Approximate Bayesian Computation methods to model Multistage Carcinogenesis

Metodi di Approximate Bayesian Computation per modellare la Cancerogenesi Multistadiale

Consuelo R. Nava, Cinzia Carota, Jordy Bollon, Corrado Magnani, Francesco Barone-Adesi

Abstract A direct modelling of Multistage Carcinogenesis (MC), avoiding mathematical approximations, is here proposed. We take advantage of Approximate Bayesian Computation methods to estimate MC unknown parameters of interest. A simulation of a fictitious cohort of people exposed to a carcinogen is proposed. We show performances of our approach with and without the use of a semi-automatic ABC selection of summary statistics.

Abstract *Si propone una modellizzazione diretta della cancerogenesi multistadiale (MC), evitando approssimazioni matematiche. Metodi di Approximate Bayesian Computation vengono utilizzati per stimare i parametri di interesse della MC. Si simula una coorte fittizia di persone esposte a un agente cancerogeno. Si mostrano le performance del nostro approccio con e senza l'uso di tecniche di selezione semiautomatica delle statistiche descrittive.*

Key words: Approximate Bayesian Computation; semi-automatic selection; rejection algorithm; Multistage Carcinogenesis.

Consuelo R. Nava
Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: c.nava@univda.it

Cinzia Carota
Università degli Studi di Torino, Dipartimento di Economia e Statistica Cognetti de Martiis, Torino (Italy) e-mail: cinzia.carota@unito.it

Jordy Bollon
Università del Piemonte Orientale, Dipartimento di Medicina Traslazionale, Novara (Italy) e-mail: jbollon94@gmail.com

Corrado Magnani
Università del Piemonte Orientale, Dipartimento di Medicina Traslazionale, Novara (Italy) e-mail: corrado.magnani@uniupo.it

Francesco Barone-Adesi
Università del Piemonte Orientale, Dipartimento di Scienze del Farmaco, Novara (Italy) e-mail: francesco.baroneadesi@uniupo.it

1 Introduction

The theory of multistage carcinogenesis (MC) assumes that the transformation of a normal cell into a neoplastic one does not take place in a single step, but rather consists of a multi-stage process [17]. In each stage, normal cells undergo a sequence of genetic mutations which gradually cause the acquisition of tumor cell characteristics [2]. A time-homogeneous birth process governs the transition probability from the i-th to the $i+1$-th stage at time t, with $t = 0,\ldots,T$ [1]. Approximated formulas of MC are available to predict cancer rates at different times [16, 3]], avoiding algebraically cumbersome computations resulting from the system of stochastic differential equations which model MC. Accurate predictions of cancer risk rates are useful [18] to plan health surveillance programs for carcinogenic agents. However, epidemiological studies using MC are presently limited, and they usually rely only on approximated formulas [12, 5, 21].

Different authors pointed out that in some situations the use of approximated formulas, indeed, can lead to an overestimation of hazard rates [16]. Hence, it would be desirable to use the "exact" MC model to evaluate the evolution of cancer risk with the age and, eventually, with long-term carcinogenic exposure. However, due to its high complexity, MC does not allow to define a likelihood function. A possible solution is represented by Bayesian methods, now increasingly used in population genetics [13]. Specifically, Approximate Bayesian Computation (ABC) methods [14, 19] compare observed data with simulated data not through the likelihood function – assumed to be unavailable – but rather with selected summary statistics, such as means, hazard ratios (etc.), obtained through simulations from the same original model – assumed to be known [11, 13]. Even if recent ABC applications can be found in population genetics [9, 23, 22], infectious disease models [15], and systems biology [20, 26], its use in epidemiology is still limited [24, 27, 6].

As an alternative to approximated formulas [5], we propose ABC methods to model MC and to estimate its unknown parameters $\boldsymbol{\theta}$ of interest. We propose a suitable Rejection algorithm [25] enriched with a semi-automatic variable selection method [10]. A code in R has been developed to model MC to estimate $\boldsymbol{\theta}$: the transition rates among the different stages (which can vary during the exposure to a carcinogen) and the elimination rates of the carginogen from the organism.

To this aim, the article is structured as follows. In Section 2, we present ABC techniques to illustrate how ABC can be integrated to model MC and how summary statistics can be selected. In Section 3, we propose a simulation example which mimics a cohort of subjects exposed to a carcinogen and we present some preliminary results. Section 4 concludes the article with some remarks and suggesting future research.

2 Methodology

In epidemiology and population genetics, limits in the use of frequentist model-based inference arise due to the necessity to include prior information and to explicit the likelihood function.

