OPEN ACCESS Freely available online

PLoS one

Parameter Identifiability and Redundancy in a General Class of Stochastic Carcinogenesis Models

Mark P. Little[1]*, Wolfgang F. Heidenreich[2], Guangquan Li[1]

1 Department of Epidemiology and Public Health, Imperial College, London, United Kingdom, **2** Institut für Strahlenschutz, Helmholtz Zentrum München, Neuherberg, Germany

Abstract

Background: Heidenreich *et al.* (*Risk Anal* 1997 **17** 391–399) considered parameter identifiability in the context of the two-mutation cancer model and demonstrated that combinations of all but two of the model parameters are identifiable. We consider the problem of identifiability in the recently developed carcinogenesis models of Little and Wright (*Math Biosci* 2003 **183** 111–134) and Little *et al.* (*J Theoret Biol* 2008 **254** 229–238). These models, which incorporate genomic instability, generalize a large number of other quasi-biological cancer models, in particular those of Armitage and Doll (*Br J Cancer* 1954 **8** 1–12), the two-mutation model (Moolgavkar *et al. Math Biosci* 1979 **47** 55–77), the generalized multistage model of Little (*Biometrics* 1995 **51** 1278–1291), and a recently developed cancer model of Nowak *et al.* (*PNAS* 2002 **99** 16226–16231).

Methodology/Principal Findings: We show that in the simpler model proposed by Little and Wright (*Math Biosci* 2003 **183** 111–134) the number of identifiable combinations of parameters is at most two less than the number of biological parameters, thereby generalizing previous results of Heidenreich *et al.* (*Risk Anal* 1997 **17** 391–399) for the two-mutation model. For the more general model of Little *et al.* (*J Theoret Biol* 2008 **254** 229–238) the number of identifiable combinations of parameters is at most $r+1$ less than the number of biological parameters, where r is the number of destabilization types, thereby also generalizing all these results. Numerical evaluations suggest that these bounds are sharp. We also identify particular combinations of identifiable parameters.

Conclusions/Significance: We have shown that the previous results on parameter identifiability can be generalized to much larger classes of quasi-biological carcinogenesis model, and also identify particular combinations of identifiable parameters. These results are of theoretical interest, but also of practical significance to anyone attempting to estimate parameters for this large class of cancer models.

Citation: Little MP, Heidenreich WF, Li G (2009) Parameter Identifiability and Redundancy in a General Class of Stochastic Carcinogenesis Models. PLoS ONE 4(12): e8520. doi:10.1371/journal.pone.0008520

Editor: Dov J. Stekel, University of Nottingham, United Kingdom

Received September 6, 2009; **Accepted** November 30, 2009; **Published** December 31, 2009

Funding: This work was funded partially by the European Commission under contract FI6R-CT-2003-508842 (RISC-RAD) and FP6-036465 (NOTE). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mark.little@imperial.ac.uk

Introduction

Models for complex biological systems may involve a large number of parameters. In principle it may well be that some of these parameters may not be observed, or be possible to be derived from observed data via regression techniques. Such parameters are said to be unidentifiable or non-identifiable, the remaining parameters being identifiable.

There is a substantial literature on identifiability in stochastic models in various contexts [1,2,3]. Catchpole and Morgan [3] considered identifiability and parameter redundancy and the relations between them in a general class of (exponential family) models. Catchpole and Morgan [3] defined a set of model parameters in an exponential family model to be *redundant* if the likelihood can be written using a strictly smaller parameter vector; otherwise they are *irredundant*. Rothenberg [1], Jacquez and Perry [4] and Catchpole and Morgan [3] also defined a notion of *local identifiability*, to mean that within a neighbourhood of each set of parameter values the likelihood differs for at least some data points. This notion has been extended by Little *et al.* [5] to *gradient weak local identifiability* and *weak local identifiability*. Little *et al.* [5] defined a set of parameters to be *weakly locally identifiable* if the maxima of the likelihood are isolated; they defined parameters to be *gradient weakly locally identifiable* if the turning points (those for which the likelihood derivative with respect to the parameters is zero) are isolated. The results obtained by Little *et al.* [5] (Corollary 2 (ii) and the subsequent Remark (ii)), show that, subject to some regulatory conditions, the number of locally identifiable or (gradient) weakly locally identifiable parameter combinations is equal to the rank of the Hessian matrix, or equivalently the rank of the Fisher information matrix. The notions of identifiability in stochastic models [1,2,3,5], within which framework this paper is set, should be contrasted with the consideration of identifiablity in non-stochastic settings considered by some [4,6,7].

Heidenreich [8] and Heidenreich *et al.* [9] considered parameter identifiability in the context of the two-mutation cancer model [10] and demonstrated that of the five biological parameters in the model, on the basis of the cancer hazard function only three could

be identified. [It should be noted that given extra information, for example on numbers and sizes of intermediate cell compartment clones, there is information on an additional parameter.]

