Predicted number of driver events per tumor strongly correlates with contribution from anthropogenic and lifestyle risk factors

Aleksey V. Belikov

School of Biological and Medical Physics, Laboratory of Innovative Medicine,
Moscow Institute of Physics and Technology (MIPT),
Institutsky per., 9, 141701 Dolgoprudny, Moscow Region, Russia.

Correspondence to: belikov.research@gmail.com

Abstract

I have recently shown that the number of rate-limiting driver events per tumor can be estimated from the age distribution of cancer incidence using the Erlang or gamma probability distribution. Here I show that this number strongly correlates with the proportion of cancer cases due to anthropogenic and lifestyle risk factors, such as air pollution, occupational hazards, ionizing radiation, smoking, alcohol, poor diet, insufficient exercise and obesity, but does not seem to correlate with the proportion of cases due to infection or ultraviolet radiation. The correlation was confirmed for three different countries, three corresponding incidence databases, and three risk estimation studies, as well as for both sexes: USA, CDC WONDER database, Islami et al. study, males [$r=0.82$, $P=0.0006$, 13 cancer types], females [$r=0.83$, $P<0.0001$, 16 cancer types]; England, ECIS database, Brown et al. study, males [$r=0.90$, $P<0.0001$, 16 cancer types], females [$r=0.67$, $P=0.002$, 19 cancer types]; Australia, CI5 database, Whiteman et al. study, males [$r=0.90$, $P=0.0004$, 10 cancer types], females [$r=0.68$, $P=0.01$, 13 cancer types]. It is thus confirmed that predictions based on interpreting the age distribution of cancer incidence as the Erlang probability distribution have biological meaning, validating the underlying Poisson process as the law governing the development of the majority of cancer types, with the possible exception of infection- and UV-induced cancers. It also suggests that the majority of driver events (70-80% in males, 50-70% in females) are induced by anthropogenic carcinogens and the lifestyle, and not by cell divisions or other internal processes.

Introduction

I have recently proposed that the age distribution of cancer incidence can be interpreted as the statistical distribution of probability to accumulate the required number of driver events by the given age [1]. I have shown that, of all standard probability distributions, the gamma distribution (and its special case – the Erlang distribution) fits the actual age distribution of incidence for 20 most prevalent cancers the best [1]. I have then shown that the gamma/Erlang distribution is the only standard distribution that, in addition, converges for all studied childhood and young adulthood cancers, thus validating it as the universal equation describing cancer incidence [2]. Importantly, the Erlang distribution describes the waiting time for the occurrence of the given number of independent random events, as it was initially devised to calculate queues at telephone stations. It is based on the Poisson process, which implies not only pure randomness of event timings but also their constant average rate. Thus, the excellent fit of the Erlang distribution to the actual incidence data implies that cancers develop according to the Poisson process, i.e. driver events occur randomly and at a constant average rate.

Interestingly, the parameter k of the Erlang distribution can be interpreted as the number of rate-limiting driver events that occur by the time of cancer diagnosis. It allows to estimate this number for any cancer type, upon fitting the Erlang distribution to the actual age distribution of incidence. I have shown that these numbers vary considerably, from 1 in retinoblastoma [2] to 41 in prostate cancer [1]. Next, it is important to show that these predictions correspond to experimentally observed variables, such as the number of driver mutations per tumor predicted from sequencing data. However, the variability of DNA alterations that can contribute to cancer progression, some of which are not yet routinely assessed, and the imperfection of algorithms for separating driver and passenger mutations severely complicate this task, as discussed in [1]. Thus, a simpler correlate is required to prove the meaningfulness of the predictions, before engaging in a full-scale confirmation effort.

