The Erlang distribution approximates the age distribution of incidence of childhood and young adulthood cancers

Aleksey V Belikov [Corresp., 1] , **Alexey D Vyatkin** [1] , **Sergey V Leonov** [1]

[1] Laboratory of Innovative Medicine, School of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia

Corresponding Author: Aleksey V Belikov
Email address: belikov.research@gmail.com

Background. It is widely believed that cancers develop upon acquiring a particular number of (epi)mutations in driver genes, but the law governing the kinetics of this process is not known. We have previously shown that the age distribution of incidence for the 20 most prevalent cancers of old age is best approximated by the Erlang probability distribution. The Erlang distribution describes the probability of several successive random events occurring by the given time according to the Poisson process, which allows an estimate for the number of critical driver events.

Methods. Here we employ a computational grid search method to find global parameter optima for 5 probability distributions on the CDC WONDER dataset of the age distribution of childhood and young adulthood cancer incidence.

Results. We show that the Erlang distribution is the only classical probability distribution we found that can adequately model the age distribution of incidence for all studied childhood and young adulthood cancers, in addition to cancers of old age.

Conclusions. This suggests that the Poisson process governs driver accumulation at any age and that the Erlang distribution can be used to determine the number of driver events for any cancer type. The Poisson process implies the fundamentally random timing of driver events and their constant average rate. As waiting times for the occurrence of the required number of driver events are counted in decades, and most cells do not live this long, it suggests that driver mutations accumulate silently in the longest-living dividing cells in the body – the stem cells.

1 **The Erlang distribution approximates the age distribution**
2 **of incidence of childhood and young adulthood cancers**

3 **Aleksey V Belikov** [Corresp., 1], **Alexey D Vyatkin** [1], **Sergey V Leonov** [1]

4 [1] Laboratory of Innovative Medicine, School of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow,
5 Russia

6 Corresponding Author: Aleksey V Belikov Email address: belikov.research@gmail.com

7   *Background.* It is widely believed that cancers develop upon acquiring a particular number of
8   (epi)mutations in driver genes, but the law governing the kinetics of this process is not known. We
9   have previously shown that the age distribution of incidence for the 20 most prevalent cancers of
10 old age is best approximated by the Erlang probability distribution. The Erlang distribution
11 describes the probability of several successive random events occurring by the given time
12 according to the Poisson process, which allows an estimate for the number of critical driver events.

13 *Methods.* Here we employ a computational grid search method to find global parameter optima for
14 5 probability distributions on the CDC WONDER dataset of the age distribution of childhood and
15 young adulthood cancer incidence.

16 *Results.* We show that the Erlang distribution is the only classical probability distribution we found
17 that can adequately model the age distribution of incidence for all studied childhood and young
18 adulthood cancers, in addition to cancers of old age.

19 *Conclusions.* This suggests that the Poisson process governs driver accumulation at any age and that
20 the Erlang distribution can be used to determine the number of driver events for any cancer type.
21 The Poisson process implies the fundamentally random timing of driver events and their constant
22 average rate. As waiting times for the occurrence of the required number of driver events are
23 counted in decades, and most cells do not live this long, it suggests that driver mutations
24 accumulate silently in the longest-living dividing cells in the body – the stem cells.

INTRODUCTION

26   Since the discovery of the connection between cancer and mutations in DNA, in the

27   middle of the 20th century, there have been multiple attempts to deduce the law of driver

28   mutation accumulation from the age distribution of cancer incidence or mortality

29   (Hornsby, Page & Tomlinson, 2007). The proposed models, however, suffer from several

30   serious drawbacks. For example, early models assume that cancer mortality increases with

31   age according to the power law (Nordling, 1953; Armitage & Doll, 1954; Knudson, 2001),

32   whilst already at that time it was known that many cancers display deceleration of

33   mortality growth at advanced age. Moreover, when large-scale incidence data have

34   accumulated, it became clear that cancer incidence not only ceases to increase with age but,

35   for at least some cancers, starts to decrease (Saltzstein, Behling & Baergen, 1998; Harding,

36   Pompei & Wilson, 2012). More recent models of cancer progression are based on multiple

37   unverified biological assumptions, consist of complex equations that incorporate many

38   empirically determined parameter values, and still have not been shown to describe the

39   decrease in cancer incidence at an advanced age (Luebeck & Moolgavkar, 2002; Little &

40   Wright, 2003; Michor, Iwasa & Nowak, 2006; Meza et al., 2008; Calabrese & Shibata, 2010;

41   Luebeck et al., 2013). It is also clear that an infinite number of such mechanistic models can

42   be created and custom-tailored to fit any set of data, leading us to question their

43   explanatory and predictive values.

44   Recently, we have proposed that the age distribution of cancer incidence is, in fact, a

45   statistical distribution of probabilities for a required number of driver events to occur

46   precisely by the given age, i.e. a probability density function (Belikov, 2017). We define

47   driver events as discrete molecular alterations that provide the cancer cell of origin and the

48 resulting clone the necessary means to overcome both the intracellular and extracellular

49 tumor suppressor mechanisms, such as cell cycle checkpoints, contact inhibition, immune

50 surveillance, and promote growth in the absence of growth factors. Thus, our concept of a

51 driver event integrates mutational, microenvironmental, immune and evolutionary

52 perspectives (Rozhok & DeGregori, 2015; Solary & Lapane, 2020). It is also important to

53 note that only the winner clone's evolutionary history determines the time of tumour

54 appearance and thus the kinetics of cancer incidence, therefore any mention of driver

55 events in this article would always refer to driver events in the winner clone. The total

56 amount of competing clones would rather affect the overall probability of having a cancer.

57   Previously. we tested the probability density functions of 16 well-known continuous

58 probability distributions for fits with the CDC WONDER data on the age distribution of

59 incidence for the 20 most prevalent cancers of old age (Belikov, 2017). The best fits were

60 observed for the gamma distribution and its special case – the Erlang distribution, with the

61 average R^2 of 0.995 (Belikov, 2017). Notably, these two distributions describe the

62 probability of several successive independent random events occurring precisely by the

63 given time. In fact, the Erlang distribution is the distribution of a sum

64 of independent exponential variables, whereas the exponential distribution is the

65 probability distribution of the time between events in a Poisson point process, i.e., a

66 process in which events occur continuously and independently at a constant average rate

67 (Birnbaum, 1954). This allowed us to estimate the number of driver events, the average

68 time interval between them and the maximal populational susceptibility, for each cancer

69 type. The results showed high heterogeneity in all three parameters amongst the cancer

70 types but high reproducibility between the years of observation (Belikov, 2017).

