from a medical point of view because PIDs are genetic
544 diseases; even though a patient might be diagnosed at a later age, the disease might have

545 affected him/her since birth or at least for some time before the diagnosis. As mentioned
546 above, however, the date of diagnosis might be a truncation variable. In contrast, setting the
547 start date to the date of diagnosis avoids the left truncation issue.

548 For the sake of clarity, we first show the Kaplan-Meier analysis of the time to death of
549 patients with CVID when the starting point is the date of diagnosis. Patients with CVID are
550 diagnosed at different ages, ranging from early childhood to late adulthood. Our analyses
551 were further stratified with respect to five age classes. When patients are diagnosed at a
552 later age (40+), the risk of death differs significantly (**Figure 6a**). Twenty years after
553 diagnosis, 45% of the patients diagnosed after the age of 40 were dead. For patients
554 diagnosed between 0-4, 5-9, 10-19, and 20-39 years of age, the death rate was 8%, 4%, 15%,
555 and 9%, respectively. However, the patients diagnosed earlier in life are more prone to die
556 for a reason unrelated to their PID; this highlights the limitations of this method and its
557 interpretation. The same analysis was conducted for six different PIDs but with birth as the
558 start date (**Figure 6b**). This time, the left truncation induced by the date of diagnosis was
559 taken into account by applying the above-described methodology. We can see that CVID and
560 non-CVID B-cell deficiency patients have similar survival curves and have a better prognosis
561 than patients with the other diseases. At 40 years of age, for example, 12% of the CVID
562 patients and 13% of the non-CVID patients are estimated to have died. Patients with an
563 innate immunodeficiency tend to have a higher survival than patients with a CID: the
564 probabilities of dying before 20 years of age are estimated to 25% for patients with an innate
565 immunodeficiency and to 45% for patients with a CID. A high proportion of patients with
566 SCID are estimated to die at a young age: the probability of dying in the first two years is
567 estimated 44%. For patients who have survived, the probability of death is low. In the
568 following examples, we will always use birth as the start date.

569 Not taking left-truncation into account in registered patients with CVID leads to
570 significant overestimation of the probability of survival (**Figure 7**). By using the correct
571 methodology, we estimate that 30% of patients will die before the age of 62.3 years. When
572 the naïve estimator is used, the equivalent age is 75.0. This comparison highlights the
573 potential consequences for public health at the population level.

574 We next studied the first occurrence of cancer in patients with PID. Here, the event
575 of interest is cancer, and so death is an obvious competing risk that needs to be handled
576 properly. Since we are interested in the occurrence of cancers associated with PIDs, we also
577 considered all curative therapies (HSCT, gene therapy, and thymus transplantation) as
578 competing risks. These curative therapies can be treated as a single composite event
579 recorded at the age when the patient first encounters a competing risk. We computed the
580 CIF for cancer in the six subgroups of PID patients and the CIF for the composite competing
581 risk (**Figure 8**).

582 CID patients are more likely to have experienced a first cancer before the age of 55
583 (**Figure 8a**). For the patients that are still alive at that age and have not yet experienced
584 cancer or a curative therapy, the CVID and non-CVID patients are the most at risk of
585 developing cancer. Clearly, these findings are strongly linked to those shown in **Figure 8b**.
586 Patients with SCID can undergo curative therapy or die very soon after they are born and so
587 are no longer at risk of developing cancer. In contrast, the patients with a CVID or non-CVID
588 B-cell deficiency have a much lower risk of death or curative therapy than the other patient
589 groups; this is because of their greater risk of cancer at older ages.

590 Lastly, we analyzed the recurrences of cancer, autoimmune disease episodes, and
591 inflammatory events on the CID patient group (**Figure 9**). Any of these three types of event is

592 defined as a recurrent event. We sought to estimate the mean number of such events
593 having occurred before any time point. As in the previous analysis, curative therapy and
594 death are considered to be a composite competing event. The mean number of recurrences
595 is low and increased slightly over time to a value of 0.74 before the age of 80. Again, this
596 value must be compared with the risk of experiencing death or a curative therapy over time,
597 which is high for these patients (0.98 before the age of 80).

598

599

4. DISCUSSION

600 In the present article, we discussed dedicated statistical methods for analyzing time-
601 to-event data in registries. Those types of data have the particularity that they are not
602 completely observed, and various approaches must be used to avoid biased estimations. In
603 particular, we discussed how to take into account right-censoring, left truncation, and
604 competing events. We also considered recurrent event situations, in which individuals can
605 experience the event of interest several times. If those particular mechanisms in the
606 collection of the data are not properly taken into account, the application of standard
607 methods for completely observed data will give rise to systematic biases. This fact was
608 highlighted in a simulation study in which the true mechanism that generated the data was
609 known in advance; and it was therefore possible to compare various methods with the truth.
610 Lastly, we used the methods presented here to analyze data from the CEREDIH registry.

611 In order to avoid bias in the analysis of time-to-event data, we refer to the three
612 pillars stated by Andersen and Keiding (23). “First, do not condition on future”. In other
613 words, no estimation should be carried out that uses events that will occur in the future . For
614 example, when one wants to compute an estimation on the age scale, one should check
615 whether all the patients have been followed up from birth. This is typically not the case
616 when age is the time scale, and the data will probably suffer from delayed entry (i.e. left
617 truncation). In the CEREDIH data, for instance, left truncation means that patients are
618 included in the study because it is known that they will be diagnosed at some time in the
619 future. Nevertheless, we have seen in this paper how left truncation can be taken into
620 account by modifying the risk set using the age at diagnosis. The second principle is “do not
621 regard individuals at risk after they have died”, and the third is “stick to this world.” In the

622 present article, these last two points apply to competing risks. Censoring a patient at the
623 time of his/her death implies that he/she will experience the event of interest postmortem,
624 which is impossible in the real world.

625 Left truncation is a specific example of immortal time bias. Immortal time bias occurs
626 when an individual is incorrectly considered to be at risk, i.e. during a period of time when
627 they cannot experience the event of interest (24). As an illustration of immortal time bias, let
628 us consider a cancer study in which the objective is to compare the risk of death for a group
629 of patients with cancer and a group of patients without cancer. If the time scale is age, then
630 cancer status will be a truncation variable. Furthermore, the cancer and non-cancer groups
631 are not well defined because it is not possible to know in advance (i.e. at birth) whether the
632 individual will remain cancer-free or will develop cancer at a later time. If one performs a
633 survival analysis by defining the two groups (cancer and non-cancer) in advance using the
634 Kaplan-Meier estimator, then the survival curve of the non-cancer group will be strongly
635 biased downwards; the risk set will not include any of the patients in the cancer group, even
636 though many of these patients will not have yet developed a cancer at some specific times
637 and should therefore be included in the risk set. Since the non-cancer risk set will be much
638 smaller than it should be at early time points, the corresponding survival curve will indicate
639 that the prognosis for the non-cancer group is worse than it truly is. In contrast, the fact that
640 the cancer group is defined prior to the onset of cancer introduces selection bias. As a result,
641 the comparison of the two curves will largely attenuate the effect of cancer on death. It is
642 important to note that with age as the time scale, the cancer group does not correspond to a
643 real situation. Individuals are not born with a cancer status, and cancer may or may not oc-
644 cur during the lifetime of a patient. When cancer does occur, the individual is no longer at
645 risk of developing cancer but might have an elevated risk of death. In fact, this is a multi-

646 state situation in which different events with different risks must be taken into account in
647 the survival analysis.

648 It was not possible for us to cover all the types of incomplete observation that can
649 arise in analyses of time-to-event data. Furthermore, we did not discuss how to analyze the
650 effect of covariates on a time-to-event response variable through regression modelling. In
651 the context of right-censoring and/or left truncation, this is usually performed with the Cox
652 model (16). The Cox model can also be applied to competing events and recurrent events,
653 notably via the survival package in R. Lastly, MSMs constitute a major topic of interest but
654 are not covered here (17). This situation arises when (i) an individual can experience differ-
655 ent events (referred to as states) during his/her lifetime and (ii) the risk of experiencing any
656 of these events varies. This is a natural extension of the competing risk situation; instead of
657 studying only two possible events (one of which is terminal), one looks at multiple events
658 between which transitions may or may not be allowed in the model. In the CEREDIH registry
659 data, for instance, patients may experience various events: severe infections, cancer, auto-
660 immune disease episodes, death, etc. By using MSMs, one can describe all the different
661 states associated with the disease and the changes from one state to another. This approach
662 can also be incorporated into a regression model in which covariates affect some states and
663 not others. These regression models can be implemented using the *{msm}* or *{mstate}* pack-
664 ages in R.

665 In this article, we presented various methods and highlighted a number of pitfalls in
666 the analysis of time-to-event data. As a consequence, we strongly encourage medical
667 researchers who study time-to-event data to collaborate closely with statisticians. Firstly,
668 registry data (especially rare disease registries) are essential for understanding a disease
669 (10). Secondly, funds for rare disease research are often limited, and it is therefore crucial to

670 use appropriate statistical methods and derive correct conclusions. Robust, high-quality
671 health data are critical for (i) enhancing healthcare delivery, medical R&D, and our
672 knowledge of disease, (ii) supporting policy and regulatory decisions, and (iii) ultimately
673 benefiting patients in particular and society more widely. Data can change lives by speeding
674 up diagnosis, improving patient care, and fostering the development of new treatments. In
675 rare diseases like PIDs, health data is even more vital for the provision of more effective,
676 high-quality, safe and personalized care. Worldwide, efforts are growing to strengthen the
677 collection and use of data through patient registries and the shaping of collaborative health
678 data ecosystems.

679

680 **REFERENCES**

- 681 1. Lankester AC, Neven B, Mahlaoui N, von Asmuth EGJ, Courteille V, Alligon M, et al.
682 Hematopoietic cell transplantation in severe combined immunodeficiency: The SCETIDE
683 2006-2014 European cohort. *J Allergy Clin Immunol*. mai 2022;149(5):1744-1754.e8.
- 684 2. Albert MH, Sirait T, Eikema DJ, Bakunina K, Wehr C, Suarez F, et al. Hematopoietic
685 stem cell transplantation for adolescents and adults with inborn errors of immunity: an
686 EBMT IEWP study. *Blood*. 6 oct 2022;140(14):1635-49.
- 687 3. Hageman IC, van Rooij IALM, de Blaauw I, Trajanovska M, King SK. A systematic over-
688 view of rare disease patient registries: challenges in design, quality management, and
689 maintenance. *Orphanet J Rare Dis*. 5 mai 2023;18(1):106.
- 690 4. Comité National des Registres. Rapport d'activité 1996-1999. 2000 p. 175. (Inserm
691 édit. Paris).
- 692 5. Comité national des Registres. Comité national des Registres [Internet]. [cité 11 juill
693 2023]. Disponible sur: [https://www.santepubliquefrance.fr/comite-d-evaluation-des-](https://www.santepubliquefrance.fr/comite-d-evaluation-des-registres)
694 registres
- 695 6. Tangye SG, Al-Herz W, Bousfiha A, Cunningham-Rundles C, Franco JL, Holland SM, et
696 al. Human Inborn Errors of Immunity: 2022 Update on the Classification from the Interna-
697 tional Union of Immunological Societies Expert Committee. *J Clin Immunol*. 24 juin 2022;
- 698 7. Bousfiha A, Moundir A, Tangye SG, Picard C, Jeddane L, Al-Herz W, et al. The 2022
699 Update of IUIS Phenotypical Classification for Human Inborn Errors of Immunity. *J Clin Im-*
700 *munol* [Internet]. 6 oct 2022; Disponible sur:
701 <http://www.ncbi.nlm.nih.gov/pubmed/36198931>
- 702 8. CEREDIH: The French PID study group. The French national registry of primary immu-