The first issue could be overcome taking advantage of Bayesian methods. Indeed, estimates of $\boldsymbol{\theta}$ given observed data $\mathcal{D}$ are obtained sampling from the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{D})$, proportional to $L(\mathcal{D}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$, respectively the likelihood function and the prior distribution of $\boldsymbol{\theta}$.

The second issue could be overcome using ABC methods as a rejection technique [19, 25, 23] to compute the likelihood function via simulation. Let's assume that an underlying stochastic process $\mathcal{M}$ given $\boldsymbol{\theta}$ generates $\mathcal{D}$ on which k summary statistics of the data $\boldsymbol{S} = \{S_1, \ldots, S_k\}$ are defined. ABC, in its simplest form, "proposes" a (pseudo)-randomly drawn parameter value $\boldsymbol{\theta}^{(n)}$ from $\pi(\boldsymbol{\theta})$ with $n = 1, \ldots, N$. Hence, simulated artificial data $\mathcal{D}^{(n)}$ are generated from $\mathcal{M}$ given $\boldsymbol{\theta}^{(n)}$. $\boldsymbol{\theta}^{(n)}$ are accepted only if $\mathcal{D}^{(n)}$ is "similar enough" to $\mathcal{D}$ and used to approximate $\pi(\boldsymbol{\theta}|\mathcal{D})$ with the ABC posterior, $\pi(\boldsymbol{\theta}|\mathcal{D}^{(n)})$. To this aim, an acceptance criterion of $\boldsymbol{\theta}^{(n)}$ might be based on $\rho(\mathcal{D}, \mathcal{D}^{(n)}) < \varepsilon$, where $\rho(\cdot,\cdot)$ is a suitable metric and $\varepsilon > 0$ is a tolerance level. Given the high-dimensional data generated in our case, the acceptance criterion should be substituted by $\rho(\boldsymbol{S}, \boldsymbol{S}^{(n)}) < \varepsilon$ with $\boldsymbol{S}^{(n)} = \{S_1^{(n)}, \ldots, S_k^{(n)}\}$, a small number of suitable summary statistics defined on $\mathcal{D}^{(n)}$. In such a way, we combine both the computational convenience of using $\boldsymbol{S}$ to approximate $\pi(\boldsymbol{\theta}|\boldsymbol{S}^{(n)})$ instead of $\pi(\boldsymbol{\theta}|\mathcal{D}^{(n)})$, and main Bayesian inference advantages in epidemiology [8].

Rejection-sampling method [19] needs a small number k of suitable summary statistics [7] to avoid a low acceptance rate or a distorting increment of ε [4] (small values of ε allow an approximately calibrated ABC [14, 10]). The complexity of MC and the long follow-up of patients do not accommodate this requirement, making the selection of $\boldsymbol{S}$ difficult and/or reducing estimation accuracy. Thus, based on a regression adjustment – namely, the robust semi-automatic ABC projection technique [10] – we added an extra stage to the algorithm in [19] in order to overcome this issue and to derive summary statistics within our ABC for MC.

The semi-automatic ABC consist of the following steps: (i) run a pilot ABC based on summary statistics chosen subjectively to identify a region of non-negligible posterior mass, i.e. a training region to simulate parameter values. This is a suitable step when uninformative or improper $\pi(\boldsymbol{\theta})$ are considered; (ii) simulate sets of parameter values $\boldsymbol{\theta}^{(n)}$ from the prior truncated to the training region and generate artificial data $\mathcal{D}^{(n)}$; (iii) use $\boldsymbol{\theta}^{(n)}$ and $\mathcal{D}^{(n)}$ to estimate summary statistics fitting regressions; (iv) select the best model, according to model selection criteria (as BIC), and run ABC with selected summary statistics.

Regressions of step (iii) are the linear ones [10] which have as dependent variables the simulated values of the i^{th} parameter, $\theta_i^{(1)}, \ldots, \theta_i^{(n)} \ldots, \theta_i^{(N)}$, with $i = 1, 2, 3$. A vector-valued function of (non-linear) transformations of the input statistics of the artificial data, $f(\boldsymbol{S}^{(n)}) = \left[\boldsymbol{S}^{(n)}, \boldsymbol{S}^{(n)2}, \boldsymbol{S}^{(n)3}, \boldsymbol{S}^{(n)4}\right]$ – here all first, second, third and fourth powers of individual data point – represents the set of explanatory variables.