71   However, four other probability distributions – the extreme value (Gumbel), normal,

72   logistic and Weibull – also showed good fits to the data, although inferior to the gamma and

73   Erlang distributions. This leaves some uncertainty regarding the exceptionality of the

74   gamma/Erlang distribution for the description of cancer incidence. Here we test these

75   shortlisted distributions on the CDC WONDER data on childhood and young adulthood

76   cancers, using a grid search computational method to find global parameter optima. We

77   show that the gamma/Erlang distribution is the only distribution that provides close fits

78   for all tested cancers without involvement of negative age values. This result, taken

79   together with our previous findings (Belikov, 2017), suggests that driver accumulation is

80   governed by the Poisson process, both in childhood and in adult cancers. This is consistent

81   with driver mutations that accumulate randomly, silently, at constant average rate and for

82   many decades, likely in stem cells.

83   **METHODS**

84   **Data acquisition**

85   Data were collected as previously described in (Belikov, 2017). Briefly, United States

86   Cancer Statistics Public Information Data: Incidence 1999 - 2012 were downloaded via CDC

87   WONDER online database (http://wonder.cdc.gov/cancer-v2012.HTML).   Results were

88   grouped by 5-year Age Groups, Crude Rates were selected as an output and All Ages were

89   selected in the Age Group box. The data were downloaded separately for each specific

90   cancer type, upon its selection in the Childhood Cancers tab. Only cancers that show

91   childhood/young adulthood incidence peaks and do not show middle/old age incidence

92   peaks were analysed further. The middle age of each age group was used as the x value, e.g.

93   17.5 for the "15-19 years" age group.

94

95   **Data analysis**

96   For data analysis, custom scripts were written in Python 3.7, now publicly available at

97   our GitHub repository: https://github.com/belikov-av/childhoodcancers The following

98   packages were used: Scikit-learn 0.22.1, Numpy 1.18.1, Scipy 1.4.1, Pandas 1.0.3, Plotly

99   4.5.4 and Matplotlib 3.1.3. Distributions were defined as shown in Table 1.

100   In order to find and visually demonstrate the globally optimal parameters for every

101   distribution, a grid search method (Jiménez, Lázaro & Dorronsoro, 2007) was chosen, as

102   opposed to standard maximum likelihood estimation methods that do not guarantee

103   finding the global optimum. Grid search allowed us to explore all combinations of values for

104   two parameters (k and b for the Erlang and Weibull distributions; μ and b for the extreme

105   value, logistic and normal distributions) within the defined limits and with the defined

106   steps (400 values for each parameter, 160000 grid nodes). For each pair of parameter

107   values, the optimum value for the third parameter (amplitude, interpreted as the maximal

108   populational susceptibility, see Belikov, 2017) was found using the golden section search

109   method (Pejic & Arsic, 2019). In order to perform the golden section search, first the area

110   under the incidence curve was estimated as the sum of actual incidence values for each age

111   group. The left and right boundaries for the golden section search were then defined as an

112   estimated area under the curve divided or multiplied by 100, respectively. After the

113   optimal amplitude parameter value was found with the golden section search, the fitted

114   distribution was generated and plotted alongside the incidence data and the goodness of fit

115   (R^2) was calculated, for each node of the grid. Finally, the node with the highest R^2 was

116   selected and its parameters deemed the global optimum.

117   **RESULTS**

118   To test the universality and exceptionality of the gamma/Erlang distribution, the

119   publicly available USA incidence data on childhood and young adulthood cancers were

120   downloaded from the CDC WONDER database (see Methods). In addition to the

121   gamma/Erlang distribution, the probability density functions of the following continuous

122   probability distributions were selected for testing based on their good fits to adult cancers

123   (Belikov, 2017): extreme value (Gumbel), logistic, normal and Weibull. We used a custom

124   grid search script to explore all combinations of parameters and find global R^2 maxima (see

125   Figure 1 for gamma/Erlang distribution, Supplemental Material for other distributions and

126   Methods for the detailed description). Results showed that the extreme value (Gumbel),

127   logistic and normal distributions fit to some childhood cancers in a way that a large part of

128   their probability density function appears in the negative range of x values (Figure 2). As

129   the x axis corresponds to patients' ages, this makes these three distributions

130   uninterpretable in the biological context. Hence, they were excluded from the further

131   analysis. The remaining gamma/Erlang and Weibull distributions are defined only for

132   positive x values, so there was no such problem (Figure 2). The gamma/Erlang distribution

133   provided closer fits to the childhood and young adulthood cancers than the Weibull

134   distribution (Figure 3, Table 2).

135   Importantly, the gamma distribution and the Erlang distribution derived from it are

136   the only classical continuous probability distributions that describe the cumulative waiting

137   time for k successive random events, with the Erlang distribution differing only in counting

138   events as integer numbers. Assuming this model describes the waiting time for real

139   discrete random events such as driver mutations, the gamma/Erlang distribution provides

140   the opportunity to get unique insights into the carcinogenesis process. We have previously

141   proposed that the shape parameter k of the gamma/Erlang distribution indicates the

142   average number of driver events that need to occur in order for a cancer to develop to a

143   stage that can be detected during clinical screening; the scale parameter b indicates the

144   average time interval (in years) between such events; and the amplitude parameter A

145   divided by 1000 estimates the maximal susceptibility (in percent) of a given population to a

146   given   type   of   cancer   (Belikov,   2017).