In this paper we consider the problem of identifiability in recently developed carcinogenesis models of Little and Wright [11] and Little *et al.* [12]. These models generalize a large number of other quasi-biological cancer models, in particular those of Armitage and Doll [13], the two-mutation model [10], the generalized multistage model of Little [14], and a recently developed cancer model of Nowak *et al.* [15] that incorporates genomic instability. We shall show that via a specific reparameterization, in the simpler model proposed by Little and Wright [11] in principle combinations of all but two of the model parameters are identifiable, thereby generalizing previous results of Heidenreich [8] and Heidenreich *et al.* [9] for the two-mutation cancer model. For the more general model of Little *et al.* [12] combinations of all but $r+1$ of the model parameters are identifiable, where r is the number of destabilization types, thereby also generalizing all these results. We also identify particular forms of identifiable parameters.

Methods

Parameter Identifiability in the Context of a Stochastic Cancer Model with Genomic Instability

We consider the problem of parameter identifiability in a particular class of stochastic cancer models, those of Little and Wright [11] and Little *et al.* [12]. The ideas used are similar to those employed by Heidenreich *et al.* [9], in particular the use of Cauchy's method of characteristics. We shall assume throughout this section that this model is embedded in a member of the exponential family so that the log-likelihood is given by $L(x|\theta)=\sum_{l=1}^{n}\left[\frac{x_l\varsigma_l-b(\varsigma_l)}{a(\phi)}+c(x_l,\phi)\right]$ where the natural parameters $\varsigma_l=\varsigma_l[(\theta_i)_{i=1}^{p},z_l]$ are functions of the model parameters $(\theta_i)_{i=1}^{p}$ and some auxiliary data $(z_l)_{l=1}^{n}$, but that the scaling parameter ϕ is not. We shall assume that the $\mu_l=b'(\varsigma_l[(\theta_i)_{i=1}^{p},z_l])=z_l\cdot h[(\theta_i)_{i=1}^{p},y_l]$, where $h[(\theta_i)_{i=1}^{p},y_l]$ is the cancer hazard function, and that the $(z_l)_{l=1}^{n}$ are all non-zero. This is generally the case, in particular when cohort data are analysed using Poisson regression models, e.g., as in Little and Wright [11] or Little and Li [16]. By the remarks following Corollary 2 of Little *et al.* [5], proving weak local identifiability of a subset of cardinality k of the biological parameters $(\theta_i)_{i=1}^{p}$ is equivalent to showing that for this subset of parameters $rk\left[\left(\frac{\partial^2 h}{\partial\theta_i\partial\theta_j}\right)_{i,j=1}^{p}\right]=k$.

The model of Little *et al.* [12], generalizing that of Little and Wright [11], which in turn generalizes the model of Little [14], assumes that cells can acquire up to k successive cancer-stage mutations, and any of r (mutually exclusive) types of destabilization mutation(s). Cells become malignant when k cancer-stage mutations have occurred, no matter how many destabilizing mutations there have been. Once a cell has acquired a