The following model is fitted using least squares

$$\theta_i = \mathbb{E}(\theta_i|\mathcal{D}) + \epsilon_i = \beta_{i0} + \boldsymbol{\beta}_i f(\boldsymbol{S}) + \epsilon_i \;\; \forall i = 1,2,3$$

where ϵ_i is a white noise error. The fitted function $\hat{\beta}_{i0} + \hat{\boldsymbol{\beta}}_i f(\boldsymbol{S})$ is an estimate of $\mathbb{E}(\theta_i|\mathcal{D})$. Neglecting the constant, the i^{th} summary statistic for ABC is $\hat{\beta}_i f(\boldsymbol{S})$. Note that the input statistics, $\boldsymbol{S}$, could include raw data and (non-linear) transformations.

In general, we assume J stages in MC, denoted with E_j, with $j = 1,\dots,J$, and a constant transition rate for the cell to go from the state E_j to the state E_{j+1}, $E_j \rightarrow E_{j+1}$ [16]. Here, r_0 is the transition rate $E_1 \rightarrow E_2$, if the individual is not exposed to a carcinogen. We assume that r_0 represents also the transition rates across all the other stages, $E_j \rightarrow E_{j+1}$, with $j = 2,\dots,J-1$. Thus, besides the E_1, transition rates are assumed to be constant (r_0). $r_1 = a \cdot r_0$ denotes the accelerated cell transition rate $E_1 \rightarrow E_2$ which is observed during the exposure to a carcinogen in E_1, while λ represents the clearance of slowly eliminated carcinogens such as asbestos [3, 5]). Hence, $\boldsymbol{\theta} = \{r_0, r_1, \lambda\}$ is the vector of unknown MC parameters of interest as proposed in (1). Given that a likelihood function cannot be defined to model MC, ABC methods allow to approximate the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{D})$.

$$\begin{array}{ll} \boxed{E_1} \xrightarrow[r_0]{} \boxed{E_2} \xrightarrow[r_0]{} \dots \xrightarrow[r_0]{} \boxed{E_j} \xrightarrow[r_0]{} \dots \xrightarrow[r_0]{} \boxed{E_J} & \text{no carcinogen exposure} \\ \boxed{E_1} \xrightarrow[r_1,\lambda]{} \boxed{E_2} \xrightarrow[r_0]{} \dots \xrightarrow[r_0]{} \boxed{E_j} \xrightarrow[r_0]{} \dots \xrightarrow[r_0]{} \boxed{E_J} & \text{carcinogen exposure} \end{array} \tag{1}$$

Under MC and assuming constant clearance λ of the internal dose, d_i, of the carcinogen overtime, the transition rate, r_t, at time t is:

$$\log(r_t) = \alpha + \beta \sum_{i=0}^{t} (d_i \, e^{-\lambda(t-i)}).$$

3 Simulations and preliminary results

We describe a model $\mathcal{M}$ for carcinogen exposure to show how the proposed ABC methods accurately estimate $\boldsymbol{\theta}$ and to construct a general approach to deal with MC. We simulate a fictitious cohort of 5,000 subjects, each of which was observed for a time $T = 100$, where $t = 0$ is the year of birth. A cancer develops if at least one cell reaches the last stage. We assume, without lack of generality, $J = 4$. The carcinogen exposure, consecutive or not, between 15 and 64 years old, could mimic, for instance, the asbestos exposure in an occupational setting [3]. We set $\boldsymbol{\theta} = \{7 \cdot 10^{-6},\, 7 \cdot 10^{-5},\, 0.2\}$ and we assume that 50% of workers were exposed to asbestos. Hazard ratios for each year are computed as summary statistics and used comparatively with the semi-automatic ABC selection method given $f(\boldsymbol{S})$. We run 200,000 simulations with $\varepsilon = 0.005$, and uninformative priors $\pi(\theta_i) = \text{Unif}(0,1)\; \forall i = 1,2,3$, given an own elaborated R code for MC which also recalls EasyABC and abctools packages. Table 1 shows the obtained ABC preliminary estimates of θ from the approximated posterior means, with and without the semi-automatic ABC selection. The former approach results to be more accurate and closer to the original values of the parameters of interest than the one with arbitrarily selected summary statistics (hazard ratio for $t = \{35, 40, 85, 95\}$). Estimations with multiple combinations of summary statistics are carried out. No meaningful improvement of the ABC performance with respect to the one here proposed were obtained.

Table 1 Posterior means approximated with ABC rejection. Standard errors are in parentheses. Summary statistics without semi-automatic selection are hazard ratio of selected years.

Semi-automatic selection	r_0	r_1	λ
No	$6.14{\cdot}10^{-6}$ $(9{\cdot}10^{-7})$	$6.17{\cdot}10^{-5}$ $(1.84{\cdot}10^{-5})$	0.114 (0.078)
Yes	$7.2{\cdot}10^{-6}$ $(7{\cdot}10^{-7})$	$6.64{\cdot}10^{-5}$ $(1.45{\cdot}10^{-5})$	0.216 (0.078)

4 Conclusion

We show that part of the appeal of the proposed ABC approach is its flexibilityWe are planning to apply the proposed methodology to a real cohort of workers exposed to asbestos to predict future mesothelioma rates. The here proposed methodology can be easily implemented to any carcinogen exposure under MC. The code written in R is general enough to accommodate other epidemiological assumptions, such as the asbestos clearance. Future research will be aimed to extend this approach to include more sophisticated ABC methods (MCMC or sequential ABC).