147   To obtain these parameter values, the gamma/Erlang distribution was fitted

148   individually to incidence of each of 10 childhood/young adulthood cancer types (Figure 3,

149   Table 3).

150   The non-integer values of the shape parameter k can be easily explained if we suppose

151   that the studied population consists of the mixture of patients with slightly different

152   numbers of driver events. The goodness of fit varied from 0.9495, for extracranial and

153   extragonadal germ cell tumours of young adulthood, to 1.000, for extracranial and

154   extragonadal germ cell tumours of childhood and retinoblastoma, with the average of

155   0.9854. The predicted number of driver events varied from 0.4, for extracranial and

156   extragonadal germ cell tumours of childhood, to 8.95, for malignant gonadal germ cell

157   tumours. The predicted average time between the events varied from 1.4 years, for

158   hepatoblastoma, to 14.85 years, for intracranial and intraspinal embryonal tumours. The

159   predicted maximal populational susceptibility varied from 0.0322%, for extracranial and

160   extragonadal germ cell tumours of childhood, to 1.966%, for malignant gonadal germ cell

161   tumours. Overall, the data show the same high heterogeneity in carcinogenesis patterns

162   revealed in the previous study (Belikov, 2017).

163   **DISCUSSION**

164

165   We have previously shown that five probability distributions – the extreme value

166   (Gumbel), gamma/Erlang, normal, logistic and Weibull – approximate the age distribution

167   of incidence for the 20 most prevalent cancers of old age (Belikov, 2017). The shape of

168   those incidence distributions resembles the bell shape of the normal distribution, with

169   some asymmetry, or at least the left part of it. However, many cancers of childhood have a

170   radically different shape of the incidence distribution, resembling the shape of the

171   exponential distribution (Figure 3). Of the five shortlisted distributions, only the

172   gamma/Erlang and Weibull distributions can assume such shape, when the parameter k

173   equals one or less. Other distributions can fit the data only by extending their probability

174   density function into the negative range of age values (Figure 2), which makes their proper

175   biological interpretation highly unlikely. Of the remaining two distributions, the

176   gamma/Erlang provides superior fit compared to Weibull. In fact, for cancers of old age, the

177   average R^2 for the Weibull distribution is 0.9938, whereas for the gamma/Erlang

178   distribution is 0.9954 (Belikov, 2017). For cancers of childhood and young adulthood, the

179   average R^2 for the Weibull distribution is 0.9707, whereas for the gamma/Erlang

180   distribution is 0.9854 (Table 2). Thus, it appears that the gamma/Erlang distribution is the

181   only classical probability distribution that fits universally well to cancers of childhood,

182   young adulthood and old age. The Weibull distribution can also approximate cancer

183   incidence (Webster, 2019), but would only be exactly correct for cancers with k =1, when it

184   becomes the exponential distribution.

185   We have proposed that the parameter k of the Erlang distribution indicates the

186   average number of driver events that need to occur in order for a cancer to develop to a

187   stage that can be detected during clinical screening (Belikov, 2017).   The Erlang

188   distribution, like the Weibull distribution, reduces to the exponential distribution when k

189   equals one, because the exponential distribution describes the waiting time for a single

190   random event. It would thus mean that cancers with the exponential shape of the age

191   distribution of incidence require only a single driver event with random time of

192   occurrence, most likely a somatic driver mutation (Pon & Marra, 2015) or epimutation

193   (Roy, Walsh & Chan, 2014). This explains their maximal prevalence in the early childhood.

194   In his seminal paper (Knudson, 1971), Alfred Knudson has proposed that hereditary

195   retinoblastoma is caused by a single somatic mutation in addition to one heritable

196   mutation. He also proposed that in the nonhereditary form of the disease, both mutations

197   should occur in somatic cells. As hereditary form is estimated to represent about 45% of all

198   cases (Knudson, 1971; Dimaras et al., 2012), the number of driver mutations predicted

199   from combined incidence data should be around 1.55. Interestingly, whilst the

200   gamma/Erlang distribution fits the incidence data excellently, with $R^2=1.0$, it predicts 1.3

201   driver events (Table 3). This yields the estimate of the hereditary form prevalence at 70%.

202   This higher value may point to the general underestimation of the hereditary component in

203   unilateral retinoblastoma, perhaps due to limitations of routine genetic screening and the

204   influence of genetic mosaicism (Chen et al., 2014). In contrast to retinoblastoma, the

205   hereditary form of neuroblastoma is estimated to comprise only 1-2% of all cases (Tolbert

206   et al., 2017), hence the exponential age distribution of incidence would mean that only one

207   somatic mutation is required. Indeed, the gamma/Erlang distribution predicts one driver

208   event (Table 3).

209   The prediction of a single driver event in cancers with the exponential age distribution

210   of incidence does not mean that only a single driver gene can be discovered in such cancer

211   types. In fact, many driver genes are redundant or even mutually exclusive, e.g. when the

212   corresponding proteins are components of the same signalling pathway (Vandin, Upfal &

213   Raphael, 2012). Thus, each tumour in such cancer types is expected to have a mutation in

214   one driver gene out of a set of several possible ones, in which all genes most likely encode

215   members of the same pathway or are responsible for the same cellular function. For

216   example, in each neuroblastoma tumour sample, a mutation was present in only one out of

217   five putative driver genes – *ALK*, *ATRX*, *PTPN11*, *MYCN* or *NRAS* (Pugh et al., 2013).

218   Another aspect to consider is that while one mutation is usually sufficient to activate an

219   oncogene, two mutations are typically required to inactivate both alleles of a tumour

220   suppressor gene. Therefore, cancers with the exponential age distribution of incidence are

221   predicted to have either a single somatic mutation in an oncogene, or a single somatic

222   mutation in a tumour suppressor gene plus an inherited mutation in the same gene. The

223   former is the case for neuroblastoma, where an amplification or an activating point

224   mutation in *ALK* is often present (Chen et al., 2008; George et al., 2008; Janoueix-Lerosey et

225   al., 2008). The latter is the case for retinoblastoma, where an inactivating mutation in one

226   allele of *RB1* is usually inherited, whereas an inactivating mutation in the other *RB1* allele

227   occurs in a somatic cell (Friend et al., 1986).