- 703 nodeficiency diseases. Clin Immunol Orlando Fla. mai 2010;135(2):264-72.
- 704 9. Alligon M, Mahlaoui N, Courteille V, Costes L, Afonso V, Randrianomenjanahary P, et
705 al. An appraisal of the frequency and severity of noninfectious manifestations in primary
706 immunodeficiencies: A study of a national retrospective cohort of 1375 patients over 10
707 years. J Allergy Clin Immunol. juin 2022;149(6):2116-25.
- 708 10. Abbott JK, Gelfand EW. Registries are shaping how we think about primary immuno-
709 deficiency diseases. J Allergy Clin Immunol. juin 2022;149(6):1943-5.
- 710 11. Sylvestre MP, Huszti E, Hanley JA. Do OSCAR winners live longer than less successful
711 peers? A reanalysis of the evidence. Ann Intern Med. 5 sept 2006;145(5):361-3; discussion
712 392.
- 713 12. Lange T, Keiding N. Skin cancer as a marker of sun exposure: a case of serious immor-
714 tality bias. Int J Epidemiol. juin 2014;43(3):971.
- 715 13. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: pos-
716 sibilities and pitfalls. Int J Epidemiol. juin 2012;41(3):861-70.
- 717 14. Kaplan, E. L., & Meier, P. Nonparametric estimation from incomplete observations.
718 Journal of the American statistical association. 1958;457-81.
- 719 15. Friedrich, S., Beyersmann, J., Winterfeld, U., Schumacher, M., & Allignol, A. Nonpar-
720 ametric estimation of pregnancy outcome probabilities. The Annals of Applied Statistics.
721 2017;840-67.
- 722 16. Cox D.R. Regression Models and Life-Tables. Journal of the Royal Statistical Society:
723 Series B (Methodological). 1972;187-202.
- 724 17. Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Methods
725 Med Res. avr 2002;11(2):91-115.
- 726 18. Therneau, T. A package for survival analysis in S. R package version. 2015;

- 727 19. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis
728 should report results on all cause-specific hazards and cumulative incidence functions. *J Clin*
729 *Epidemiol.* juin 2013;66(6):648-53.
- 730 20. Nelson W. Confidence Limits for Recurrence Data—Applied to Cost or Number of
731 Product Repairs. *Technometrics.* 1 mai 1995;37(2):147-57.
- 732 21. Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. *Biometrics.*
733 juin 2000;56(2):554-62.
- 734 22. Schroder J, Bouaziz O, Agner BR, Martinussen T, Madsen PL, Li D, et al. Recurrent
735 event survival analysis predicts future risk of hospitalization in patients with paroxysmal and
736 persistent atrial fibrillation. *PloS One.* 2019;14(6):e0217983.
- 737 23. Andersen PK, Keiding N. Interpretability and importance of functionals in competing
738 risks and multistate models. *Stat Med.* 20 mai 2012;31(11-12):1074-88.
- 739 24. Hanley JA, Foster BJ. Avoiding blunders involving « immortal time ». *Int J Epidemiol.*
740 juin 2014;43(3):949-61.

741

742

743

744

745 **List of Tables**

746 Table 1: the first 10 individuals simulated

747 Table 2: the first five rows of the simulated counting process database

748

749

750 **Tables**

751 Table 1: the first 10 individuals simulated

id	T	C	Tobs	status
1	23.20864	40.058249	23.208645	1
2	47.81562	32.885624	32.885624	0
3	19.73251	6.162220	6.162220	0
4	25.09012	61.071086	25.090124	1
5	17.23864	79.453411	17.238643	1
6	12.54956	20.671471	12.549556	1
7	29.33067	78.726083	29.330674	1
8	30.65042	5.784648	5.784648	0
9	37.83960	31.833046	31.833046	0
10	41.52067	77.230945	41.520666	1

752

753

754 Table 2: the first five rows of the simulated counting process database

id	start	stop	status	terminal
1	0.00000	13.437412	1	0
1	13.43741	23.040031	1	0
1	23.04003	24.159902	1	0
1	24.15990	26.020856	0	1
2	0.00000	6.701158	0	0

755

756

757 **List of Figures**

758 Figure 1: Comparison of the estimated survival function using the naïve approach (in which
759 censored observations are removed) and the Kaplan-Meier estimator.

760 Figure 2: Comparison of the estimated survival function using the Kaplan-Meier estimator
761 that takes left truncation into account and the KM estimator that ignores left truncation.

762 Figure 3: Comparison of the estimated cumulative incidence function using the competing
763 risk method and the Kaplan-Meier estimator, which treats the competing risk as censoring.

764 Figure 4: Comparison of the estimated mean number of recurrent events using censored
765 data with competing events, and the naïve estimator that does not take into account censor-
766 ing and the competing event. Left panel: the mean number of recurrent events. Right panel:
767 the competing event (the terminal event).

768 Figure 5: Distribution of PID categories in the CEREDIH registry (n=7,753 patients) as of June
769 22nd, 2022

770 Figure 6: Survival probabilities with different starting points. (a) Post-diagnosis survival for
771 patients with CVID, by age at diagnosis. (b) Overall survival from birth, by PID category.

772

773 Figure 7: Comparison of the estimation of the survival function on the CVID population using
774 the Kaplan-Meier estimator that takes left truncation into account and the KM estimator
775 that ignores left truncation.

776

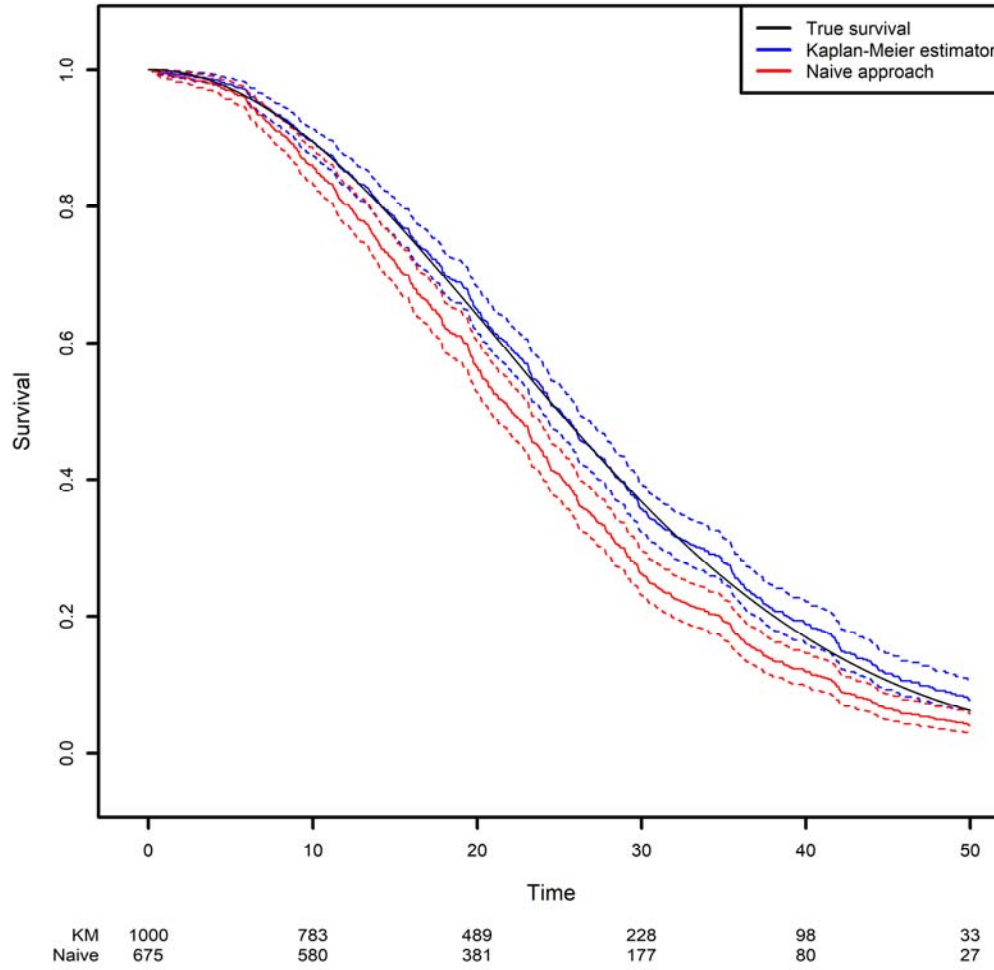
777 Figure 8: The probabilities of developing a first cancer and its competing event, according to
778 the main PID categories. (a) the probability of developing a first cancer, according to the
779 main PID categories. (b) the probability of death or receiving curative therapy (the
780 competing risk), whichever comes first and according to the main PID categories.

781

782 Figure 9: (a) the mean number of recurrent malignant, autoimmune, and inflammatory
783 events, taking death and curative therapies as competing risks. (b) the competing event
784 survival function, where the competing event is defined as death or curative therapy
785 (whichever comes first).

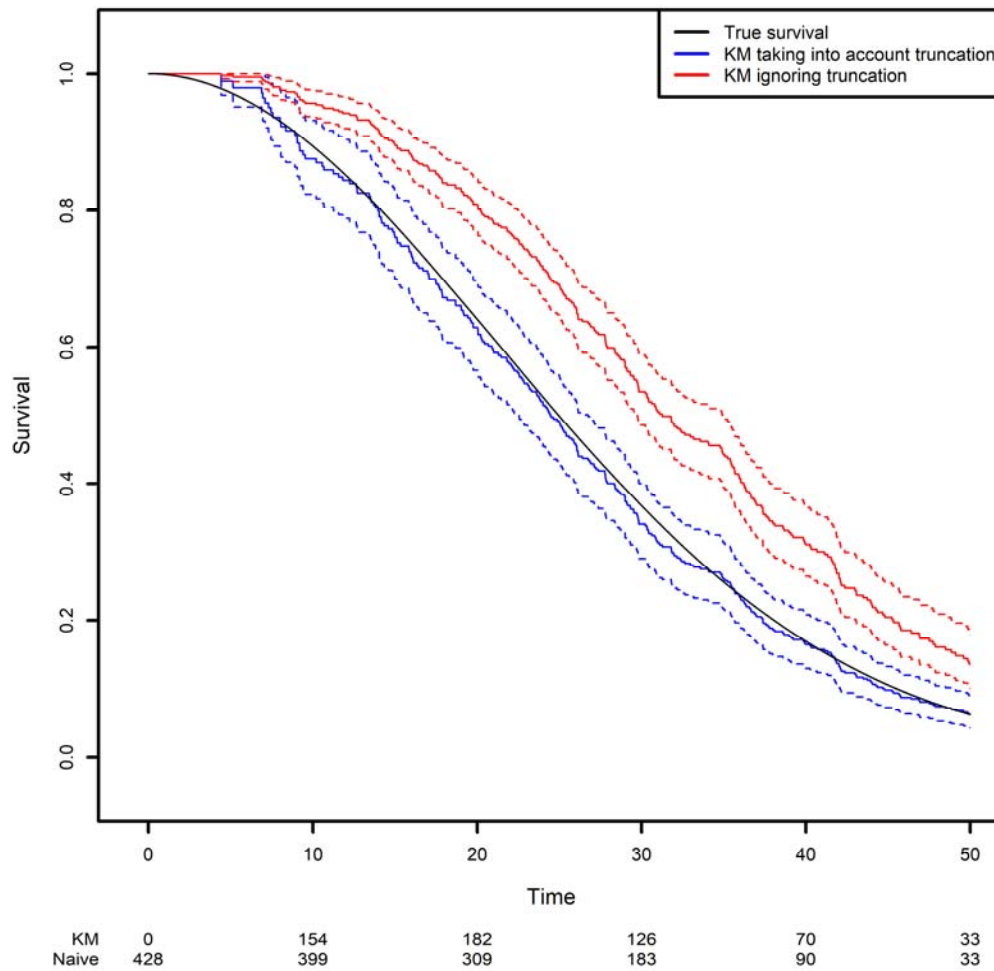
786 **Figures**

787 Figure 1: Comparison of the estimated survival function using the naïve approach (in which
788 censored observations are removed) and the Kaplan-Meier estimator.