Here I identify such correlate as the percentage of cancer cases due to modifiable risk factors. This is an often-used parameter in epidemiological studies, and is also called the population attributable fraction (PAF). It shows, for example, what percentage of lung cancer cases are caused by smoking tobacco. Combined PAF shows the overall contribution of all potentially modifiable risk factors, which usually include air pollution, occupational hazards, ionizing radiation, smoking, alcohol, poor diet, insufficient exercise, obesity, infection and

ultraviolet radiation. Here I show that the numbers of driver events per tumor predicted by the Erlang distribution strongly correlate with combined PAFs for most cancers, with the exception of cancers with the large contribution from infection and ultraviolet radiation. This confirms that predictions obtained from the Erlang distribution are meaningful, validating the Poisson process as the law governing the development of most cancer types and fostering the search for correlations with tumor sequencing data. Moreover, it suggests that up to 80% of driver events are caused by the environment and lifestyle, and not, for example, by stem cell divisions, as has been recently proposed [3, 4].

Methods

I. Data acquisition

a) Population attributable fractions data

Population attributable fractions (PAFs) combining all risk factors were obtained from published open-access articles separately for each cancer type and sex. PAFs for USA were obtained from the publication by Islami et al., Table 2 [5]. PAFs for England were obtained from the publication by Brown et al., Table 2 [6]. PAFs for Australia were obtained from the publication by Whiteman et al., Table 2 [7].

b) USA incidence data

United States Cancer Statistics Public Information Data: Incidence 1999–2012 was downloaded from the Centers for Disease Control and Prevention Wide-ranging OnLine Data for Epidemiologic Research (CDC WONDER) online database (http://wonder.cdc.gov/cancer-v2012.HTML) in November 2018 [8]. The United States Cancer Statistics (USCS) are the official federal statistics on cancer incidence from registries having high-quality data for 50 states and the District of Columbia. Data are provided by The Centers for Disease Control and Prevention National Program of Cancer Registries (NPCR) and The National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program. Results were grouped by 5-year Age Groups and Crude Rates were selected as output. Crude Rates are calculated as the number of new cancer cases reported each calendar year per 100,000 population in each 5-year age group. The data were downloaded separately for males and females for each cancer type listed in the publication by Islami et al., Table 2 [5].

c) England incidence data

England cancer incidence data were downloaded from the European Cancer Information System (ECIS) Data explorer (https://ecis.jrc.ec.europa.eu/explorer.php?$0-1$1-UK$2-224$4-1,2$3-All$6-5,84$5-1999,2012$7-2$CRatesByCancer$X0_10-ASR_EU_NEW) in November 2018 [9]. The ECIS database contains the aggregated output and the results computed from data submitted by population-based European cancer registries participating in Europe to the European Network of Cancer Registries – Joint Research Centre (ENCR-JRC) project on "Cancer Incidence and Mortality in Europe". Years of observation were limited to 1999-2012 period, to match the USA data. Incidence is calculated as the number of new cancer cases reported each calendar year per 100,000 population in each 5-year age group. The data were downloaded separately for males and females for each cancer type listed in the publication by Brown et al., Table 2 [6], except for vulva and vagina cancers, as their selection was not possible in ECIS Data explorer.

d) Australia incidence data

Australia cancer incidence data were downloaded from the Cancer Incidence in Five Continents (CI5) Volume XI Age-specific curves Online Analysis tool (http://ci5.iarc.fr/CI5-XI/Pages/age-specific-curves_sel.aspx) in November 2018 [10]. CI5 is published approximately every five years by the International Agency for Research on Cancer (IARC) and the International Association of Cancer Registries (IACR) and provides comparable high quality statistics on the incidence of cancer from cancer registries around the world. Volume XI contains information from 343 cancer registries in 65 countries for cancers diagnosed from 2008 to 2012. Incidence is calculated as the number of new cancer cases reported each calendar year per 100,000 population in each 5-year age group. The data were downloaded separately for males and females for each cancer type listed in the publication by Whiteman et al., Table 2 [7].