228   The number of driver events predicted by the Erlang distribution refers exclusively to

229   rate-limiting events responsible for cancer progression. For example, it was shown that

230   inactivation of both alleles of *RB1* leads first to retinoma, a benign tumour with genomic

231   instability that easily transforms to retinoblastoma upon acquiring additional mutations

232 (Dimaras et al., 2008). In this case, two mutations in *RB1* are rate-limiting, whereas

233 mutations in other genes are not, because genomic instability allows them to occur very

234 quickly. In neuroblastoma, frequent *MYCN* amplification and chromosome 17q gain are

235 found only in advanced stages of the disease (Brodeur et al., 1984; Bown et al., 1999), so

236 they are unlikely to be the initiating rate-limiting events.

237   Application of the gamma/Erlang distribution to childhood and young adulthood

238 cancers showed its exceptionality amongst other probability distributions. The fact that it

239 can successfully describe the radically different age distributions of incidence for cancers of

240 any age and any type suggests that the underlying Poisson process is the universal

241 principle governing cancer development. The Poisson process implies the fundamentally

242 random timing of driver events and their constant average rate (Belikov, 2017). Indeed, it

243 has been shown that the number of mutations in cancers (Tomasetti, Vogelstein &

244 Parmigiani, 2013) and stem cells (Blokzijl et al., 2016) from various tissues increases

245 linearly with age, i.e. with constant rates. The gamma/Erlang distribution allows to

246 estimate, by multiplying the number of driver events by the average time interval between

247 them, that an average person needs from 73 to 324 years to accumulate the required

248 number of driver alterations, depending on the cancer type (Belikov, 2017). These

249 estimates correspond to peaks in cancer incidence, because accumulating mutations faster

250 or slower than at an average rate is less likely. Notably, only the susceptible part of the

251 population will develop a given type of cancer, hence the area under the incidence curve

252 never reaches the total population size. Moreover, the finding that driver mutations

253 accumulate for decades is consistent with the silent accumulation of driver mutations in

254 stem cells before the terminal clonal expansion (Tomasetti & Vogelstein, 2015; Blokzijl et

255 al., 2016; Gerstung et al., 2020), because this is the only type of dividing cells surviving for

256 so long in the body, and DNA damage requires cellular division to be fixed in the form of

257 mutations. For childhood and young adulthood cancers, these estimates range from one to

258 35 years (see Table 3), but the mechanism is likely the same.

259   Intriguingly, the parameters k (the number of driver events) and b (the average

260 interval between them, in years) demonstrate inverse correlation in adult cancers (see

261 Table 1 in Belikov, 2017). This could be explained by the observation that all incidence

262 peaks of adult cancers are located at the age of 73 years old or greater, therefore cancer

263 types with high mutation rates (and thus low b values) would accumulate more driver

264 mutations (k) by that time than cancer types with low mutation rates (and thus high b

265 values). Although this answers the initial question it raises a new one: why cancers with

266 high mutation rates do not simply peak at earlier ages and with lower k, if such low k is

267 sufficient for low mutation rate cancers? One likely explanation is that an evolutionary

268 pressure of natural selection ensured that cancers do not routinely appear during the

269 reproductive period of life and therefore increased the number of tumour suppressor

270 mechanisms to overcome by high mutation rate cancers. Thus, common cancers can appear

271 only at the old age when evolutionary pressure is absent or minimal. Recently. it has been

272 proposed that age-related degradation of stem cell niche microenvironment is the major

273 reason for positive selection of driver mutations in stem cells and for late-life cancer

274 incidence peaks (Rozhok & DeGregori, 2015; Rozhok & De Gregori, 2019). That theory,

275 however, has yet to explain the decline in incidence after the peak.

276   Simulations suggest that the number of human (hematopoietic) stem cells increases

277 from birth until adolescence and then plateaus (Catlin et al., 2011). This should lead to the

278 gradual increase in childhood cancer incidence with age, as the target population for

279 mutations grows. On the other hand, modelling suggests that the replication rate of human

280 (hematopoietic) stem cells decreases from ~17 times in the first year of life, to ~2.5

281 times/year between the ages of 3 and 13 years, to ~0.6 times/year in adults (Sidorov et al.,

282 2009). This should lead to gradual decrease in childhood cancer incidence with age, as the

283 stem cell mutation rate heavily depends on the cell division rate. What is the net effect of

284 these two processes with opposing effects on childhood cancer incidence is not clear

285 (Rozhok & DeGregori, 2015). Our modelling with the constant driver event rate parameter

286 closely approximates the actual incidence of multiple childhood cancers, which suggests

287 that increases in stem cell pool sizes and decreases in stem cell division rates indeed

288 counterbalance each other and have negligible net effect on childhood cancer incidence.

289   Another interesting prediction of our model would be that the number of driver

290 mutations per stem cell or per tumour would have a Poisson distribution across the

291 patients' population of any given age. Finally, as the gamma/Erlang distribution allows to

292 predict the number and rate of driver events in any cancer subtype for which the data on

293 the age distribution of incidence are available, it may help to optimize the algorithms for

294 distinguishing between driver and passenger mutations (Raphael et al., 2014), leading to

295 the development of more effective targeted therapies.

296

297
298 **Acknowledgements**

299   AVB acknowledges MIPT 5-100 program support for early career researchers.

300 **References**

301 Armitage P, Doll R. 1954. The age distribution of cancer and a multi-stage theory of

302 carcinogenesis. *British Journal of Cancer* 8. DOI: 10.1038/bjc.1954.1.

303 Belikov AV. 2017. The number of key carcinogenic events can be predicted from cancer

304 incidence. *Scientific Reports* 7. DOI: 10.1038/s41598-017-12448-7.

305 Birnbaum A. 1954. Statistical Methods for Poisson Processes and Exponential

306 Populations. *Journal of the American Statistical Association* 49. DOI:

307 10.1080/01621459.1954.10483504.

308 Blokzijl F, De Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E,

309 Prins P, Nijman IJ, Martincorena I, Mokry M, Wiegerinck CL, Middendorp S, Sato T, Schwank

310 G, Nieuwenhuis EES, Verstegen MMA, Van Der Laan LJW, De Jonge J, Ijzermans JNM, Vries

311 RG, Van De Wetering M, Stratton MR, Clevers H, Cuppen E, Van Boxtel R. 2016. Tissue-

312 specific mutation accumulation in human adult stem cells during life. *Nature* 538. DOI:

313 10.1038/nature19768.