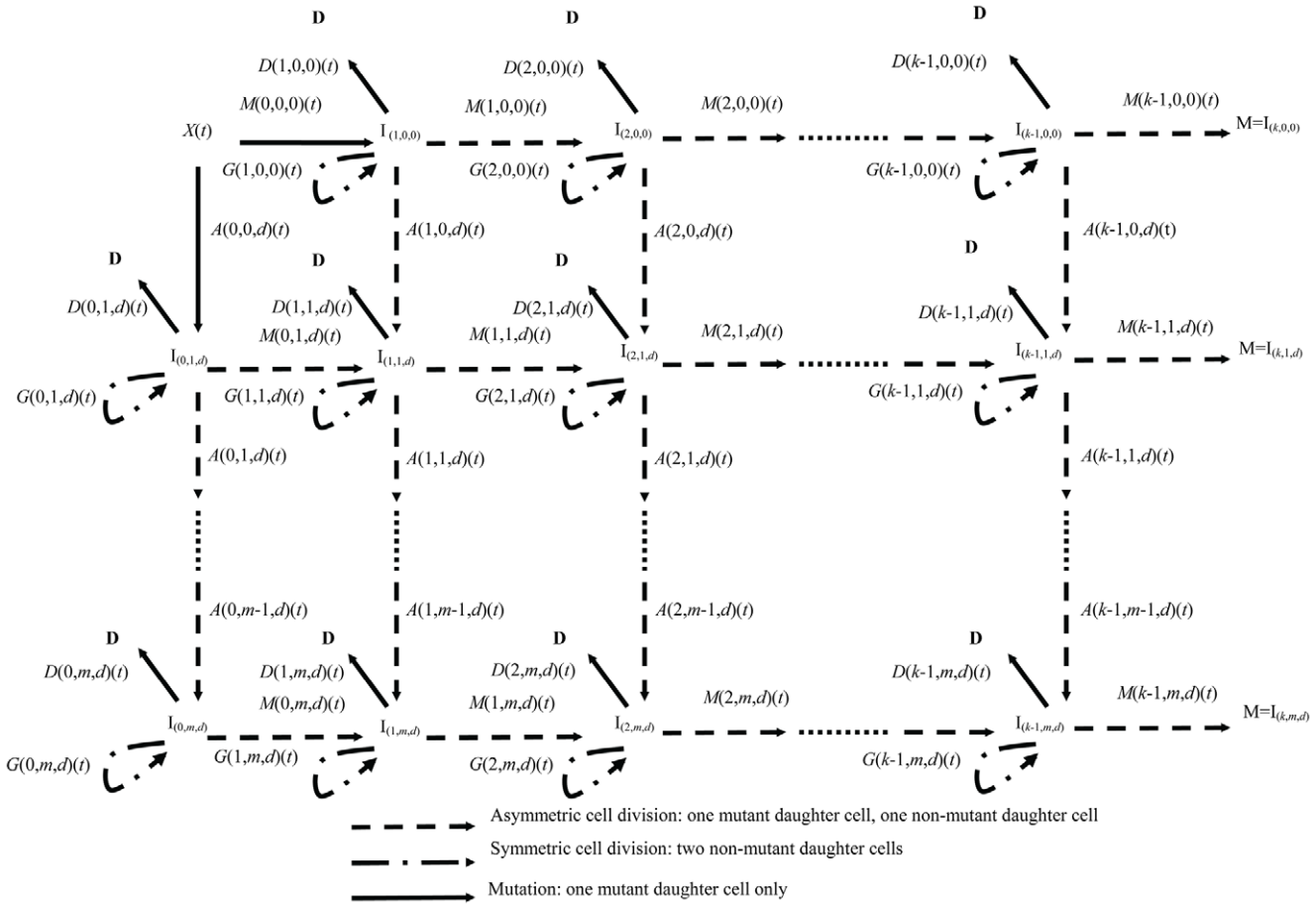

Figure 1. Diagram of cancer model with k cancer-stage mutations and m destabilizing mutations, as in [12].
doi:10.1371/journal.pone.0008520.g001

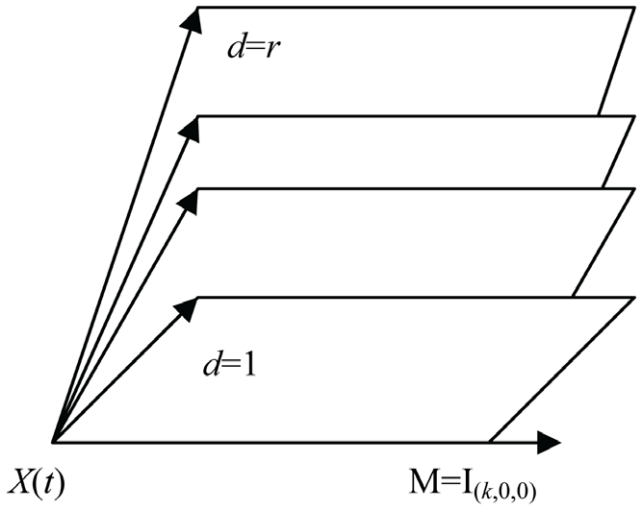

Figure 2. Destabilizing-mutation planes in model, each plane with structure of Figure 1, as in [12].
doi:10.1371/journal.pone.0008520.g002

destabilizing mutation of type d ($1 \le d \le r$), it and its daughter cells can acquire up to $m_d - 1$ further destabilizing mutations of the same type. We define r to be the multiplicity of destabilization mutation types. It is to be expected that the more destabilizing mutations cells acquire of each type, the higher the cancer stage mutation rate is, but this is not intrinsic to the model. We write $(m_1 - m_2 - \ldots - m_r)$ as the *signature of the destabilizing mutation types*. We habitually describe this model as of type $k - r - (m_1 - m_2 - \ldots - m_r)$ for short. The model is illustrated schematically in Figures 1 and 2. Table 1 lists the biological parameters that are used in the model, and their multiplicity.

Cells at different stages of the process are labelled by $I_{(\alpha,\beta,d)}$, where the first subscript, α, represents the number of cancer stage mutations that the cell has accumulated, the second subscript, β, represents the number of destabilizing mutations acquired, their type being given by the third subscript, d. At all stages other than $I_{(0,0,0)}$, cells are allowed to divide symmetrically or differentiate (or undergo apoptosis) at rates $G(\alpha,\beta,d)$ and $D(\alpha,\beta,d)$, respectively.

Table 1. The number of biological parameters in a model with k cancer stages, r7 types of GI and m_d ($d = 1, \cdots, r$) levels of destabilizations.