II. Data selection and analysis

a) Estimation of the number of driver events per tumor

For analysis, the incidence data were imported into GraphPad Prism 6. The following age groups were selected: “5–9 years”, “10–14 years”, “15–19 years”, “20–24 years”, “25–29 years”, “30–34 years”, “35–39 years”, “40–44 years”, “45– 49 years”, “50–54 years”, “55–59 years”, “60–64 years “, “65–69 years”, “70–74 years”, “75–79 years” and “80–84 years”. Prior

age groups were excluded due to possible contamination by childhood subtype incidence, and "85+ years" was excluded due to an undefined age interval. If in the first several age groups ("5–9 years", "10–14 years", "15–19 years") incidence initially decreased with age, reflecting contamination by childhood subtype incidence, these values were removed until a steady increase in incidence was detected. The middle age of each age group was used as the x value, e.g. 17.5 for the "15–19 years" age group. Data were analyzed with Nonlinear regression. The following User-defined equation was created for the gamma distribution:

$$Y = A*(x^{(k-1)})*(\exp(-x/b))/((b^k)*\text{gamma}(k))$$

The parameter *A* was constrained to "Must be between zero and 100000.0" and parameters *b* and *k* to "Must be greater than 0.0". "Initial values, to be fit" for all parameters were set to 1.0. All other settings were kept at default values, e.g. Least squares fit and No weighting.

The numerical value of the *k* parameter is interpreted as the number of driver events per tumor.

b) Correlation of the predicted numbers of driver events per tumor with PAFs

Obtained *k* values were correlated to population attributable fractions (PAFs) in GraphPad Prism 6 using the inbuilt Correlation tool at default settings, e.g. Pearson correlation with two-tailed P value. Cancer types were sorted into two groups – anthropogenic and non-anthropogenic, and correlation was performed separately for each group. Cancer types in which infection (*Helicobacter pylori*, Hepatitis B virus, Hepatitis C virus, Human herpes virus type 8: Kaposi sarcoma herpes virus, Human immunodeficiency virus and Human papillomavirus) or ultraviolet radiation contributed to more than 30% of cases, for a given country according to the published PAF data [5-7], were assigned to the non-anthropogenic group. The rest were assigned to the anthropogenic group, which included cancers with substantial contribution from air pollution, occupational exposure, exposure to ionizing radiation, smoking and exposure to secondhand smoke, alcohol intake, poor diet (red and processed meat, insufficient fiber, vegetables, fruit and calcium), excess body weight, insufficient physical activity, insufficient breastfeeding, postmenopausal hormone therapy and oral contraceptives, according to the published PAF data [5-7].

Results and discussion

While plotting the correlation of the number of driver events per tumor predicted from the Erlang distribution with the estimated percentage of cases due to modifiable risk factors obtained from the published studies, I have noticed that cancers appear to cluster in two groups. One group, which included the majority of cancers, demonstrated the linear correlation, whereas the other clustered in the upper left corner in a cloud-like fashion. Investigation of the "cloud" group revealed that it consists entirely of cancers with substantial (>30%) contribution of infection to their pathogenesis, plus the melanoma cancer. I therefore named this group "non-anthropogenic", as infections and ultraviolet radiation existed long before the human civilization. Interestingly, all cancers in the other group were induced by factors that arose with human civilization, such as air pollution, occupational hazards, ionizing radiation, smoking, alcohol, poor diet, insufficient exercise, obesity, insufficient breastfeeding, postmenopausal hormone therapy and oral contraceptives. Therefore, I termed this group "anthropogenic cancers".

The possible explanation for this dichotomy is that the human body managed to evolve some protective countermeasures against risk factors that were present for millions of years, whereas our recent anthropogenic "developments" take it by surprise. For example, ultraviolet radiation has been present on Earth since the beginning, and although melanocytes cannot completely protect their DNA, and a lot of DNA damage occurs, it is likely that they developed a very slow division rate [11] to avoid conversion of this damage into mutations for as long as possible. This may explain why only few rate-limiting steps are predicted for melanoma despite lots of DNA damage than melanocytes receive – rate-limiting in this case is cell division, and not the DNA damage. Similarly, the human body had plenty of time to adapt to viruses, which may explain why the incidence rates of virus-induced cancers are low and less driver events are predicted than would be expected from the linear correlation. It is also clear that viruses are inducing cancer via different mechanisms than chemical carcinogens [12, 13], and thus the development of such cancers may not be described by the Poisson process. Indeed, many of the virus-induced cancers have rather poor fits of the Erlang distribution to their age distributions of incidence.