314 Bown N, Cotterill S, Łastowska M, O'Neill S, Pearson ADJ, Plantaz D, Meddeb M, Danglot

315 G, Brinkschmidt C, Christiansen H, Laureys G, Nicholson J, Bernheim A, Betts DR,

316 Vandesompele J, Van Roy N, Speleman F. 1999. Gain of Chromosome Arm 17q and Adverse

317 Outcome in Patients with Neuroblastoma. *New England Journal of Medicine* 340. DOI:

318 10.1056/nejm199906243402504.

319 Brodeur GM, Seeger RC, Schwab M, Varmus HE, Michael Bishop J. 1984. Amplification of

320 N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science*

321 224. DOI: 10.1126/science.6719137.

322 Calabrese P, Shibata D. 2010. A simple algebraic cancer equation: Calculating how

323 cancers may arise with normal mutation rates. *BMC Cancer* 10. DOI: 10.1186/1471-2407-

324 10-3.

325 Catlin SN, Busque L, Gale RE, Guttorp P, Abkowitz JL. 2011. The replication rate of

326 human hematopoietic stem cells in vivo. *Blood* 117:4460–4466. DOI: 10.1182/blood-2010-

327 08-303537.

328 Chen Z, Moran K, Richards-Yutz J, Toorens E, Gerhart D, Ganguly T, Shields CL, Ganguly

329 A. 2014. Enhanced sensitivity for detection of low-level germline mosaic RB1 mutations in

330   sporadic retinoblastoma cases using deep semiconductor sequencing. *Human Mutation* 35.

331   DOI: 10.1002/humu.22488.

332   Chen Y, Takita J, Choi YL, Kato M, Ohira M, Sanada M, Wang L, Soda M, Kikuchi A, Igarashi

333   T, Nakagawara A, Hayashi Y, Mano H, Ogawa S. 2008. Oncogenic mutations of ALK kinase in

334   neuroblastoma. *Nature* 455. DOI: 10.1038/nature07399.

335   Dimaras H, Khetan V, Halliday W, Orlic M, Prigoda NL, Piovesan B, Marrano P, Corson

336   TW, Eagle RC, Squire JA, Gallie BL. 2008. Loss of RB1 induces non-proliferative retinoma:

337   Increasing genomic instability correlates with progression to retinoblastoma. *Human*

338   *Molecular Genetics* 17. DOI: 10.1093/hmg/ddn024.

339   Dimaras H, Kimani K, Dimba EAO, Gronsdahl P, White A, Chan HSL, Gallie BL. 2012.

340   Retinoblastoma. DOI: 10.1016/S0140-6736(11)61137-9.

341   Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, Dryja TP. 1986. A

342   human DNA segment with properties of the gene that predisposes to retinoblastoma and

343   osteosarcoma. *Nature* 323. DOI: 10.1038/323643a0.

344   George RE, Sanda T, Hanna M, Fröhling S, Luther W, Zhang J, Ahn Y, Zhou W, London WB,

345   McGrady P, Xue L, Zozulya S, Gregor VE, Webb TR, Gray NS, Gilliland DG, Diller L, Greulich

346   H, Morris SW, Meyerson M, Look AT. 2008. Activating mutations in ALK provide a

347   therapeutic target in neuroblastoma. *Nature* 455. DOI: 10.1038/nature07397.

348   Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ,

349   Rubanova Y, Anur P, Yu K, Tarabichi M, Deshwar A, Wintersinger J, Kleinheinz K, Vázquez-

350   García I, Haase K, Jerman L, Sengupta S, Macintyre G, Malikic S, Donmez N, Livitz DG, Cmero

351   M, Demeulemeester J, Schumacher S, Fan Y, Yao X, Lee J, Schlesner M, Boutros PC, Bowtell

352   DD, Zhu H, Getz G, Imielinski M, Beroukhim R, Sahinalp SC, Ji Y, Peifer M, Markowetz F,

353   Mustonen V, Yuan K, Wang W, Morris QD, Yu K, Adams DJ, Campbell PJ, Cao S, Christie EL,

354   Cun Y, Dawson KJ, Drews RM, Eils R, Fittall M, Garsed DW, Ha G, Lee-Six H, Martincorena I,

355   Oesper L, Peto M, Raphael BJ, Salcedo A, Shi R, Shin SJ, Spiro O, Stein LD, Vembu S, Wheeler

356   DA, Yang TP, Spellman PT, Wedge DC, Van Loo P. 2020. The evolutionary history of 2,658

357   cancers. *Nature* 578. DOI: 10.1038/s41586-019-1907-7.

358   Harding C, Pompei F, Wilson R. 2012. Peak and decline in cancer incidence, mortality,

359   and prevalence at old ages. *Cancer* 118. DOI: 10.1002/cncr.26376.

360   Hornsby C, Page KM, Tomlinson IP. 2007. What can we learn from the population

361   incidence of cancer? Armitage and Doll revisited. *Lancet Oncology* 8. DOI: 10.1016/S1470-
362   2045(07)70343-1.

363   Janoueix-Lerosey I, Lequin D, Brugières L, Ribeiro A, De Pontual L, Combaret V, Raynal V,
364   Puisieux A, Schleiermacher G, Pierron G, Valteau-Couanet D, Frebourg T, Michon J, Lyonnet
365   S, Amiel J, Delattre O. 2008. Somatic and germline activating mutations of the ALK kinase
366   receptor in neuroblastoma. *Nature* 455. DOI: 10.1038/nature07398.

367   Jiménez ÁB, Lázaro JL, Dorronsoro JR. 2007. Finding Optimal Model Parameters by
368   Discrete Grid Search. In: Corchado E, Corchado JM, Abraham A eds. *Innovations in Hybrid*
369   *Intelligent Systems*. Advances in Soft Computing. Berlin, Heidelberg: Springer Berlin
370   Heidelberg, 120–127. DOI: 10.1007/978-3-540-74972-1_17.

371   Knudson AG. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings*
372   *of the National Academy of Sciences of the United States of America* 68. DOI:
373   10.1073/pnas.68.4.820.

374   Knudson AG. 2001. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* 1.
375   DOI: 10.1038/35101031.

376   Little MP, Wright EG. 2003. A stochastic carcinogenesis model incorporating genomic
377   instability fitted to colon cancer data. *Mathematical Biosciences* 183. DOI: 10.1016/S0025-
378   5564(03)00040-3.