Model parameter descriptions	Model parameters	Number of such parameters in the model
Stem cell population number	$X(t)$	1
Growth rate	$A(\alpha,\beta,d)(t)$	$k - 1 + k \cdot \sum_{d=1}^{r} m_d$
Death/differentiation rate	$D(\alpha,\beta,d)(t)$	$k - 1 + k \cdot \sum_{d=1}^{r} m_d$
Cancer-stage mutation rate	$M(\alpha,\beta,d)(t)$	$k + k \cdot \sum_{d=1}^{r} m_d$
Destabilizing mutation rate	$A(\alpha,\beta,d)(t)$	$k \cdot \sum_{d=1}^{r} m_d$
Total		$3 \cdot k - 1 + 4 \cdot k \cdot \sum_{d=1}^{r} m_d$

doi:10.1371/journal.pone.0008520.t001

Each cell can divide into an equivalent daughter cell and another cell with an extra cancer stage mutation at rate $M(\alpha,\beta,d)$. Likewise, cells can also divide into an equivalent daughter cell and another cell with an additional destabilizing mutation of type d at rate $A(\alpha,\beta,d)$. The model assumes that there are $X(t)$ susceptible stem cells at age t. Further details on derivation of the hazard function are given in the paper of Little *et al.* [12].

Results

In Text S1 Section B we derive the hazard function and show that it can be written in terms of certain combinations of the biological parameters given in Table 1. From equations (B12)–(B16) in Text S1 Section B it is seen that the characteristics and ψ are governed by certain parameter combinations. Table 2 summarizes the maximum number of identifiable parameter combinations and their forms associated with each cell compartment. The maximum number of identifiable parameters associated with each destabilization zone, $I_{\alpha,\beta,d}$, are 4 when $\alpha < k-1$ and $0 < \beta < m_d$; 4 when $\alpha = k-1$ and $0 < \beta < m_d$; 3 when $\alpha < k-1$ and $\beta = m_d$ and 2 when $\alpha = k-1$ and $\beta = m_d$. The function ψ is governed by at most $r+1$ parameter combinations. Therefore, we have shown that the hazard function $h(\theta)$ can be written as $h(G_1(\theta), G_2(\theta), \ldots, G_N(\theta))$ for some scalar functions $G_1(.), G_2(.), \ldots, G_N(.)$, where $N = (k-2)\cdot(3+r) + (3+r)\cdot 1 + (1+r)\cdot 1 + 4\cdot(k-1)\cdot\sum_{d=1}^{r}(m_d - 1) + 4\cdot\sum_{d=1}^{r}(m_d - 1) + 3\cdot(k-1)\cdot r + 2\cdot r = 3k - 2 - r + 4k\cdot\sum_{d=1}^{r} m_d$ (Table 2). Assuming that the cancer model is embedded in a member of the exponential family (in the sense outlined in Text S1 Section C) the same will be true of the total log-likelihood $L(x|\theta) = L(x|G_1(\theta), G_2(\theta), \ldots, G_N(\theta))$. By means of the Chain Rule we obtain $\frac{\partial^2 L(x|\theta)}{\partial\theta_i \partial\theta_j} = \sum_{l,k=1}^{N} \frac{\partial^2 L(x|G_1,\ldots,G_N)}{\partial G_l \partial G_k} \frac{\partial G_l}{\partial\theta_i}\frac{\partial G_k}{\partial\theta_j} + \sum_{l=1}^{N} \frac{\partial L(x|G_1,\ldots,G_N)}{\partial G_l}\frac{\partial^2 G_l}{\partial\theta_i\partial\theta_j}$, so that the Fisher information matrix is given by

$$I(\theta) = -E_\theta\left[\frac{\partial^2 L(x|\theta)}{\partial\theta_i\partial\theta_j}\right] = -E\left[\sum_{l,k=1}^{N}\frac{\partial^2 L(x|G_1,\ldots,G_N)}{\partial G_l \partial G_k}\frac{\partial G_l}{\partial\theta_i}\frac{\partial G_k}{\partial\theta_j}\right] = -\sum_{l,k=1}^{N}\frac{\partial G_l}{\partial\theta_i} E\left[\frac{\partial^2 L(x|G_1,\ldots,G_N)}{\partial G_l\partial G_k}\right]\frac{\partial G_k}{\partial\theta_j} \quad (1)$$

which therefore has rank at most N. A similar argument shows that if one were to reparameterise (via some invertible C^2 mapping $\theta = f(\omega)$) then the embedded log-likelihood $L(x|f^{-1}(\theta)) = L(x|\omega)$ associated with $h(f^{-1}(\theta)) = h(\omega)$ must also have Fisher information matrix of rank at most N. By Theorems 1 and 3 of Catchpole and Morgan [3], for this embedded exponential family model therefore there can be at most N irredundant parameters. Therefore, of the theoretically available $1 + 2\cdot[k - 1 + k\cdot\sum_{d=1}^{r} m_d] + k + 2\cdot k\cdot\sum_{d=1}^{r} m_d = 3k - 1 + 4k\cdot\sum_{d=1}^{r} m_d$ biological parameters (Table 1), at most $N = 3k - 2 - r + 4k\cdot\sum_{d=1}^{r} m_d$ parameter combinations are identifiable, indicating a minimum of $(r+1)$ parameter redundancies in the model. Also, from the results obtained by Little *et al.* [5] (Corollary 2 (ii) and the subsequent
Remark (ii)), subject to some regulatory conditions, the number of locally identifiable or (gradient) weakly locally identifiable parameter combinations is equal to the rank of the Fisher

Table 2. Parameter combinations associated with each cell compartment. The forms of these combinations are extracted from equations (B12)–(B16) in Text S1.