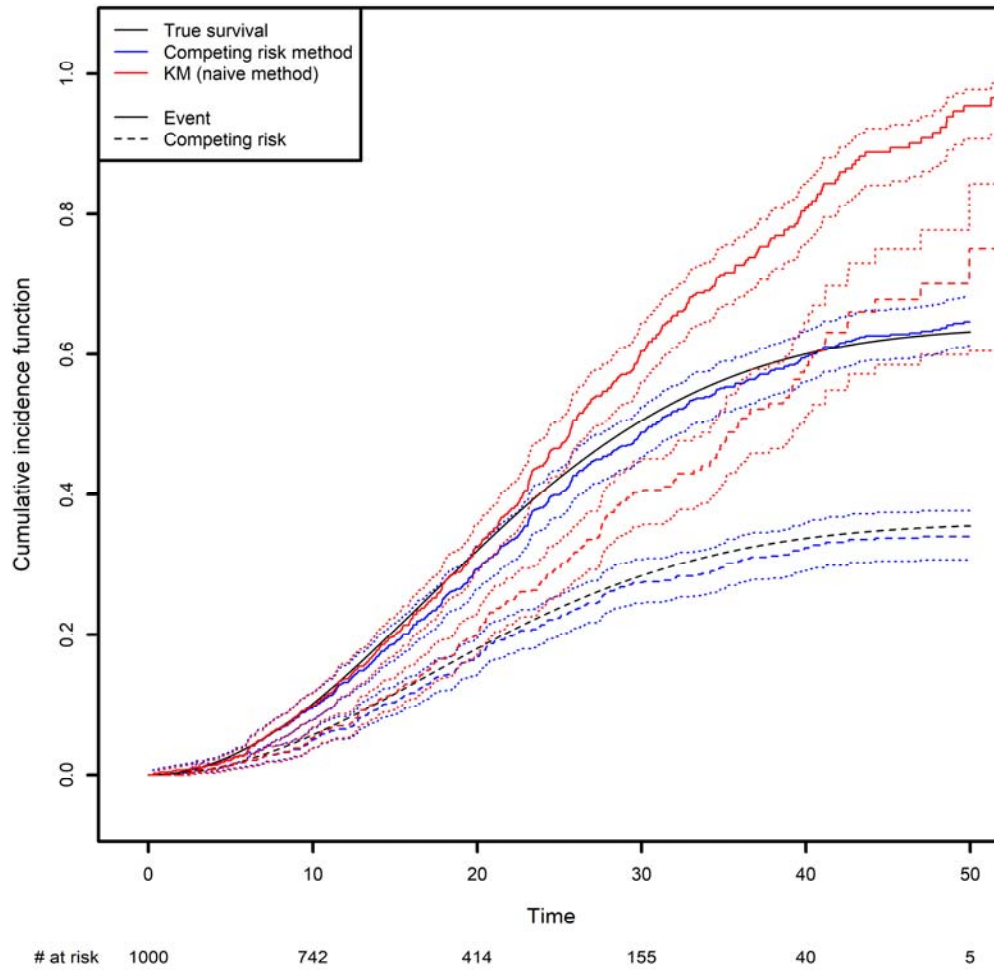
789

790 Figure 2: Comparison of the estimated survival function using the Kaplan-Meier estimator
791 that takes left truncation into account and the KM estimator that ignores left truncation.

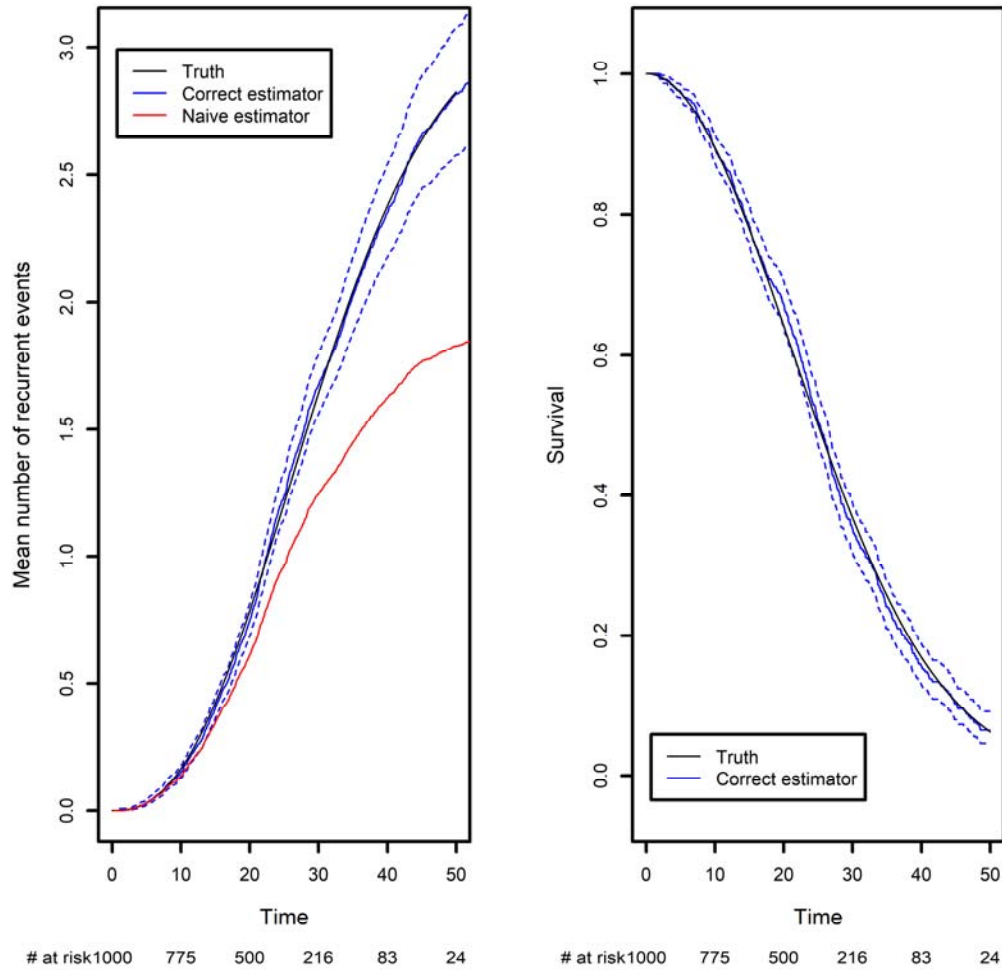


792

793 Figure 3: Comparison of the estimated cumulative incidence function using the competing
794 risk method and the Kaplan-Meier estimator, which treats the competing risk as censoring.

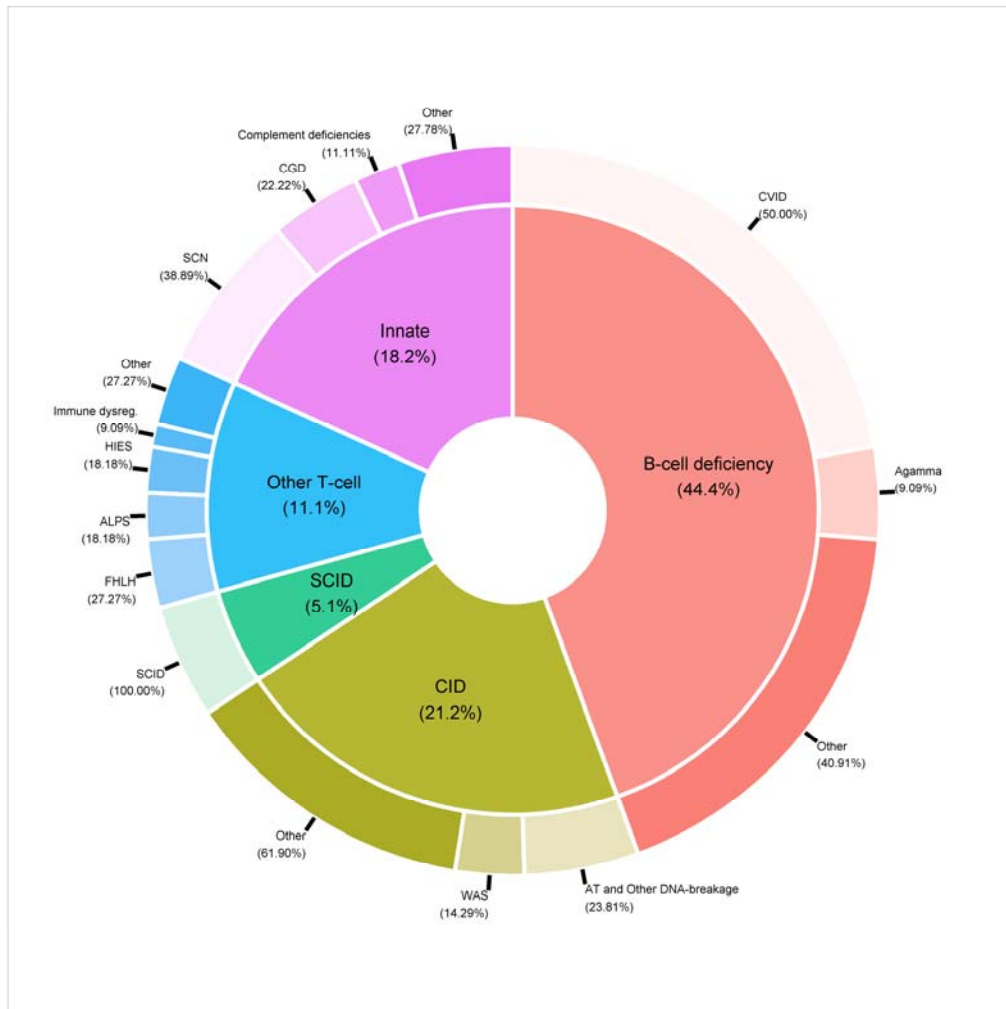


796 Figure 4: Comparison of the estimated mean number of recurrent events using censored
797 data with competing events, and the naïve estimator that does not take into account censor-
798 ing and the competing event. Left panel: the mean number of recurrent events. Right panel:
799 the competing event (the terminal event).



800

801 Figure 5: Distribution of PID categories in the CEREDIH registry (n=7,753 patients) as of June
802 22nd, 2022

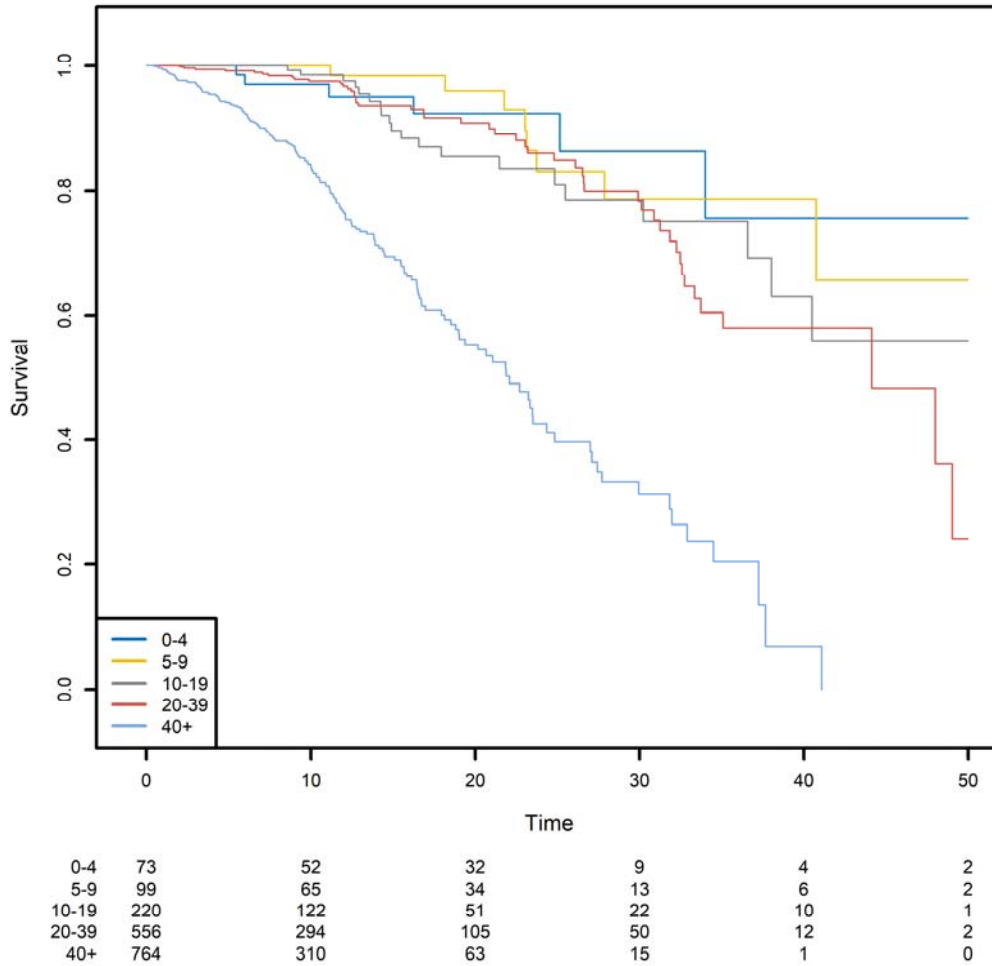


803

804 Figure 6. Survival probabilities with different starting points. (a) Post-diagnosis survival for
 805 patients with CVID, by age at diagnosis. (b) Overall survival from birth, by PID category.