References

1. Armitage, P.: Multistage models of carcinogenesis. Environ Health Perspect. 63, 195-201 (1985)
2. Armitage, P., Doll, R.: The age distribution of cancer and multi-stage theory of carcinogenesis. Br. J. Cancer. 8.1, 1-12 (1954)
3. Barone-Adesi, F., Ferrante, D., Bertolotti, M., Todesco, A., Mirabelli, D., Terracini, B., Magnani, C.: Long-term mortality from pleural and peritoneal cancer after exposure to asbestos: Possible role of asbestos clearance. Int. J. Cancer. 123.4, 912-916 (2008)
4. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. Genetics, 162.4, 2025-2035 (2002)
5. Berry, G.: Prediction of mesothelioma, lung cancer, and asbestosis in former Wittenoom asbestos workers. Occup. Environ. Med. 48.12, 793-802 (1991)
6. Dehideniya, M.B., Drovandi, C.C., McGree, J.M.: Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology. Comput. Stat. Data Anal. 124, 277-297 (2018)
7. Didelot, X., Everitt, R.G., Johansen, A.M., Lawson, D.J.: Likelihood-free estimation of model evidence. Bayesian Anal. 6.1, 49-76 (2011)

8. Dunson, D.B.: Commentary: practical advantages of Bayesian analysis of epidemiologic data. Am. J. Epidemiol. 153.12, 1222-1226 (2001)
9. Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L.: Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. 104.45, 17614-17619 (2007)
10. Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC (with discussion). J R Stat Soc Series B Stat Methodol. 74.3, 419-474 (2012)
11. Grelaud, A., Marin, J.M., Robert, C., Rodolphe, F., Tally, F.: Likelihood-free methods for model choice in Gibbs random fields. Bayesian Anal. 3, 427-442 (2009)
12. Magnani, C., Ferrante, D., Barone-Adesi, F., Bertolotti, M., Todesco, A., Mirabelli, D., Terracini, B.: Cancer risk after cessation of asbestos exposure: a cohort study of Italian asbestos cement workers. Occup. Environ. Med. 65.3,164-170 (2008)
13. Marjoram, P., Tavaré, S.: Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. 7.10, 759-770 (2006)
14. Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. Stat. Comput. 22.6, 1167-1180 (2012)
15. McKinley, T., Cook, A.R., Deardon, R.: Inference in epidemic models without likelihoods. Int. J. Biostat. 5.1 (2009)
16. Moolgavkar, S.H.: The multistage theory of carcinogenesis and the age distribution of cancer in man. JNCI. 61.1, 49-52 (1978)
17. Nordling, C.O.: A new theory on the cancer inducing mechanism. Br J Cancer. 7.1, 68-72 (1953)
18. Peto, J., Decarli, A., La Vecchia, C., Levi, F., Negri, E.: The European mesothelioma epidemic. Br. J. Cancer. 79.3-4, 666-672 (1999)
19. Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol. 16.12, 1791-1798 (1999)
20. Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M.P.H., Richardson, S., Wiuf C.: Using likelihood free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. PLoS Comput. Biol. 3.11, 2266-2278 (2007)
21. Reid, A., de Klerk, N.H., Magnani, C., Ferrante, D., Berry, G., Musk, A.W., Merler, E.: Mesothelioma risk after 40 years since first exposure to asbestos: a pooled analysis. Thorax. 69.9, 843-850 (2014)
22. Saulnier, E., Gascuel, O., Alizon, S.: Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLoS Comput Biol. 13.3, (2017)
23. Sottoriva, A., Tavaré, S.: Population Genetics of Neoplasms. In: Frontiers in Cancer Research, pp. 31-42. Springer, New York (2016)
24. Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A.: Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics. 173.3, 1511-1520 (2006)
25. Tavaré, S., Balding, D.J., Griffiths, R.C., Donnel, P.: Inferring coalescence times from DNA sequence data. Genetics. 145.2, 505-518 (1997)
26. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. Roy. Soc. Interface. 6.31, 187-202 (2007)
27. Walker, D.M., Allingham, D., Lee, H.W.J., Small, M.: Parameter inference in small world network disease models with approximate Bayesian computational methods. PHYSICA A. 389.3, 540-548 (2010)