Compartment $I_{\alpha,\beta,d}$	Number of such compartments	Forms of identifiable parameter combinations	Maximum number of identifiable parameter combinations	Total maximum number of identifiable parameter combinations
Principal axis (non-destabilization) $I_{\alpha,0,0}$ ($\alpha=0,\ldots,k-1$, $\beta=d=0$)				
$0<\alpha<k-1$	$(k-2)$	$G(\alpha,0,0)$, $D(\alpha,0,0)-G(\alpha,0,0)$, $\frac{M(\alpha,0,0)}{G(\alpha+1,0,0)}$, $\left(\frac{A(\alpha,0,d')}{G(\alpha,1,d')}\right)d'=1^r$	$3+r$	$(k-2)\cdot(3+r)$
$\alpha=k-1$	1	$G(\alpha,0,0)$, $D(\alpha,0,0)-G(\alpha,0,0)+M(\alpha,0,0)$, $M(\alpha,0,0)$, $\left(\frac{A(\alpha,0,d')}{G(\alpha,1,d')}\right)d'=1^r$	$3+r$	$3+r$
ψ ($\alpha=\beta=d=0$)	1	$\frac{X\cdot M(0,0,0)}{G(1,0,0)}$, $\left(\frac{X\cdot A(0,0,d')}{G(0,1,d')}\right)d'=1^r$	$1+r$	$1+r$
r destabilization zones ($0\le\alpha\le k-1$, $1\le\beta\le m_d$, $1\le d\le r$)				
$\alpha<k-1$, $1\le\beta<m_d$	$(k-1)\cdot\sum_{d=1}^{r}(m_d-1)$	$G(\alpha,\beta,d)$, $D(\alpha,\beta,d)-G(\alpha,\beta,d)$, $\frac{A(\alpha,\beta,d)}{G(\alpha,\beta+1,d)}$, $\frac{M(\alpha,\beta,d)}{G(\alpha,\beta+1,d)}$	4	$4\cdot(k-1)\cdot\sum_{d=1}^{r}(m_d-1)$
$\alpha=k-1$, $1\le\beta<m_d$	$\sum_{d=1}^{r}(m_d-1)$	$G(\alpha,\beta,d)$, $M(\alpha,\beta,d)$, $D(\alpha,\beta,d)-G(\alpha,\beta,d)+M(\alpha,\beta,d)$, $\frac{A(\alpha,\beta,d)}{G(\alpha,\beta+1,d)}$	4	$4\cdot\sum_{d=1}^{r}(m_d-1)$
$\alpha<k-1$, $\beta=m_d$	$(k-1)\cdot r$	$\frac{M(\alpha,\beta,d)}{G(\alpha+1,\beta,d)}$, $D(\alpha,\beta,d)-G(\alpha,\beta,d)$, $G(\alpha,\beta,d)$	3	$3\cdot(k-1)\cdot r$
$\alpha=k-1$, $\beta=m_d$	r	$D(\alpha,m_d,d)-G(\alpha,m_d,d)+M(\alpha,m_d,d)$, $G(\alpha,m_d,d)\cdot M(\alpha,m_d,d)$	2	$2\cdot r$
Total				$3k-2-r+4k\cdot\sum_{d=1}^{r}m_d$

doi:10.1371/journal.pone.0008520.t002

information matrix, so $\le N$. For example, in the case of the familiar two-mutation model [10], with $k=2$, $r=1$, $d=0$ and $m_d=0$, there are $k\cdot(m+1)-1=2\cdot1-1=1$ G's (namely $G(1,0,0)$), $k\cdot(m+1)-1=2\cdot1-1=1$ D's (namely $D(1,0,0)$), $k\cdot m=2\cdot0=0$ A's, $k\cdot(m+1)=2\cdot1=2$ M's (namely $M(0,0,0),M(1,0,0)$), and a single X, giving a total of five biological parameters. It is known from the results of Heidenreich *et al.* [8,9] that for the two-mutation model only three combinations of these are estimable, i.e., that there are two redundancies, precisely in agreement with the result given here for $r=1$. This result therefore precisely generalizes the results and approach of Heidenreich *et al.* [8,9]. Unfortunately, analytical methods for proving that precisely this number of parameters are estimable, including some recently outlined [17], cannot be used for the model considered here. Nevertheless, we conjecture that in fact precisely this number of parameters are estimable, so that the upper bound on the number of estimable parameter combinations that we have proved above is in fact sharp. This is supported by numerical evaluation of the Hessian in a couple of example cases, which we now outline.