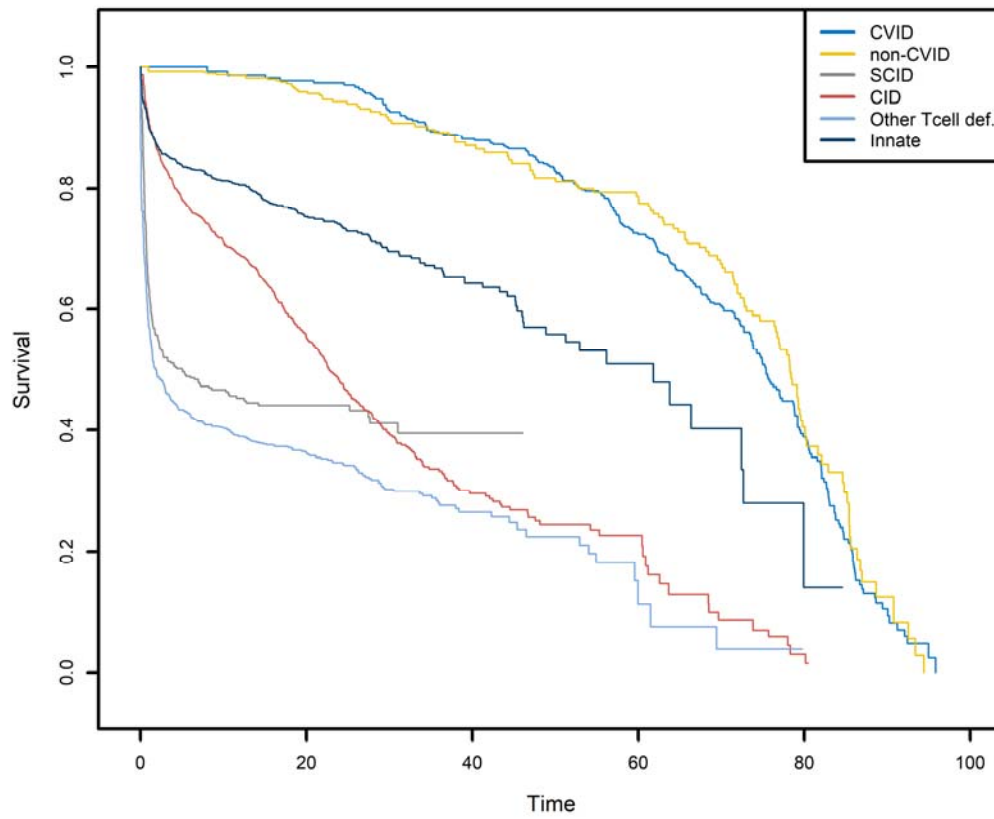
806

807 Figure 6a



808

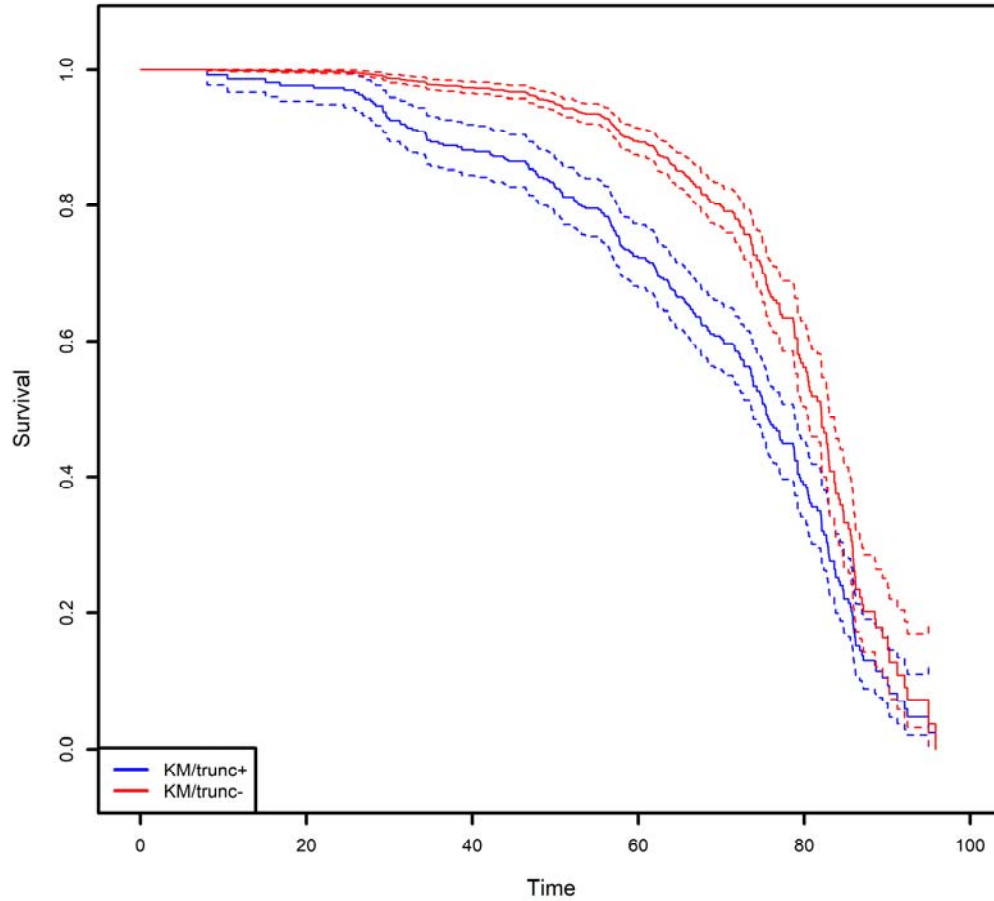
809 Figure 6b.



CVID	0	264	370	296	74	0
non-CVID	0	335	157	122	39	0
SCID	0	77	3	0	0	0
CID	0	401	64	16	2	0
OtherT	0	210	43	3	0	0
Innate	0	403	105	19	1	0

810

811 Figure 7: Comparison of the estimation of the survival function on the CVID population using
812 the Kaplan-Meier estimator that takes left truncation into account and the KM estimator
813 that ignores left truncation.



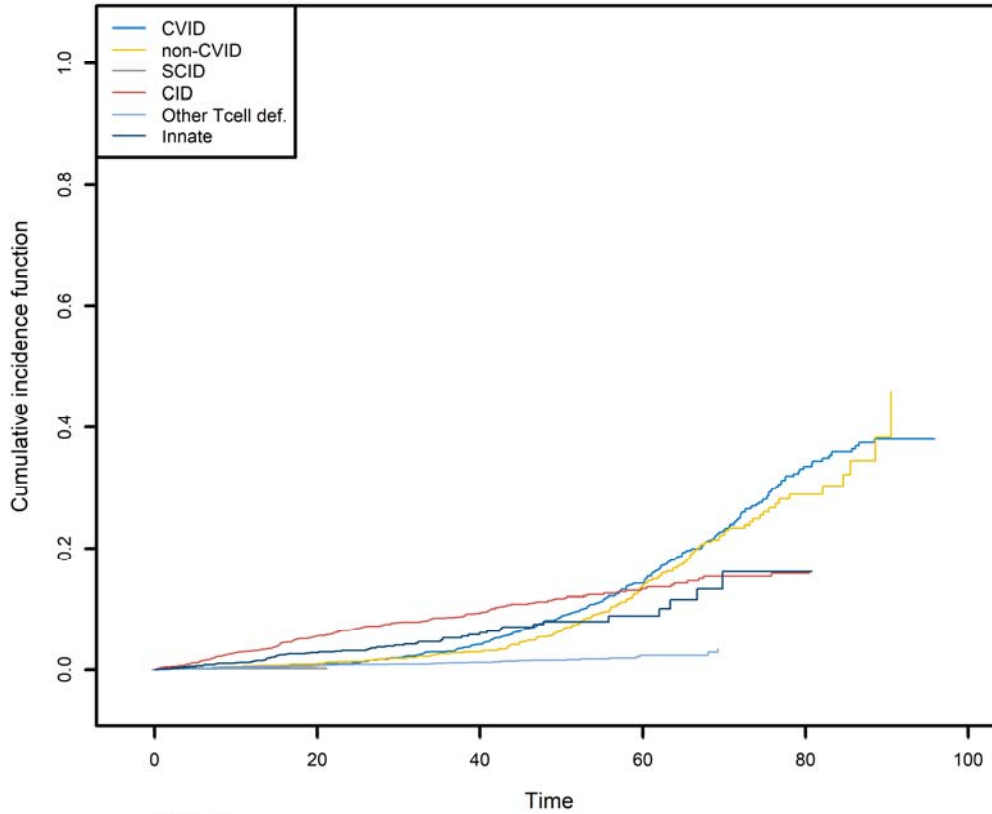
KM/trunc+	0	264	370	296	74	0
KM/trunc-	1712	1584	1134	568	86	0

814

815 Figure 8. The probabilities of developing a first cancer and its competing event, according to
 816 the main PID categories. (a) the probability of developing a first cancer, according to the
 817 main PID categories. (b) the probability of death or receiving curative therapy (the
 818 competing risk), whichever comes first and according to the main PID categories.

819

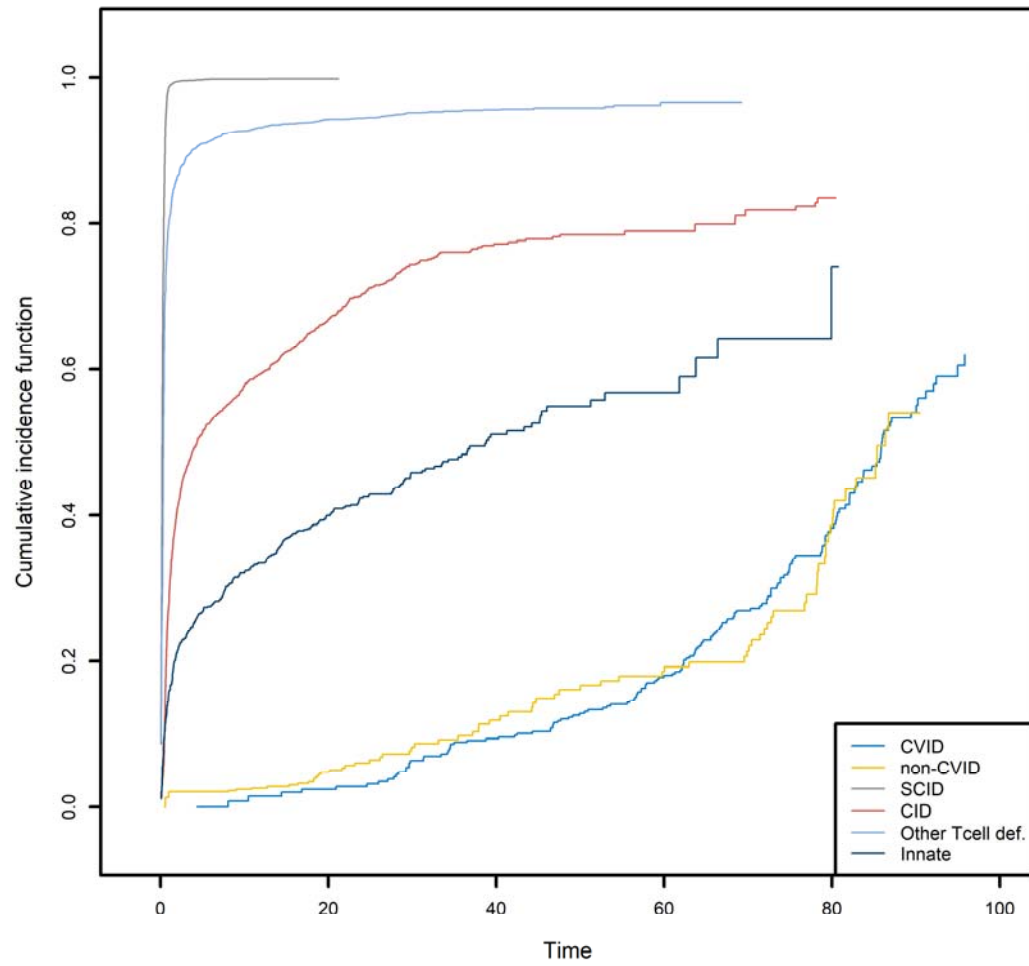
820 Figure 8a.