379   Luebeck EG, Curtius K, Jeon J, Hazelton WD. 2013. Impact of tumor progression on
380   cancer incidence curves. *Cancer Research* 73. DOI: 10.1158/0008-5472.CAN-12-2198.

381   Luebeck EG, Moolgavkar SH. 2002. Multistage carcinogenesis and the incidence of
382   colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of*
383   *America* 99. DOI: 10.1073/pnas.222118199.

384   Meza R, Jeon J, Moolgavkar SH, Georg Luebeck E. 2008. Age-specific incidence of cancer:
385   Phases, transitions, and biological implications. *Proceedings of the National Academy of*
386   *Sciences of the United States of America* 105. DOI: 10.1073/pnas.0801151105.

387   Michor F, Iwasa Y, Nowak MA. 2006. The age incidence of chronic myeloid leukemia can
388   be explained by a one-mutation model. *Proceedings of the National Academy of Sciences of*
389   *the United States of America* 103. DOI: 10.1073/pnas.0607006103.

390   Nordling CO. 1953. A new theory on the cancer-inducing mechanism. *British Journal of*
391   *Cancer* 7. DOI: 10.1038/bjc.1953.8.

392 Pejic D, Arsic M. 2019. Minimization and Maximization of Functions: Golden-Section

393 Search in One Dimension. In: Milutinovic V, Kotlar M eds. *Exploring the DataFlow*

394 *Supercomputing Paradigm*. Computer Communications and Networks. Cham: Springer

395 International Publishing, 55–90. DOI: 10.1007/978-3-030-13803-5_3.

396 Pon JR, Marra MA. 2015. Driver and Passenger Mutations in Cancer. *Annual Review of*

397 *Pathology: Mechanisms of Disease* 10:25–50. DOI: 10.1146/annurev-pathol-012414-

398 040312.

399 Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K,

400 Hanna M, Kiezun A, Kim J, Lawrence MS, Lichenstein L, Mckenna A, Pedamallu CS, Ramos

401 AH, Shefler E, Sivachenko A, Sougnez C, Stewart C, Ally A, Birol I, Chiu R, Corbett RD, Hirst

402 M, Jackman SD, Kamoh B, Khodabakshi AH, Krzywinski M, Lo A, Moore RA, Mungall KL,

403 Qian J, Tam A, Thiessen N, Zhao Y, Cole KA, Diamond M, Diskin SJ, Mosse YP, Wood AC, Ji L,

404 Sposto R, Badgett T, London WB, Moyer Y, Gastier-Foster JM, Smith MA, Auvil JMG, Gerhard

405 DS, Hogarty MD, Jones SJM, Lander ES, Gabriel SB, Getz G, Seeger RC, Khan J, Marra MA,

406 Meyerson M, Maris JM. 2013. The genetic landscape of high-risk neuroblastoma. *Nature*

407 *Genetics* 45. DOI: 10.1038/ng.2529.

408 Raphael BJ, Dobson JR, Oesper L, Vandin F. 2014. Identifying driver mutations in

409 sequenced cancer genomes: Computational approaches to enable precision medicine.

410 *Genome Medicine* 6. DOI: 10.1186/gm524.

411 Roy DM, Walsh LA, Chan TA. 2014. Driver mutations of cancer epigenomes. *Protein and*

412 *Cell* 5. DOI: 10.1007/s13238-014-0031-6.

413 Rozhok A, De Gregori J. 2019. A generalized theory of age-dependent carcinogenesis.

414 *eLife* 8. DOI: 10.7554/eLife.39950.

415 Rozhok AI, DeGregori J. 2015. Toward an evolutionary model of cancer: Considering the

416 mechanisms that govern the fate of somatic mutations. *Proceedings of the National*

417 *Academy of Sciences* 112:8914. DOI: 10.1073/pnas.1501713112.

418 Saltzstein SL, Behling CA, Baergen RN. 1998. Features of cancer in nonagenarians and

419 centenarians. *Journal of the American Geriatrics Society* 46. DOI: 10.1111/j.1532-

420 5415.1998.tb02755.x.

421 Sidorov I, Kimura M, Yashin A, Aviv A. 2009. Leukocyte telomere dynamics and human

422 hematopoietic stem cell kinetics during somatic growth. *Experimental Hematology* 37:514–

423   524. DOI: 10.1016/j.exphem.2008.11.009.

424   Solary E, Lapane L. 2020. The role of host environment in cancer evolution. *Evolutionary*

425   *Applications* 13. DOI: 10.1111/eva.13039.

426   Tolbert VP, Coggins GE, Maris JM, Maris JM. 2017. Genetic susceptibility to

427   neuroblastoma. *Current Opinion in Genetics and Development* 42. DOI:

428   10.1016/j.gde.2017.03.008.

429   Tomasetti C, Vogelstein B. 2015. Variation in cancer risk among tissues can be explained

430   by the number of stem cell divisions. *Science* 347:78–81. DOI: 10.1126/science.1260825.

431   Tomasetti C, Vogelstein B, Parmigiani G. 2013. Half or more of the somatic mutations in

432   cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the*

433   *National Academy of Sciences of the United States of America* 110. DOI:

434   10.1073/pnas.1221068110.

435   Vandin F, Upfal E, Raphael BJ. 2012. De novo discovery of mutated driver pathways in

436   cancer. *Genome Research* 22. DOI: 10.1101/gr.120477.111.

437   Webster AJ. 2019. Multi-stage models for the failure of complex systems, cascading

438   disasters, and the onset of disease. *PLOS ONE* 14:e0216422. DOI:

439   10.1371/journal.pone.0216422.