The correlation of the predicted number of driver events per tumor with the estimated percentage of cases due to modifiable risk factors for cancers in males is shown in Figure 1 and Table 1, and in females in Figure 2 and Table 2. It can be seen that anthropogenic cancers indeed exhibit the strong correlation for all studied countries and

databases, and for both sexes, whereas non-anthropogenic cancers exhibit the correlation in none of the cases. Amongst anthropogenic cancers, the correlation is stronger and more significant for males than for females. Interestingly, the correlation is stronger and more significant for American females [r=0.83, P<0.0001] than for English [r=0.67, P=0.002] and Australian [r=0.68, P=0.01] females, but weaker and less significant for USA males [r=0.82, P=0.0006] than for English [r=0.90, P<0.0001] and Australian [r=0.90, P=0.0004] males. These differences are likely explained by different exposures to risk factors between countries and sexes, as well as by the different sets of cancers included in the studies from which PAFs were obtained, and the different methodologies of those studies. The role of population genetics also cannot be ruled out.

The strong positive correlation of the predicted number of driver events per tumor with the contribution from anthropogenic risk factors suggests that the majority of driver events are *caused* by those factors. Indeed, as r^2 is called "the coefficient of determination" and describes the proportion of the variance in one variable that is explained by the other variable, we can calculate that anthropogenic risk factors explain 67%, 81% and 81% of the variance in the number of driver events per tumor for males and 69%, 45% and 46% of the variance for females, living in USA, England and Australia, respectively. This is in accord with the mainstream view that the environment and lifestyle are the major contributors to carcinogenesis, but conflicts with the recently proposed view that the majority of cancers develop due to replicative mutations occurring during stem cell division [3, 4]. The latter view is based on predominantly mouse data hand-picked from varied publications and processed through calculations with unobvious assumptions, and thus has been severely criticized [14-17].

Overall, the correlations identified here serve as the validation of the hypothesis that most cancers develop according to the Poisson process and that the Erlang distribution can be used to predict the number of driver events per tumor for each cancer type. This has numerous implications, from the fundamental understanding of the carcinogenesis process to the improvement in driver prediction algorithms.

References

1. Belikov, A.V., *The number of key carcinogenic events can be predicted from cancer incidence.* Sci Rep, 2017. **7**(1): p. 12170.
2. Belikov, A.V. *The Poisson process is the universal law of cancer development: driver mutations accumulate randomly, silently, at constant rate and for many decades, likely in stem cells*. bioRxiv, 2018. DOI: https://doi.org/10.1101/231027.
3. Tomasetti, C. and B. Vogelstein, *Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions.* Science, 2015. **347**(6217): p. 78-81.
4. Tomasetti, C., L. Li, and B. Vogelstein, *Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention.* Science, 2017. **355**(6331): p. 1330-1334.
5. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA Cancer J Clin, 2018. **68**(1): p. 31-54.
6. Brown, K.F., et al., *The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015.* Br J Cancer, 2018. **118**(8): p. 1130-1141.
7. Whiteman, D.C., et al., *Cancers in Australia in 2010 attributable to modifiable factors: summary and conclusions.* Aust N Z J Public Health, 2015. **39**(5): p. 477-84.
8. *United States Cancer Statistics: 1999 - 2012 Archive Incidence, WONDER Online Database*. 2015, United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute: http://wonder.cdc.gov/cancer-v2012.html.
9. *ECIS - European Cancer Information System*. 2018, European Union: https://ecis.jrc.ec.europa.eu.
10. Bray F, C.M., Mery L, Piñeros M, Znaor A, Zanetti R and Ferlay J, editors, *Cancer Incidence in Five Continents, Vol. XI (electronic version)*. 2017, International Agency for Research on Cancer: http://ci5.iarc.fr.
11. Halaban, R., et al., *Human melanocytes cultured from nevi and melanomas.* J Invest Dermatol, 1986. **87**(1): p. 95-101.
12. Butel, J.S., *Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease.* Carcinogenesis, 2000. **21**(3): p. 405-26.