	Risk table					
CVID	1712	1575	1104	528	67	0
nonCVID	1762	1073	652	301	34	0
SCID	408	1	0	0	0	0
CID	1660	496	126	27	1	0
OtherT	881	258	68	4	0	0
Innate	1329	539	146	29	1	0

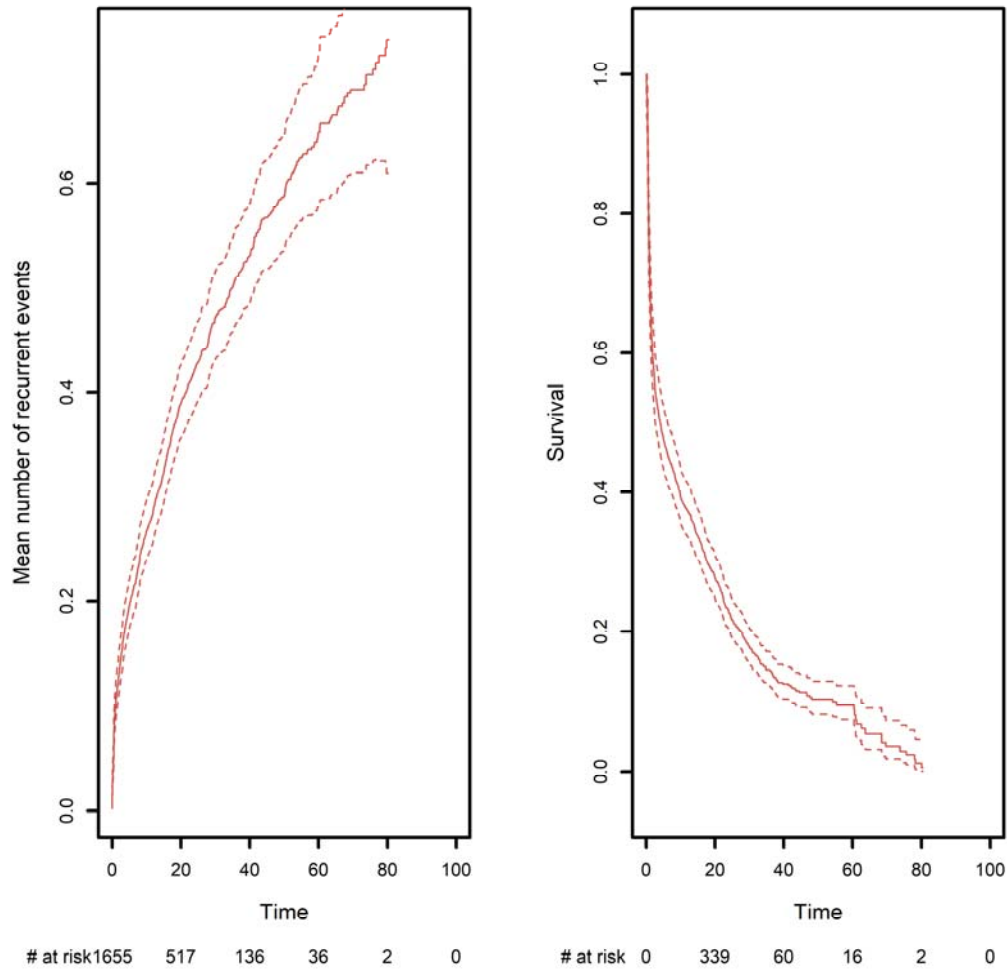
821

822 Figure 8b.



823

824 Figure 9: (a) the mean number of recurrent malignant, autoimmune, and inflammatory
825 events, taking death and curative therapies as competing risks. (b) the competing event
826 survival function, where the competing event is defined as death or curative therapy
827 (whichever comes first).



828

829 **List of Supplementary Material**

830 Supplementary data 1: True cumulative incidence function in a competing risk situation,
831 using two Weibull distributions for the simulation

832

833 **Appendix**

834 <https://github.com/Malligon/Pitfalls-in-Time-to-Event-Analysis-for-Registry-Data>

835 Supplementary data 1: True cumulative incidence function in a competing risk situation,
 836 using two Weibull distributions for the simulation

837 The two competing risks are simulated by two Weibull distributions. k_1 and k_2 are shape
 838 parameters, and λ_1 and λ_2 are scale parameters.

$$\begin{aligned}
 P(T_1 \leq t \cap T_1 \leq T_2) &= E(1_{T_1 \leq t} \& T_1 \leq T_2) \\
 &= \iint_{+\infty}^{t \wedge v} 1_{u \leq t} 1_{u \leq v} f_{T_1}(u) f_{T_2}(v) du dv \\
 &= \int_0^{+\infty} \int_0^t f_{T_1}(u) du f_{T_2}(v) dv \\
 &= \int_0^{+\infty} \left[-e^{-\left(\frac{u}{\lambda_1}\right)^{k_1}} \right]_0^t f_{T_2}(v) dv \tag{1}
 \end{aligned}$$

$$= 1 - \int_0^{+\infty} \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} e^{-\left(\frac{t \wedge v}{\lambda_1}\right)^{k_1}} dv \tag{2}$$

$$= 1 - \left(\int_0^t \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} e^{-\left(\frac{v}{\lambda_1}\right)^{k_1}} dv \right. \\
 \left. + \int_t^{+\infty} \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} e^{-\left(\frac{t}{\lambda_1}\right)^{k_1}} dv \right)$$

$$\begin{aligned}
 \int_t^{+\infty} \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} e^{-\left(\frac{t}{\lambda_1}\right)^{k_1}} dv &= e^{-\left(\frac{t}{\lambda_1}\right)^{k_1}} \left[-e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} \right]_t^{+\infty} \\
 &= e^{-\left(\frac{t}{\lambda_1}\right)^{k_1}} \left(e^{-\left(\frac{t}{\lambda_2}\right)^{k_2}} - e^{-\infty} \right) = e^{-\left(\frac{t}{\lambda_1}\right)^{k_1} - \left(\frac{t}{\lambda_2}\right)^{k_2}}
 \end{aligned}$$

839 Calculation of $\int_0^t \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} e^{-\left(\frac{v}{\lambda_1}\right)^{k_1}} dv$

$$u'e^u = \left(-\frac{k_1}{\lambda_1} \left(\frac{v}{\lambda_1}\right)^{k_1-1} - \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} \right) e^{-\left(\frac{v}{\lambda_1}\right)^{k_1} - \left(\frac{v}{\lambda_2}\right)^{k_2}}$$

$$u' = -\frac{k_1}{\lambda_1} \left(\frac{v}{\lambda_1}\right)^{k_1-1} - \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1}$$

$$u' = -\frac{k}{x\lambda_2} \left(\frac{v}{x\lambda_2}\right)^{k-1} - \frac{k}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k-1}$$

$$u' = -\frac{k}{\lambda_2} \left(\frac{1}{x} \left(\frac{v}{x\lambda_2}\right)^{k-1} + \left(\frac{v}{\lambda_2}\right)^{k-1} \right)$$

$$u' = -\frac{k}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k-1} \left(\frac{1}{x^k} + 1 \right)$$

840

$$\begin{aligned}
 \int_0^t \frac{k}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k-1} e^{-\left(\frac{v}{\lambda_2}\right)^k} e^{-\left(\frac{v}{\lambda_1}\right)^k} dv &= \int_0^t -\frac{1}{\left(\frac{1}{x^k} + 1\right)} u'e^u dv \\
 &= -\frac{1}{\left(\frac{1}{x^k} + 1\right)} \int_0^t u'e^u dv
 \end{aligned}$$

$$\begin{aligned}
 &= -\frac{1}{\left(\frac{1}{x^k} + 1\right)} \left[e^{-\left(\frac{v}{\lambda_1}\right)^k - \left(\frac{v}{\lambda_2}\right)^k} \right]_0^t \\
 &= -\frac{1}{\left(\frac{1}{x^k} + 1\right)} \left(e^{-\left(\frac{t}{\lambda_1}\right)^k - \left(\frac{t}{\lambda_2}\right)^k} - 1 \right) \\
 &= -\frac{1}{\left(\frac{\lambda_2^k}{\lambda_1^k} + 1\right)} \left(e^{-\left(\frac{t}{\lambda_1}\right)^k - \left(\frac{t}{\lambda_2}\right)^k} - 1 \right) \\
 P(T_1 \leq t \cap T_1 \leq T_2) &= 1 - \left(e^{-\left(\frac{t}{\lambda_1}\right)^k - \left(\frac{t}{\lambda_2}\right)^k} + \frac{1}{\left(\frac{\lambda_2^k}{\lambda_1^k} + 1\right)} \left(e^{-\left(\frac{t}{\lambda_1}\right)^k - \left(\frac{t}{\lambda_2}\right)^k} - 1 \right) \right)
 \end{aligned}$$

841

$$\begin{aligned}
 P(T_1 \leq t \cap T_1 \leq T_2) &= 1 - \left(\frac{1}{\left(\frac{\lambda_2^k}{\lambda_1^k} + 1\right)} + e^{-\left(\frac{t}{\lambda_1}\right)^k - \left(\frac{t}{\lambda_2}\right)^k} \left(-\frac{1}{\left(\frac{\lambda_2^k}{\lambda_1^k} + 1\right)} + 1 \right) \right) \\
 &= 1 - \left(\frac{\lambda_1^k}{\lambda_1^k + \lambda_2^k} + e^{-\left(\frac{\lambda_1^k + \lambda_2^k}{\lambda_1^k \lambda_2^k}\right) t^k} \left(1 - \frac{\lambda_1^k}{\lambda_1^k + \lambda_2^k} \right) \right)
 \end{aligned}$$

842

843 (1)

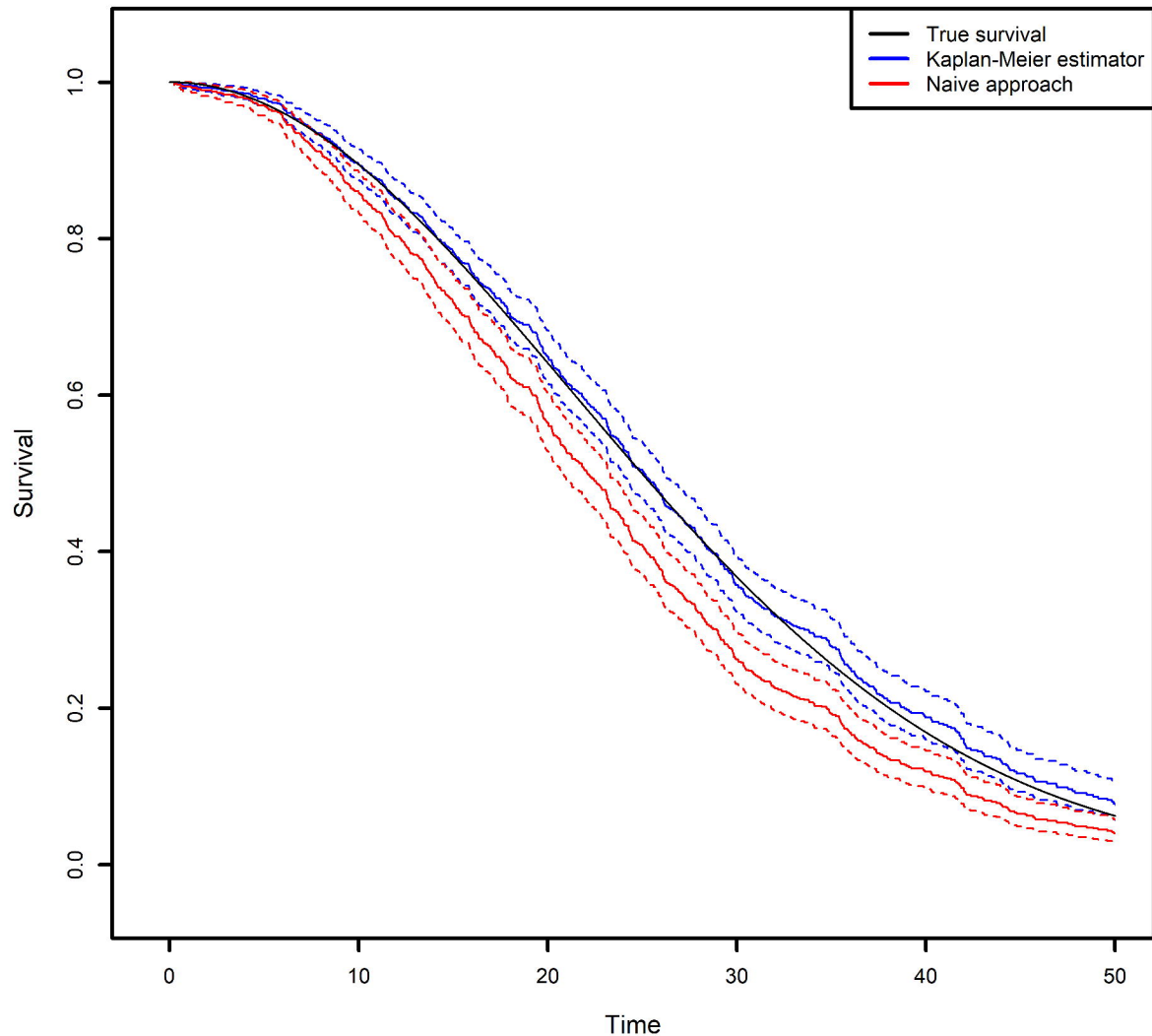
$$f_{T_i}(t) = \frac{k_i}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k_i-1} e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}} = -u' e^u \text{ with } u = -\left(\frac{t}{\lambda_i}\right)^{k_i}$$