440

Probability density functions and Python code for statistical distributions

Distribution	Probability density function	Python code
Gamma/Erlang	$$f(t) = \frac{1}{b\Gamma(k)}\left(\frac{t}{b}\right)^{k-1}e^{-\frac{t}{b}}$$	def Erlang_pdf(k, b, t): return (t**(k-1) * np.exp(-t/float(b))) / (b**k * gamma(k))
Weibull	$$f(t) = \frac{k}{b}\left(\frac{t}{b}\right)^{k-1}e^{-\left(\frac{t}{b}\right)^k}$$	def Weibull_pdf(k, b, t): return (k/b) * (t/b)**(k-1) * np.exp(-(t/b)**k)
Extreme value	$$f(t) = \frac{1}{b}e^{-\left(\frac{t-\mu}{b}\right)}e^{-e^{-\left(\frac{t-\mu}{b}\right)}}$$	def Extreme_value_pdf(mu, b, t): return np.exp((mu - t) / b) * (1/b) * np.exp(-np.exp((mu - t) / b))
Logistic	$$f(t) = \frac{e^{\left(\frac{t-\mu}{b}\right)}}{b\left(1 + e^{\left(\frac{t-\mu}{b}\right)}\right)^2}$$	def Logistic_pdf(mu, b, t): return (1/b) * np.exp((t - mu) / b) / np.square(1 + np.exp((t - mu) / b))
Normal	$$f(t) = \frac{1}{b\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{t-\mu}{b}\right)^2}$$	def Normal_pdf(mu, b, t): return (1/(b * np.sqrt(2 * pi))) * np.exp(-0.5 * np.square((t - mu)/b))

Comparison of the goodness of fit (R^2) of the gamma/Erlang and Weibull distributions to the actual age distributions of childhood/young adulthood cancer incidence.

The best fit for each cancer type is highlighted in bold. See Figure 3 for graphical representation.

Cancer type	gamma/Erlang	Weibull
Ewing tumour and related sarcomas of bone	**0.9516**	0.9476
Extracranial and extragonadal germ cell tumours of childhood	**1.000**	**1.000**
Extracranial and extragonadal germ cell tumours of young adulthood	**0.9495**	0.8689
Hepatoblastoma	**0.9996**	**0.9996**
Intracranial and intraspinal embryonal tumours	**0.9738**	**0.9738**
Intracranial and intraspinal germ cell tumours	**0.9890**	0.9674
Malignant gonadal germ cell tumours	**0.9941**	0.9530
Nephroblastoma and other nonepithelial renal tumours	**0.9970**	0.9969
Neuroblastoma and ganglioneuroblastoma	**0.9998**	**0.9998**
Retinoblastoma	**1.000**	**1.000**
Average	**0.9854**	0.9707

Estimated carcinogenesis parameters for 10 childhood/young adulthood cancer types.

The parameters are determined for the gamma/Erlang distribution fitted to actual cancer incidence data (see Figure 3).

Cancer type	k	b	A/1000
	Number of driver events	Average time between events, years	Maximal populational susceptibility, %
Ewing tumour and related sarcomas of bone	4.3	4.25	0.0846
Extracranial and extragonadal germ cell tumours of childhood	0.4	2.35	0.0322
Extracranial and extragonadal germ cell tumours of young adulthood	6.5	5.3	0.111
Hepatoblastoma	1.55	1.4	0.0338
Intracranial and intraspinal embryonal tumour	1.0	14.85	0.179
Intracranial and intraspinal germ cell tumours	5.65	3.05	0.0426
Malignant gonadal germ cell tumours	8.95	3.9	1.966
Nephroblastoma and other nonepithelial renal tumours	1.75	2.2	0.124
Neuroblastoma and ganglioneuroblastoma	1.0	2.5	0.179
Retinoblastoma	1.3	1.45	0.072

Figure 1

Goodness of fit of the gamma/Erlang distribution to the age distributions of incidence of 10 childhood/young adulthood cancer types as a function of various parameter combinations.

The **k** and **b** parameters were sampled with 0.05 interval, and for each pair the amplitude (**A**, not shown) and goodness of fit (R^2) were calculated. See Supplemental Material for the other distributions.
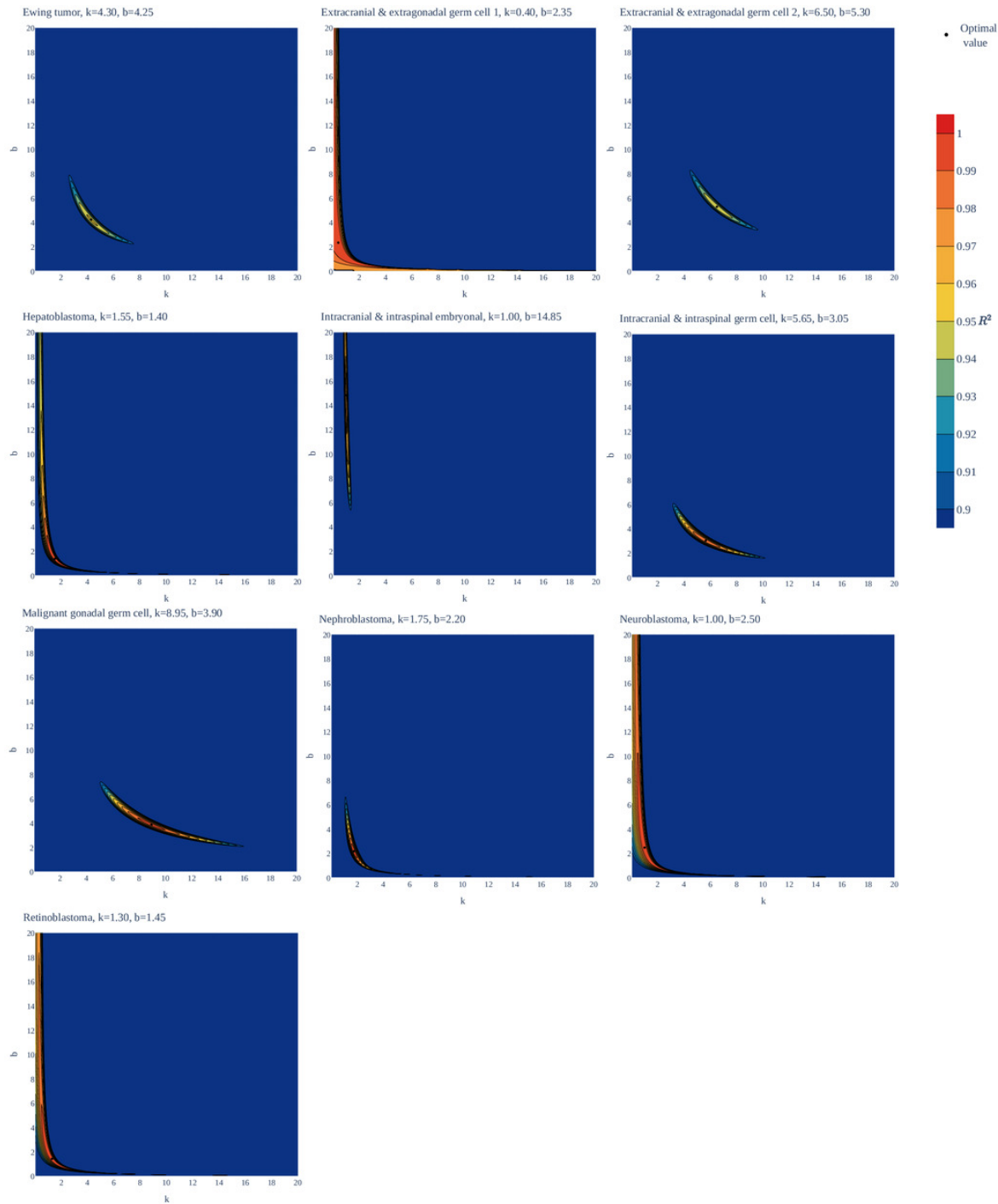
Figure 2