13. Mesri, E.A., M.A. Feitelson, and K. Munger, *Human viral oncogenesis: a cancer hallmarks analysis.* Cell Host Microbe, 2014. **15**(3): p. 266-82.
14. Giovannucci, E.L., *Are Most Cancers Caused by Specific Risk Factors Acting on Tissues With High Underlying Stem Cell Divisions?* J Natl Cancer Inst, 2015. **108**(3).
15. Rozhok, A.I., G.M. Wahl, and J. DeGregori, *A Critical Examination of the "Bad Luck" Explanation of Cancer Risk.* Cancer Prev Res (Phila), 2015. **8**(9): p. 762-4.
16. Wu, S., et al., *Substantial contribution of extrinsic risk factors to cancer development.* Nature, 2016. **529**(7584): p. 43-7.
17. Wensink, M.J., J.W. Vaupel, and K. Christensen, *Stem Cell Divisions Per Se Do Not Cause Cancer.* Epidemiology, 2017. **28**(4): p. e35-e37.

Tables

Table 1. Predicted numbers of driver events per tumor and estimated percentages of cases due to anthropogenic and lifestyle risk factors for cancers in males.

Cancer type	England		USA		Australia	
	Predicted number of driver events per tumour	Estimated percentage of cases due to modifiable risk factors [5]	Predicted number of driver events per tumor	Estimated percentage of cases due to modifiable risk factors [6]	Predicted number of driver events per tumour	Estimated percentage of cases due to modifiable risk factors [7]
Mesothelioma	29	97	-	ND	-	ND
Lung	24	82	28	89	22	86
Larynx	20	73	22	84	21	84
Bladder	20	51	20	49	14	34
Colorectum	18	57	12	58	18	56
Oral cavity	16	53	-	ND	-	ND
Liver	15	53	-	NA	-	NA
Oesophagus	14	61	19	75	14	74
Pancreas	14	34	15	26	12	31
Kidney	13	32	15	52	12	39
Myeloma	12	16	16	11	-	ND
Gallbladder	12	13	18	33	11	16
Brain	9	0	-	ND	-	ND
Leukaemias	9	11	-	ND	8	7
Myeloid leukaemias	-	ND	8	17	-	ND
Non-Hodgkin lymphomas	8	3	7	14	7	4
Thyroid	3	10	6	12	-	ND
Hodgkin lymphoma	-	NA	2	8	-	NA

ND – no data in the source publication, NA – assigned to the non-anthropogenic group due to the strong contribution of the viral infection.

Table 2. Predicted numbers of driver events per tumor and estimated percentages of cases due to anthropogenic and lifestyle risk factors for cancers in females.

Cancer type	England		USA		Australia	
	Predicted number of driver events per tumour	Estimated percentage of cases due to modifiable risk factors [5]	Predicted number of driver events per tumor	Estimated percentage of cases due to modifiable risk factors [6]	Predicted number of driver events per tumour	Estimated percentage of cases due to modifiable risk factors [7]
Lung	23	75	30	83	22	79
Uterus	23	34	20	71	22	33
Mesothelioma	22	83	-	ND	-	ND
Larynx	18	66	25	79	28	78
Pancreas	15	29	15	25	15	28
Bladder	14	43	17	39	11	26
Myeloma	14	11	16	12	-	ND
Kidney	13	36	14	56	10	24
Gallbladder	12	23	14	37	15	13
Ovary	13	11	8	4	7	7
Colorectum	12	51	9	51	11	42
Liver	12	39	-	NA	-	NA
Oesophagus	11	54	17	68	11	76
Non-Hodgkin lymphomas	11	3	10	2	7	3
Oral cavity	9	34	-	ND	-	ND
Breast	9	23	10	29	11	23
Brain	8	5	-	ND	-	ND
Leukaemias	7	13	-	ND	8	2
Myeloid leukaemias	-	ND	6	13	-	ND
Thyroid	3	9	5	13	-	ND
Hodgkin lymphoma	-	NA	2	2	-	NA

ND – no data in the source publication, NA – assigned to the non-anthropogenic group due to the strong contribution of the viral infection.

Figure 1. Correlation of the predicted numbers of driver events per tumor with the estimated percentages of cases due to modifiable risk factors for cancers in males.