844 (2)

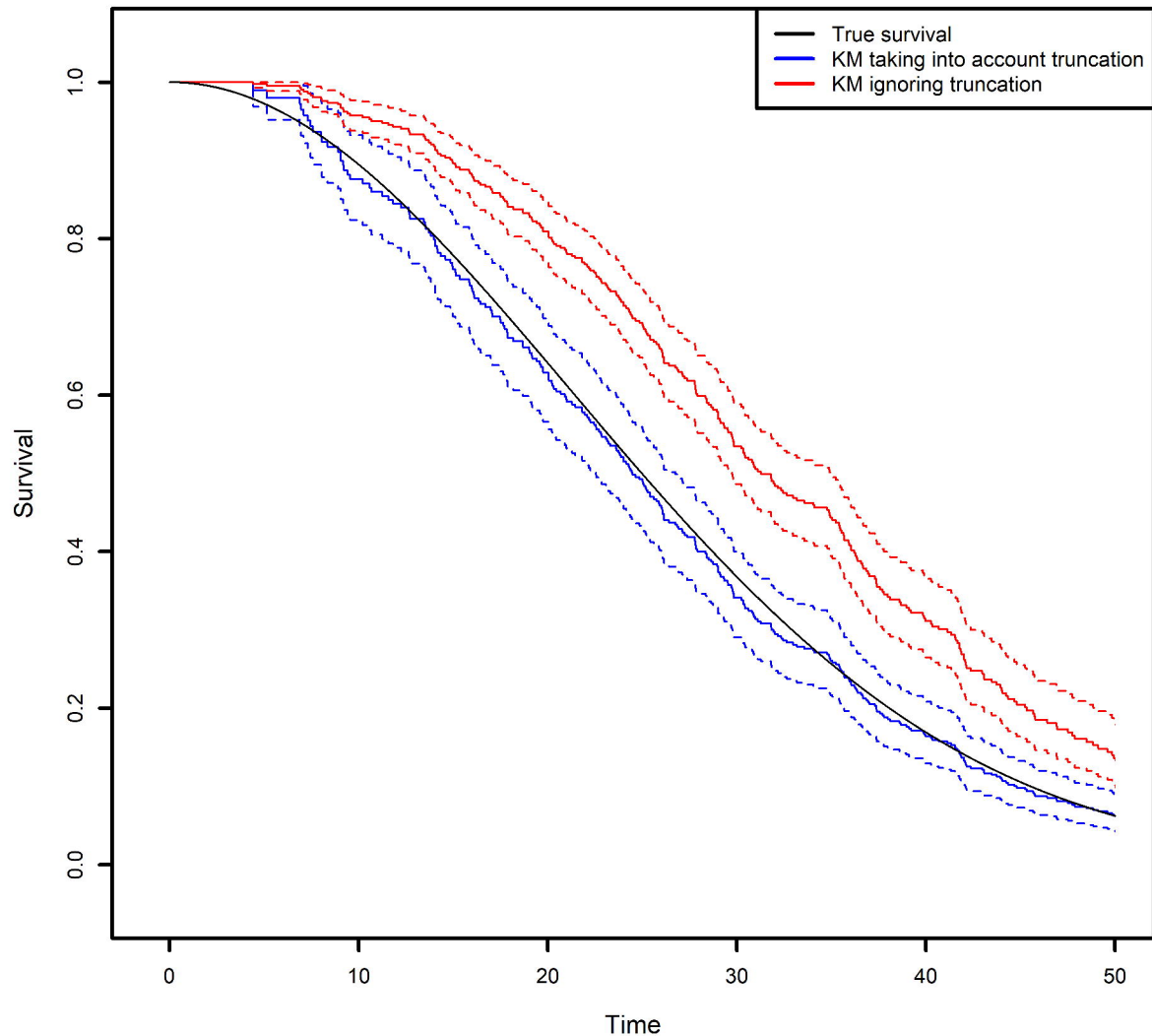
$$845 \left[-e^{-\left(\frac{u}{\lambda_1}\right)^{k_1}} \right]_0^{t^{\wedge} v} = 1 - e^{-\left(\frac{t^{\wedge} v}{\lambda_1}\right)^{k_1}}$$

$$\text{and } \int_0^{+\infty} \frac{k_2}{\lambda_2} \left(\frac{v}{\lambda_2}\right)^{k_2-1} e^{-\left(\frac{v}{\lambda_2}\right)^{k_2}} dv = 1 - e^{-\left(\frac{+\infty}{\lambda_2}\right)^{k_2}} = 1 - e^{-\infty} = 1$$