Numerical Evaluation of Hessian and Determination of Its Rank

That there are likely to be exactly this number of estimable parameters is supported by numerical evaluation of the Hessian matrix of the hazard function. We make use of the solution of the system of ordinary differential equations defining the Hessian, outlined in Text S1 Section D. We will show in two cases that the Hessian has rank two less than the number of biological parameters, w. By the above-mentioned results of Catchpole and Morgan [3] and Little *et al* [5] this suggests that precisely $w-2$ parameters are (gradient) weakly locally identifiable. In order to show that the Hessians are of rank two less than the number of biological parameters, w, we evaluate the eigenvalues of the Hessian matrix, and establish that the smallest eigenvalue among the $w-2$ largest eigenvalues in absolute value exceeds the likely magnitude of the error by at least an order of magnitude. We know the likely size of the error in numerical evaluations of each element, h_{ij}, of the Hessian from the Boerlisch-Stoer integrator that is employed, namely $\max(10^{-10},10^{-10}\cdot|h_{ij}|:1\le i,j\le w)$ (**bsstep** routine, Press *et al.* [18], p.722). It is known that if two symmetric matrices H and $\tilde{H}$ have eigenvalues $\lambda_1\le\lambda_2\le\ldots\lambda_{w-1}\le\lambda_w$ and $\tilde{\lambda}_1\le\tilde{\lambda}_2\le\ldots\tilde{\lambda}_{w-1}\le\tilde{\lambda}_w$ then $|\lambda_i-\tilde{\lambda}_i|\le||H-\tilde{H}||_2$, $1\le i\le w$, where $||H||_2=\sup[||Hx||_2/||x||_2:x\ne0]$ [19](p.396). Since the approximate Hessian that we calculate, $\tilde{H}$, differs from the true Hessian, H, by an amount $||H-\tilde{H}||_2\le\sqrt{w}\cdot\max[|h_{ij}-\tilde{h}_{ij}|:1\le i,j\le w]$, we know that:

$$|\lambda_i-\tilde{\lambda}_i|\le\sqrt{w}\cdot\max[|h_{ij}-\tilde{h}_{ij}|:1\le i,j\le w]\le\sqrt{w}\cdot\max[10^{-10},10^{-10}\cdot|\tilde{h}_{ij}|:1\le i,j\le w] \quad (2)$$

There is also the issue of numerical roundoff error in the QR algorithm (Numerical Algorithms Group (NAG) routine **F02FAF** [20]) used to compute eigenvalues. If we write now $\tilde{\lambda}_i,\hat{\tilde{\lambda}}_i$ for the true and approximate eigenvalues associated with the approximate Hessian, $\tilde{H}$, this is known to be bounded by:

$$|\tilde{\lambda}_i-\hat{\tilde{\lambda}}_i|\le c(w)\cdot\varepsilon\cdot||\tilde{H}||_2\le c(w)\cdot\varepsilon\cdot\sqrt{w}\cdot\max[|\tilde{h}_{ij}|:1\le i,j\le w],\quad 1\le i\le w \quad (3)$$

where $c(w)$ is a modestly increasing function of the dimension, w, of the approximate Hessian $\tilde{H}$ and ε is the machine precision [19](Chapter 8). Since the machine precision (in double precision) is of the order 10^{-15} this expression (3) will be dominated by the error associated with the approximation to the Hessian, given by expression (2).

Table 3. Example coefficients of model with three cancer stage mutations and one destabilizing mutation.

Coefficient	Value
$G(1,0,0)$	$8.64714335947694\times10^{-2}$
$G(2,0,0)$	$1.06188950764276\times10^{-3}$
$D(1,0,0)$	$4.25556779736062\times10^{-2}$
$D(2,0,0)$	$2.68975909218019\times10^{-1}$
$M(0,0,0)$	$1.33167380928588\times10^{-2}$
$M(1,0,0)$	$1.08841503240502\times10^{0}$
$M(2,0,0)$	$9.79093689335407\times10^{-2}$
$A(0,0,1)$	$1.33537580655960\times10^{-1}$
$A(1,0,1)$	$7.65789029061483\times10^{-2}$
$A(2,0,1)$	$3.73742902997137\times10^{-2}$
$G(0,1,1)$	$5.31044255713088\times10^{-1}$
$G(1,1,1)$	$1.32418227810710\times10^{1}$
$G(2,1,1)$	$6.88863709884594\times10^{-2}$
$D(0,1,1)$	$1.14118194976730\times10^{-2}$
$D(1,1,1)$	$2.99644035332771\times10^{-1}$
$D(2,1,1)$	$8.92155178101449\times10^{-1}$
$M(0,1,1)$	$7.55711980917015\times10^{0}$
$M(1,1,1)$	$6.58304546585478\times10^{0}$
$M(2,1,1)$	$4.33636256393215\times10^{-3}$
X	$4.06993305645860\times10^{0}$