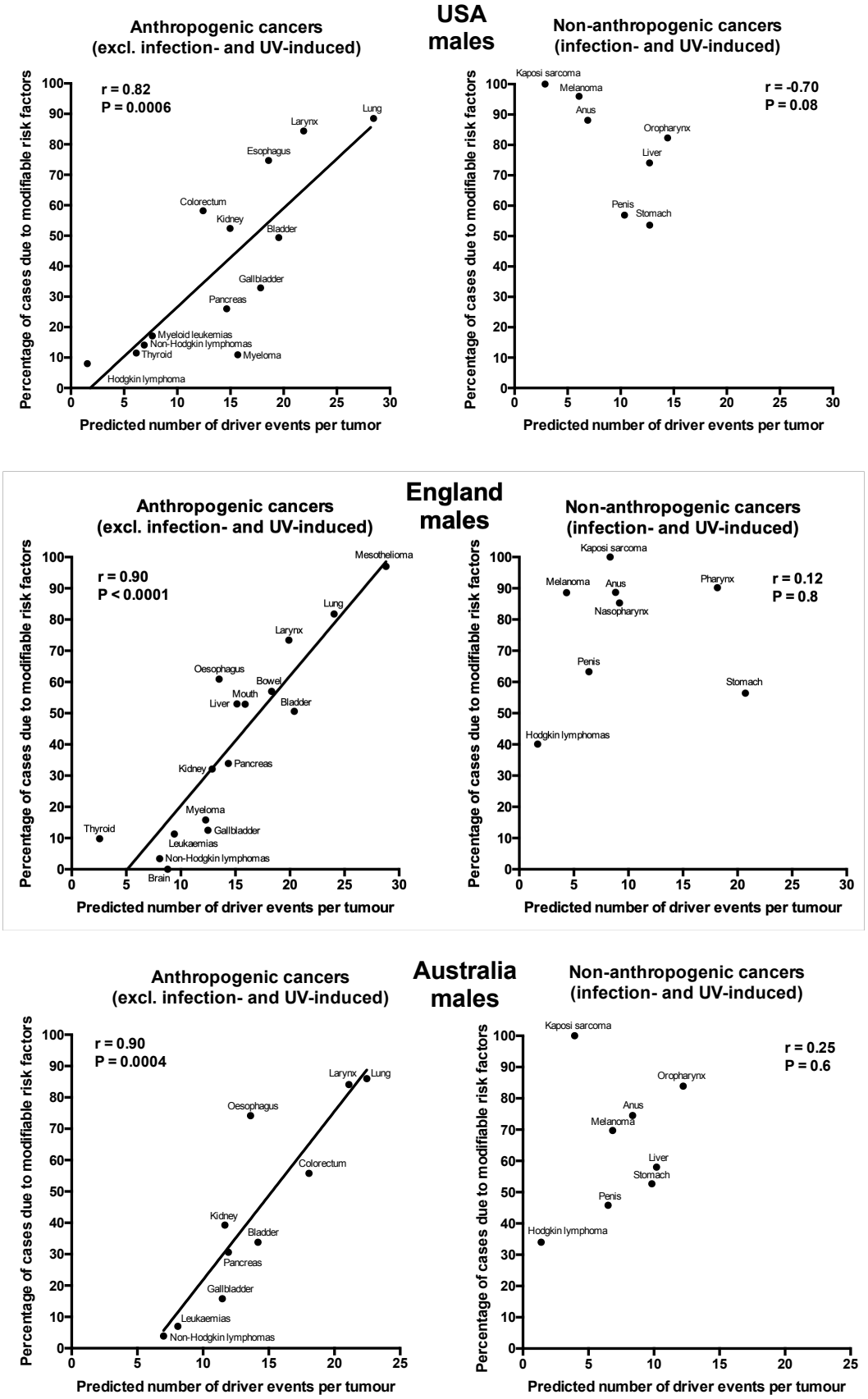

Figure 2. Correlation of the predicted numbers of driver events per tumor with the estimated percentages of cases due to modifiable risk factors for cancers in females.