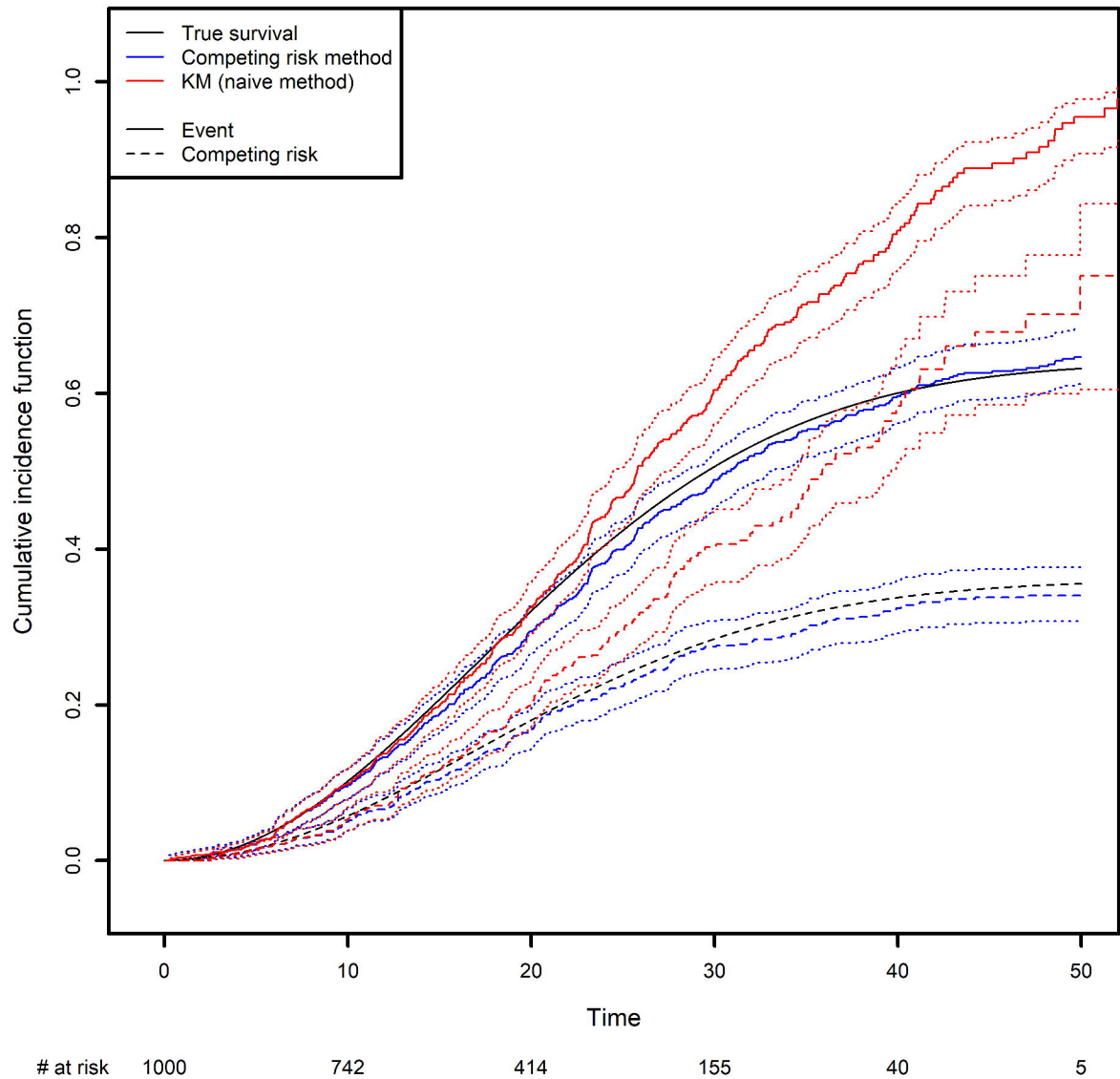
846

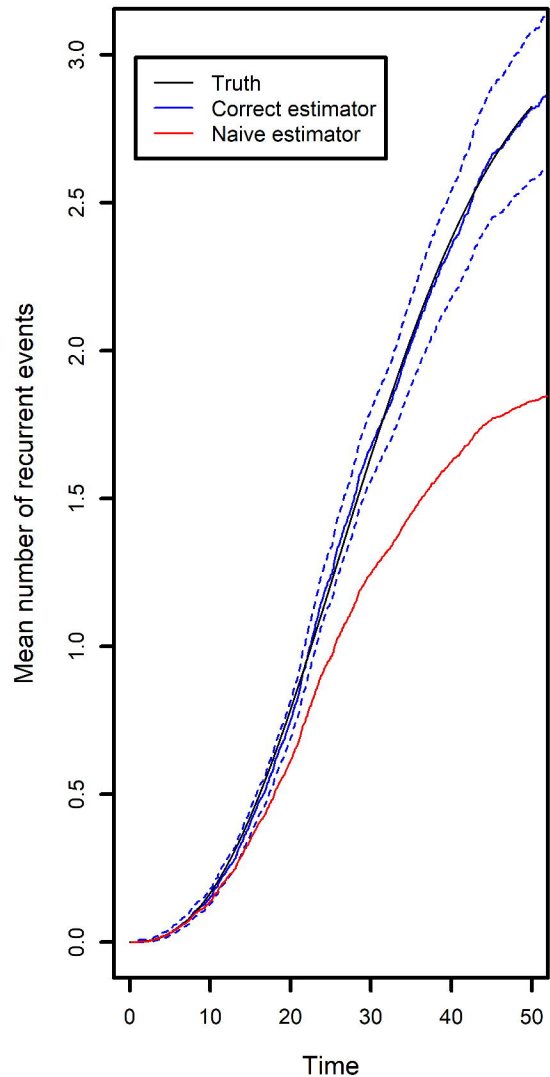


KM	1000	783	489	228	98	33
Naive	675	580	381	177	80	27

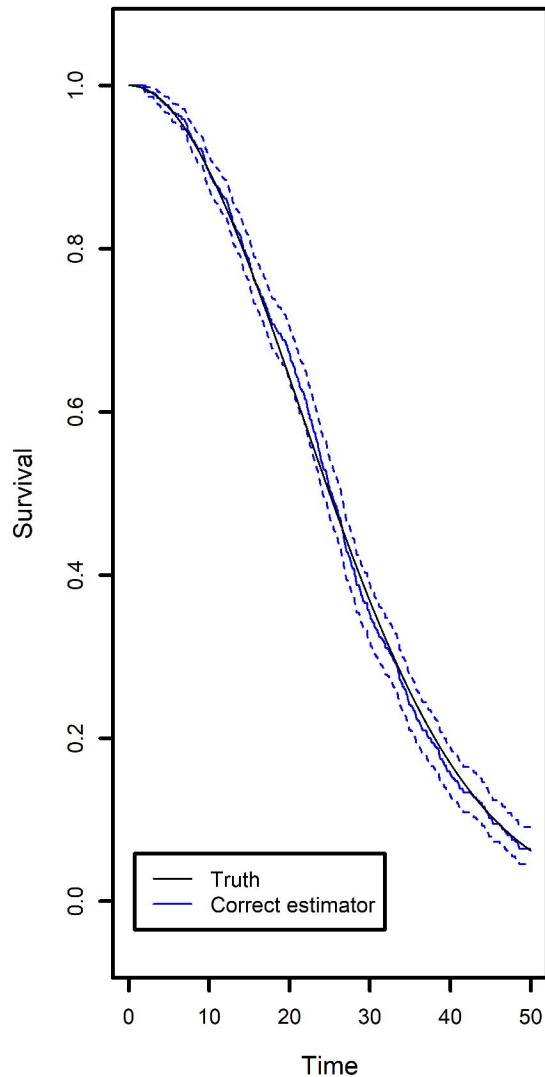


KM	0	154	182	126	70	33
Naive	428	399	309	183	90	33

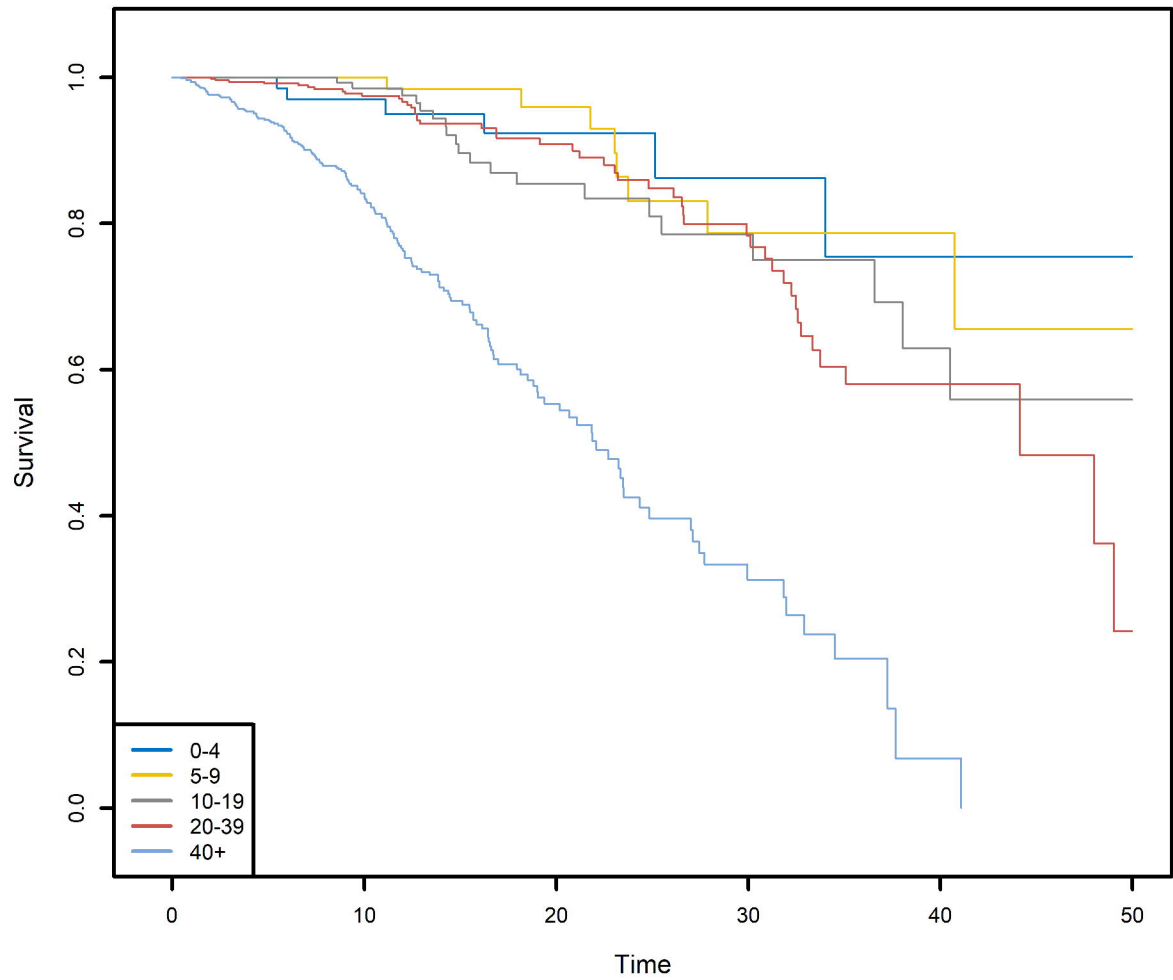




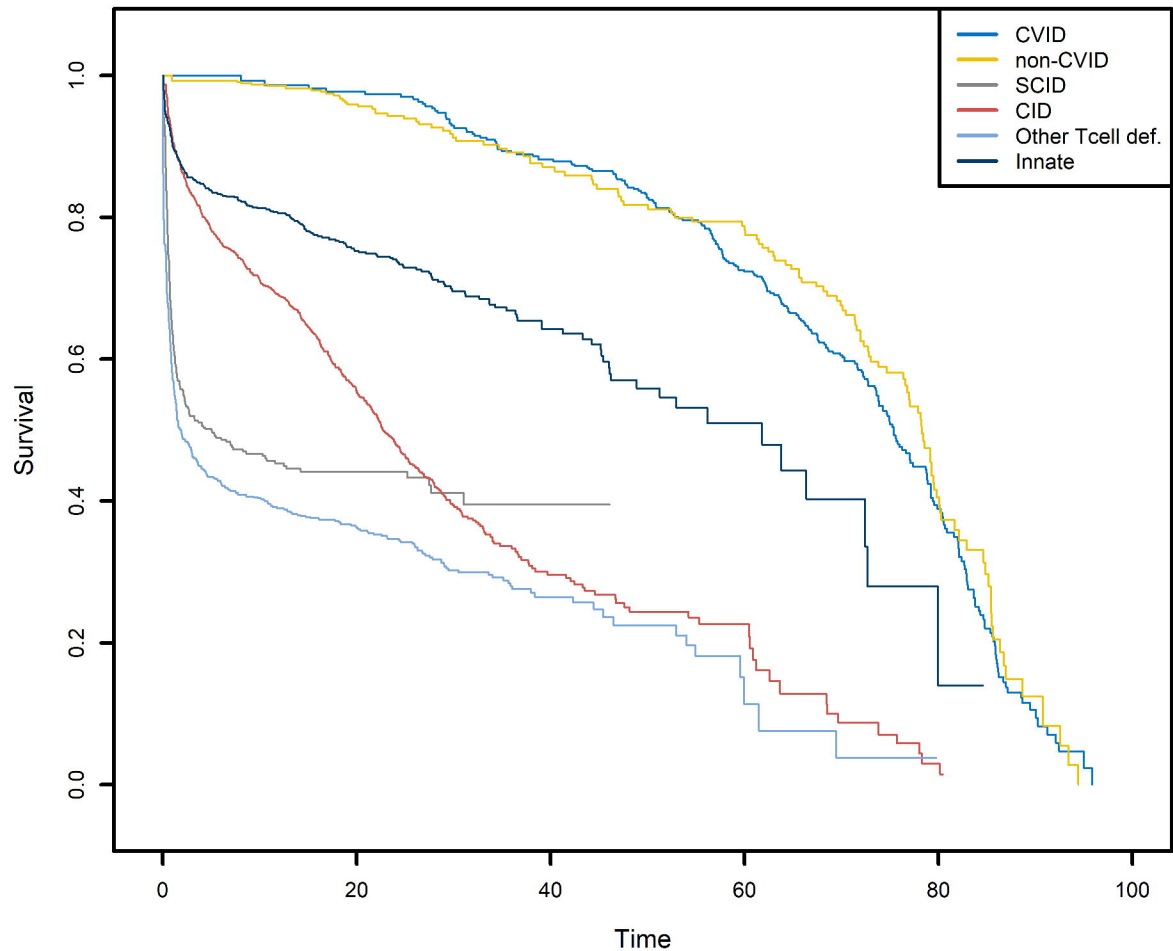
at risk 1000 775 500 216 83 24



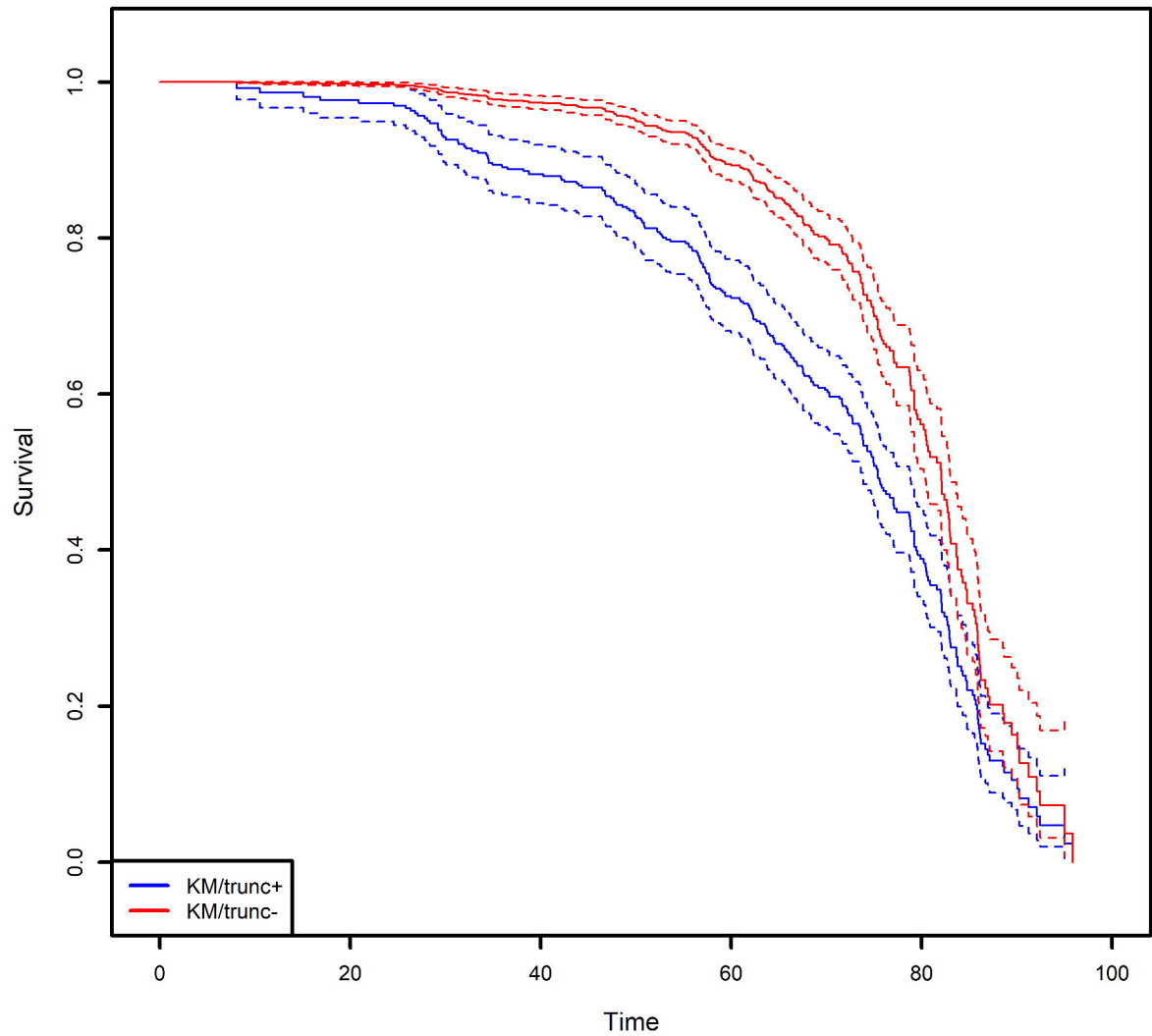
at risk 1000 775 500 216 83 24



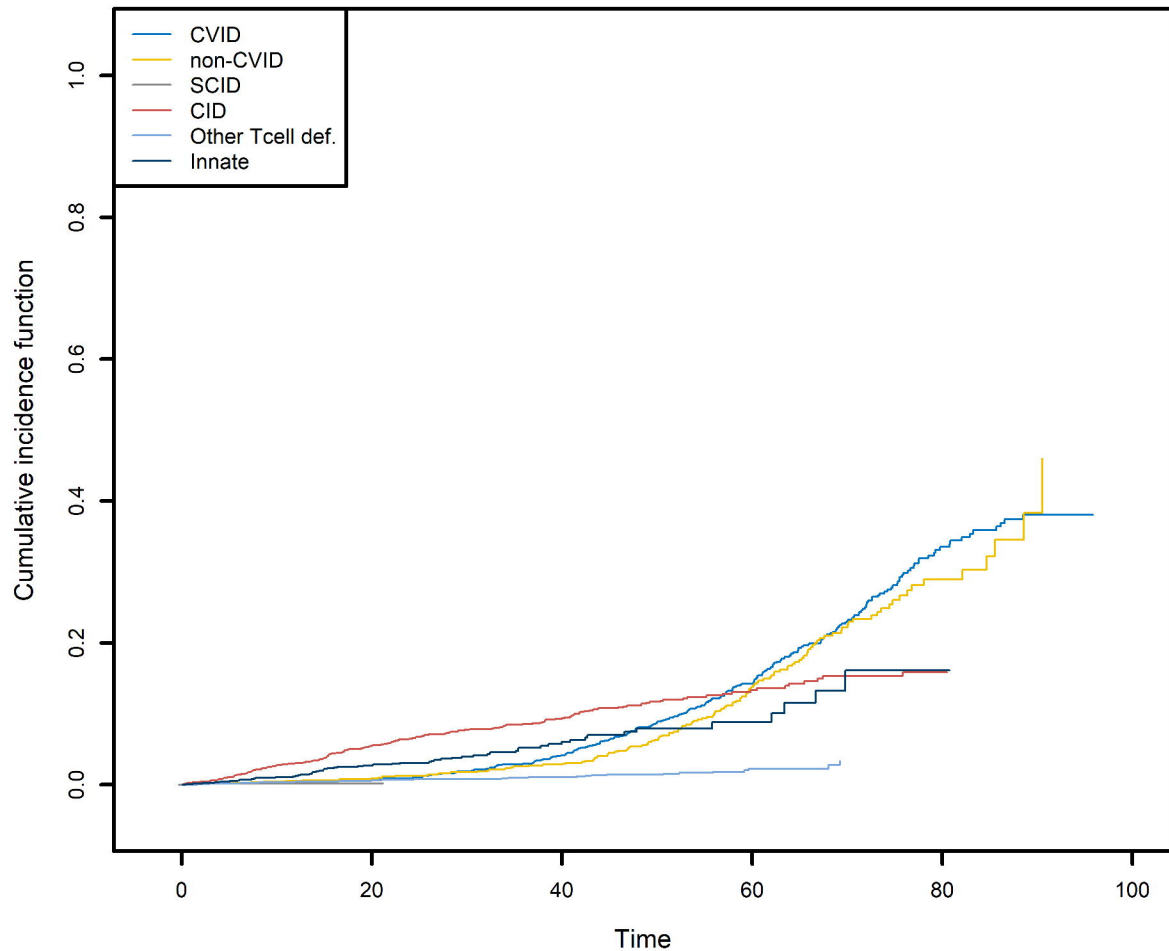
	0	10	20	30	40	50
0-4	73	52	32	9	4	2
5-9	99	65	34	13	6	2
10-19	220	122	51	22	10	1
20-39	556	294	105	50	12	2