doi:10.1371/journal.pone.0008520.t003

We evaluated the Hessian matrix for a model with three cancer-stage mutations and one destabilizing mutation, and a model with two cancer-stage mutations and one destabilizing mutation; log-normal perturbations of all parameters were performed, assuming a geometric standard deviation (GSD) of 4, centred on models with cancer-stage mutation rates of 4.0×10^{-3} year^{-1}, destabilizing mutation rates of 3.0×10^{-3} year^{-1}, intermediate cell proliferation rates of 1.0×10^{-1} year^{-1}, and intermediate cell death rates of 5.0×10^{-1} year^{-1}. For each of 1000 random sets of parameters we evaluated the Hessian by numerical integration, as outlined in Text S1 Section D. We calculated the eigenvalues of the Hessian using the QR algorithm, specifically the NAG FORTRAN subroutine **F02FAF** [20]. For each model we selected the set of random parameters for which the ratio of minimum to maximum among the $w-2$ largest eigenvalues (w being the number of biological parameters) in absolute value was greatest. These are given in Tables 3 and 4, for the three-stage and two-stage models, respectively. The associated eigenvalues are given in Table 5. The absolute value of the $w-2$th smallest eigenvalue associated with each set exceeds the error bound (2) by at least an order of magnitude in each case. This strongly suggests that the Hessians calculated for these two examples really are of rank $w-2$ for each model.

Table 4. Example coefficients of model with two cancer stage mutations and one destabilizing mutation.

Coefficient	Value
$G(1,0,0)$	$2.22095885699822\times10^{-3}$
$D(1,0,0)$	$1.31378739613141\times10^{-6}$
$M(0,0,0)$	$8.12022029775447\times10^{-4}$
$M(1,0,0)$	$1.40674010365097\times10^{-5}$
$A(0,0,1)$	$2.06668108660923\times10^{-1}$
$A(1,0,1)$	$4.57214970326658\times10^{-3}$
$G(0,1,1)$	$1.56644835664010\times10^{-2}$
$G(1,1,1)$	$3.16379145991048\times10^{-4}$
$D(0,1,1)$	$1.29917705679554\times10^{0}$
$D(1,1,1)$	$1.92969737536413\times10^{-1}$
$M(0,1,1)$	$9.58173133172697\times10^{0}$
$M(1,1,1)$	$2.26339224702545\times10^{-1}$
X	$2.78141105650539\times10^{-1}$

doi:10.1371/journal.pone.0008520.t004

Table 5. Eigenvalues in ascending order of Hessian matrix associated with a model with three cancer stage mutations and one destabilizing mutation (as in Table 3), and with a model with two cancer stage mutations and one destabilizing mutation (as in Table 4).

Number	Eigenvalues (Table 3)	Eigenvalues (Table 4)
1	$-1.20726415206490\times10^{1}$	$-1.45810346778189\times10^{0}$
2	$-4.92487558715060\times10^{0}$	$-7.77741441881355\times10^{-1}$
3	$-1.11648980088601\times10^{0}$	$-2.77127189259301\times10^{-1}$
4	$-2.44711976272777\times10^{-1}$	$-6.66243518532325\times10^{-3}$
5	$-9.84288250086772\times10^{-2}$	$-3.53209777682867\times10^{-4}$
6	$-1.23814589706358\times10^{-2}$	$-2.86471102388267\times10^{-4}$
7	$-2.95522329598474\times10^{-3}$	**$-9.25930409562877\times10^{-6}$**
8	$-1.53669876331947\times10^{-3}$	**$-1.78637642487767\times10^{-11}$**
9	$-9.80139032107413\times10^{-5}$	$2.74342908757636\times10^{-4}$
10	$-3.36238129341872\times10^{-5}$	$4.98697524563660\times10^{-4}$
11	$-2.14105771381677\times10^{-6}$	$1.11215731049368\times10^{-2}$
12	**$-1.86967299054058\times10^{-7}$**	$8.18426507233826\times10^{-1}$
13	**$5.01559183858810\times10^{-12}$**	$1.45195703291853\times10^{0}$
14	$9.44044820094881\times10^{-7}$	-
15	$4.05661818962605\times10^{-4}$	-
16	$1.92220119614334\times10^{-3}$	-
17	$1.11042617352459\times10^{-2}$	-
18	$1.03277102432191\times10^{-1}$	-
19	$1.12667702944003\times10^{0}$	-
20	$1.08248991510735\times10^{1}$	-

Non-significant eigenvalues are underlined in bold.

doi:10.1371/journal.pone.0008520.t005

Discussion

We have shown that in the class of stochastic cancer models incorporating genomic instability developed by Little and Wright [11] the number of identifiable combinations of parameters is at most two less than the number of biological parameters, thereby generalizing previous results of Heidenreich *et al.* [8,9] and Hanin *et al.* [21,22] for the two-mutation model, a special case of this model. For the more general genomic-instability cancer model of Little *et al.* [12] the number of identifiable combinations of parameters is at most $r+1$ less than the number of biological parameters, where r is the number of destabilization types, thereby

also generalizing all these results. Numerical evaluations in two special cases (with $r=1$) suggest that this bound is tight: a combination of parameters with cardinality two less than the number of biological parameters is of full rank, and so is not redundant.