Only the gamma/Erlang and the Weibull distributions fit the actual age distributions of childhood and young adulthood cancer incidence without requiring negative age values.

Dots indicate crude incidence rates for 5-year age groups, curves indicate probability density functions fitted to the incidence data for various childhood/young adulthood cancer types. The middle age of each age group is plotted. See Supplemental Material for the optimal parameter values.
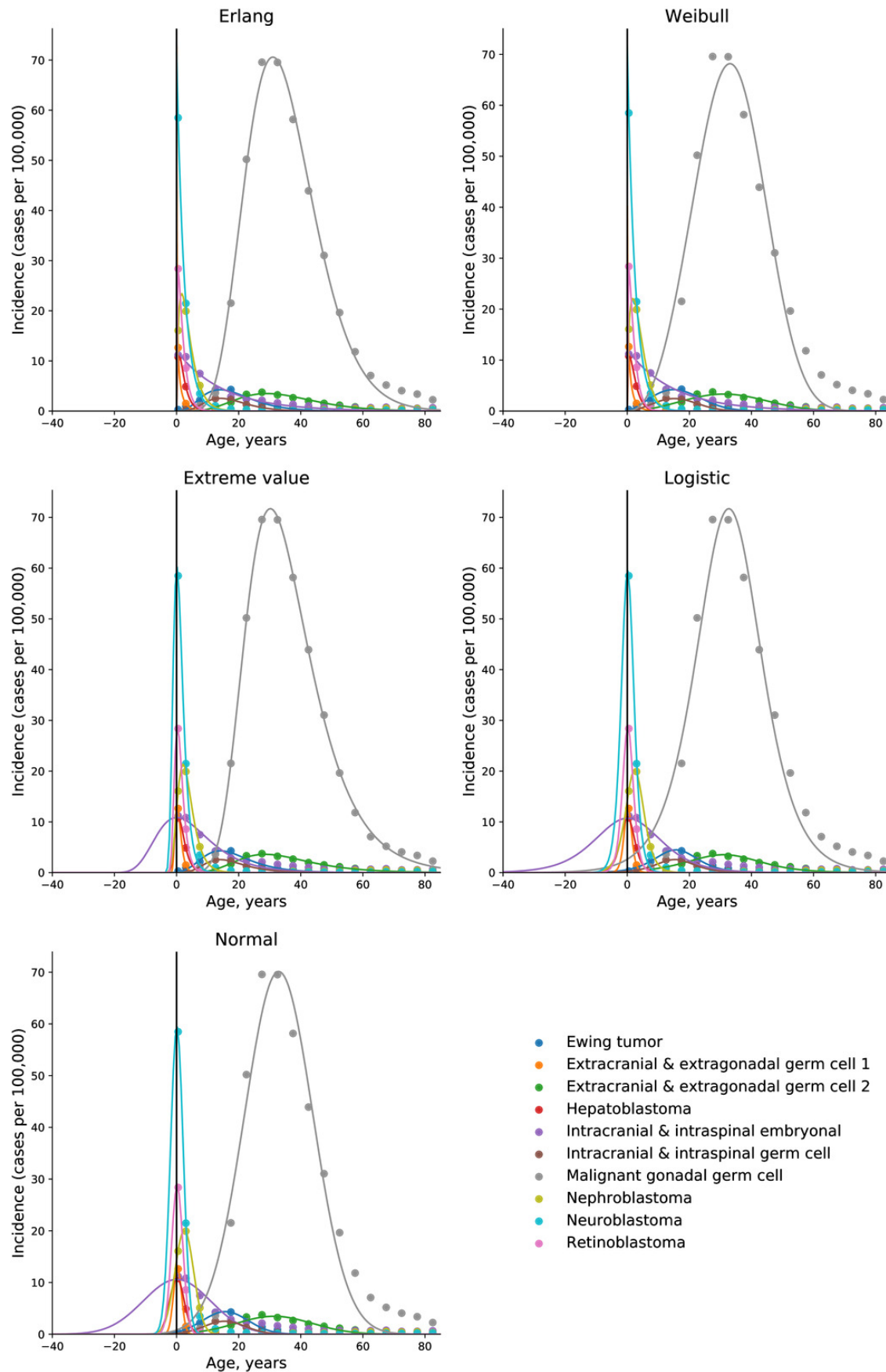
Figure 3

The gamma/Erlang distribution approximates the age distribution of incidence for childhood and young adulthood cancers better than the Weibull distribution.

Dots indicate crude incidence rates for 5-year age groups, curves indicate the probability density function of the gamma/Erlang (red) or Weibull (blue) distribution fitted to the incidence data (see Table 2 for R^2 comparison and Table 3 for estimated parameters). The middle age of each age group is plotted. Extracranial and extragonadal germ cell tumours of childhood and young adulthood are shown on the same plot.