40+	764	310	63	15	1	0



CVID	0	264	370	296	74	0
non-CVID	0	335	157	122	39	0
SCID	0	77	3	0	0	0
CID	0	401	64	16	2	0
OtherT	0	210	43	3	0	0
Innate	0	403	105	19	1	0

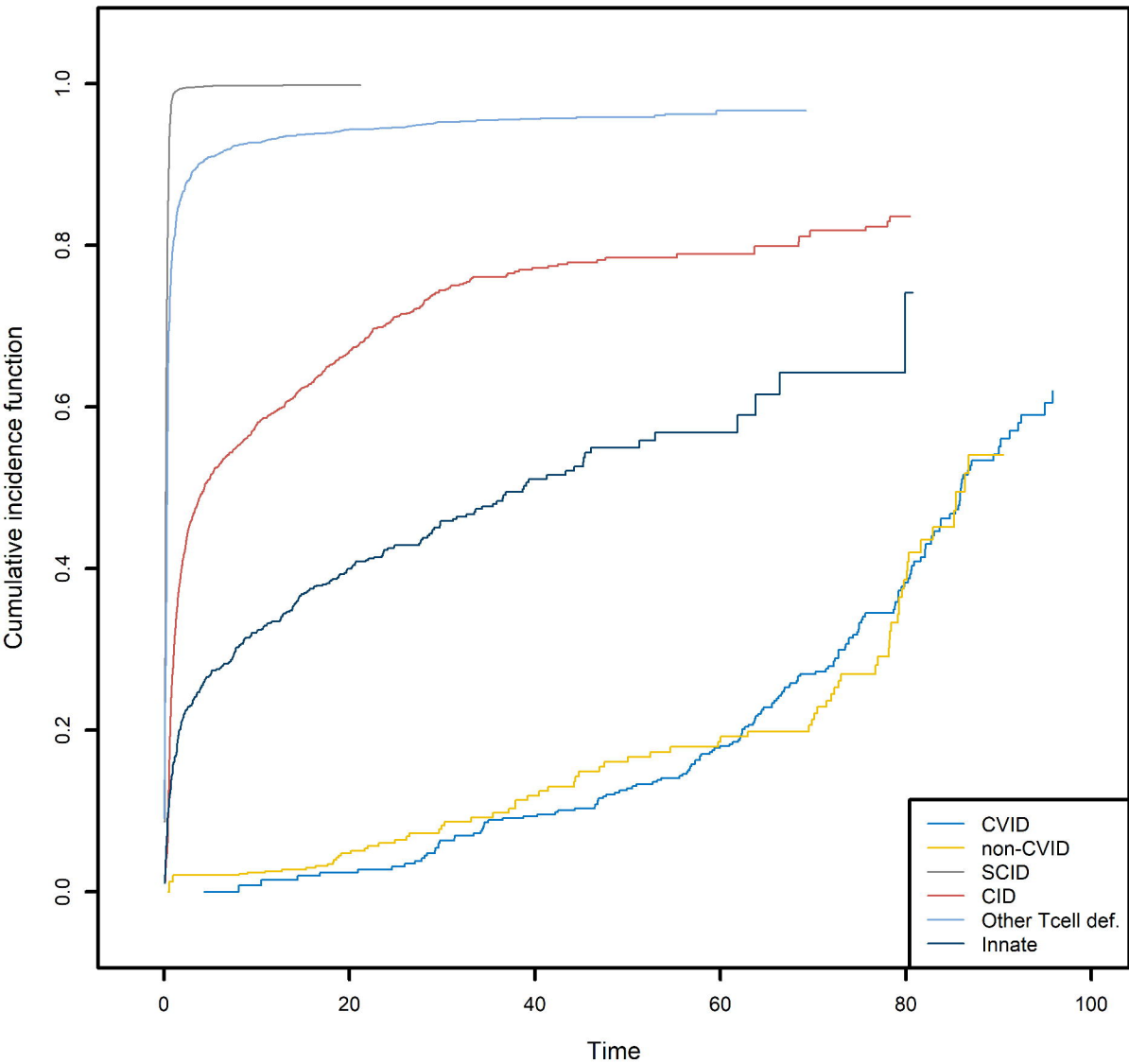


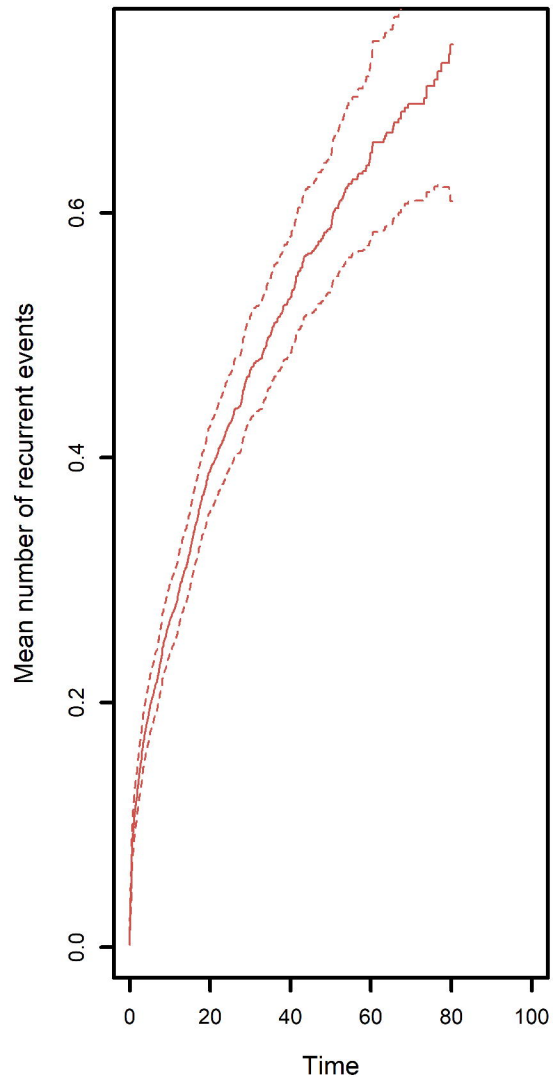
KM/trunc+	0	264	370	296	74	0
KM/trunc-	1712	1584	1134	568	86	0



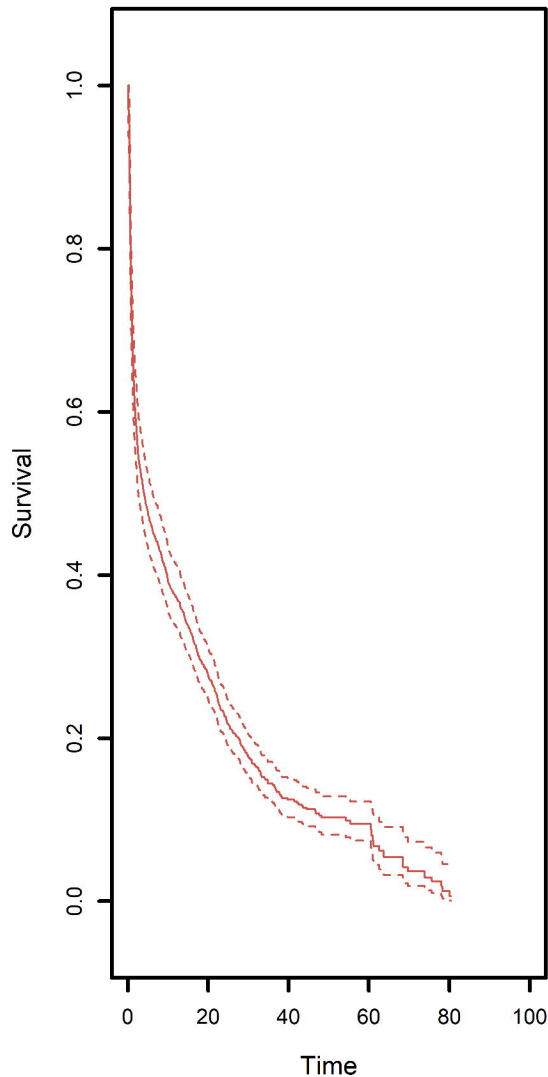
Risk table

CVID	1712	1575	1104	528	67	0
nonCVID	1762	1073	652	301	34	0
SCID	408	1	0	0	0	0
CID	1660	496	126	27	1	0
OtherT	881	258	68	4	0	0
Innate	1329	539	146	29	1	0





at risk 1655 517 136 36 2 0



at risk 0 339 60 16 2 0