A weakness of the paper is that one cannot be absolutely sure (because of the uncertainty implicit in any numerical evaluation) that the bound demonstrated by the mathematics of section 3 and Text S1 Section B is sharp. Nevertheless, we have clearly established a maximum number of identifiable parameter combinations. We have also specified particular combinations of identifiable parameters, and these should be used in model fitting to avoid obvious numerical problems, of lack of convergence and absence of a unique set of parameters maximizing the likelihood.

These results have obvious implications for the large number of other quasi-biological cancer models that are special cases of these models, in particular those of Armitage and Doll [13], the two-mutation model [10], the generalized multistage model of Little [14], and a recently developed cancer model of Nowak *et al.* [15] that incorporates genomic instability. It should be noted that the results given here are for the fully stochastic solution of the model, and would not be applicable, for example, to the deterministic approximation of the multistage model of Armitage and Doll [13] that is often employed in applications.

Our results imply that for the general class of cancer models considered here, only certain specific parameter combinations should be estimated in principle, and this is the case whatever the size of the dataset being considered. Whether for complex models for even this theoretically available number of parameters there is useful information is of course uncertain, and may well depend on the particular dataset and on the likely size of the parameters to be estimated. However, fits to a large population-based registry of colon cancer, as recently analysed by Little and Li [16], suggests that, for example, the model with two cancer-stage and one destabilizing mutations can be fitted to the dataset and yields stable parameter estimates for certain combinations of 11 parameters, in accordance with the results of this paper.

Supporting Information

Text S1 Text S1
Found at: doi:10.1371/journal.pone.0008520.s001 (0.42 MB DOC)

Acknowledgments

The authors are very grateful for the comments of Professor Byron Morgan on an advanced draft of the paper. The authors are also grateful for the detailed and helpful comments of the two referees.

Author Contributions

Conceived and designed the experiments: MPL WFH GL. Performed the experiments: MPL. Analyzed the data: MPL GL. Wrote the paper: MPL WFH GL.

References

1. Rothenberg TJ (1971) Identification in parametric models. Econometrica 39: 577–591.
2. Silvey SD (1975) Statistical inference. London: Chapman and Hall. 192 p.
3. Catchpole EA, Morgan BJT (1997) Detecting parameter redundancy. Biometrika 84: 187–196.
4. Jacquez JA, Perry T (1990) Parameter estimation: local identifiability of parameters. Am J Physiol 258: E727–E736.
5. Little MP, Heidenreich WF, Li G (2009) Parameter identifiability and redundancy: theoretical considerations. PLoS ONE, submitted (also available on arXiv:0812.4701).
6. Chappell MJ, Gunn RN (1998) A procedure for generating locally identifiable reparameterisations of unidentifiable non-linear systems by the similarity transformation approach. Math Biosci 148: 21–41.
7. Evans ND, Chappell JM (2000) Extensions to a procedure for generating locally identifiable reparameterisations of unidentifiable systems. Math Biosci 168: 137–159.
8. Heidenreich WF (1996) On the parameters of the clonal expansion model. Radiat Environ Biophys 35: 127–129.
9. Heidenreich WF, Luebeck EG, Moolgavkar SH (1997) Some properties of the hazard function of the two-mutation clonal expansion model. Risk Anal 17: 391–399.
10. Moolgavkar SH, Venzon DJ (1979) Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. Math Biosci 47: 55–77.
11. Little MP, Wright EG (2003) A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. Math Biosci 183: 111–134.
12. Little MP, Vineis P, Li G (2008) A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data. J Theoret Biol 254: 229–238. 255: 268.
13. Armitage P, Doll R (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 8: 1–12.
14. Little MP (1995) Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the multistage model of Armitage and Doll. Biometrics 51: 1278–1291.
15. Nowak MA, Komarova NL, Sengupta A, Jallepalli PV, Shih I-M, et al. (2002) The role of chromosomal instability in tumor initiation. Proc Natl Acad Sci U S A 99: 16226–16231.
16. Little MP, Li G (2007) Stochastic modelling of colon cancer: is there a role for genomic instability? Carcinogenesis 28: 479–487.
17. Cole D, Morgan BJT (2009) Determining the parametric structure of Non-linear models, University of Kent School of Mathematics, Statistics and Actuarial Science preprint, downloadable from http://www.kent.ac.uk/ims/personal/djc24/para.pdf.
18. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in FORTRAN, the art of scientific computing (2nd ed.). Cambridge: Cambridge University Press. 500 p.
19. Golub GH, van Loan CF (1996) Matrix computations (3rd ed.). Baltimore: The Johns Hopkins University Press. 728 p.
20. Numerical Algorithms Group (2006) NAG Library, Mark 21. Oxford: Numerical Algorithms Group.
21. Hanin LG, Yakovlev AY (1996) A nonidentifiability aspect of the two-stage model of carcinogenesis. Risk Anal 16: 711–715.
22. Hanin LG (2002) Identification problem for stochastic models with application to carcinogenesis, cancer detection and radiation biology. Discrete Dyn Nature Soc 7: 177–189.
23. McCullagh P, Nelder JA (1989) Generalized linear models (2nd ed.). London: Chapman and Hall. 511 p.