Radiotherapy and Oncology 114 (2015) 310–313

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

SNP prediction of morbidity

A simulated SNP experiment indicates a high risk of over-fitting and false positive results when a predictive multiple SNP model is established and tested within the same dataset

Christian Nicolaj Andreassen *

Department of Experimental Clinical Oncology, Aarhus University Hospital, Denmark

ARTICLE INFO

Article history:
Received 9 January 2015
Accepted 3 February 2015
Available online 19 February 2015

Keywords:
Normal tissue toxicity
SNP
Predictive model

ABSTRACT

Several relatively small studies have established predictive models for normal tissue radiosensitivity based on multiple SNPs. Even though these models yielded statistically significant results, the models were often inconsistent with each other. This can presumably be attributed to certain methodological problems related to the way these models were established and tested. In order to explore this potential problem, we conducted 10 simulated SNP experiments based on randomly assigned 'SNP genotypes' applied to a set of real clinical data. In 8 out of 10 times, a significant result was found for the model. This clearly demonstrates that the process of fitting the model to the dataset is indeed per se capable of producing nominally significant results. Thus, great caution should be taken when a multiple SNP model is established and tested within the same patient cohort.

 Radiotherapy and Oncology 114 (2015) 310–313

Since the human genome was sequenced at the turn of the millennium, great interest was taken in genetics and genomics in various scientific fields [1,2]. Normal tissue radiobiology is no exception [3]. During the last decade, more than a hundred published studies have addressed possible associations between single nucleotide polymorphisms (SNPs) and risk of radiation-induced normal tissue toxicity (Fig. 1). The ultimate aim of these efforts is to establish a predictive test for normal tissue radiosensitivity [4]. Some of these studies establish predictive models based on multiple SNPs. Examples are given in references [5–12]. For several reasons, such approach seems attractive. First of all, it is in accordance with prevailing assumption that normal tissue radiosensitivity is a so-called complex trait [1,2] dependent on the combined influence of a number of different loci [3,13]. Furthermore, it may offer a solution to a nagging problem in radiogenomics: an increasing amount of evidence indicates that the typical impact of the individual SNP is rather small often corresponding to genotype relative risks below 1.5 and often well below 1.2 [1,2]. Dependent on the exact genotype distribution and the proportion of 'reactors' in the study cohort, a sample size of approximately a thousand patients will usually be needed to detect such small differentials [13]. In normal tissue radiobiology, it often represents a challenge to establish such large cohorts of patients that are well-characterized in terms of treatment parameters and normal tissue outcome [14]. At first sight, the 'multiple SNP model' may seem like a welcomed method to circumvent this problem: although the study is not powered to detect the impact of each individual SNP, the combined influence of several SNPs might be sufficiently strong to be detected even in a relatively small study. And if the model works, one may argue that this provides an indirect proof that the underlying SNPs are in fact associated with normal tissue complication risk.

We have taken a closer look at four different studies that established predictive models based on multiple SNPs [5–8]. All studies were relatively small with sample sizes between 37 and 69 patients. Three studies addressed radiation-induced fibrosis whereas one addressed late toxicity in broader terms. The studies utilized a similar methodology: they assessed a limited number of SNPs. For some of the SNPs, a so-called risk allele was defined and the studies then looked for an association between the total number of risk alleles and normal tissue complication risk. Some of these models had a number of SNPs in common (XRCC1 codon 399, XRCC3 codon 241, TGFB1 codon 10 and ATM codon 1853). Furthermore, all four models yielded significant results with *p*-values as low as 0.0005 (Table 1). Nevertheless, a closer look reveals substantial inconsistencies between the models. For the XRCC1 codon 399 Arg/Gln SNP, the Arg allele was appointed as risk allele in two studies whereas the Gln allele was appointed as risk allele in the other two studies. Similar contradictory findings were observed for the other SNPs (Table 1) [15]. So, even though each

* Address: Department of Experimental Clinical Oncology, Aarhus University Hospital, Noerrebrogade 44, 8000 Aarhus C, Denmark.
E-mail address: nicolaj@oncology.au.dk

http://dx.doi.org/10.1016/j.radonc.2015.02.004
0167-8140/

model seems to work brilliantly the models are completely incompatible with each other. This presumably relates to the way these models were constructed. The process basically had three steps: (1) for each of the included SNPs, a risk allele (minority vs. majority allele) was defined based on the observation that it was (usually non-significantly) associated with the outcome parameter of the study (radiosensitivity). (2) A model was established based on these risk alleles. (3) A statistical test was carried out to determine if the number of risk alleles was significantly associated with radiosensitivity (the same parameter as used for the selection of the risk alleles). By doing so, a kind of circularity is introduced into the analysis that makes it likely that random fluctuations (for the individual SNPs) are amplified into significant associations (for the entire model) simply due to the way the risk alleles were selected in the first place. When the models were established, either of the two possible alleles for each SNP could be defined as the risk allele. This provides the opportunity to 'flip over' associations that are 'pointing in the wrong direction'. Furthermore, SNPs not having any strong association with radiosensitivity in the dataset could be omitted from the model. When several factors pointing in the same direction are combined, a significant association is likely to occur. Thus, the models may have an inherent tendency to produce false positive findings. In order to further explore this potential problem, we conducted a simulated SNP experiment based on randomly assigned 'SNP genotypes' applied to a set of real clinical data.

Material and methods

The present study was based on the dataset originally used to establish the multiple SNP model published by Andreassen et al. in 2003. The clinical material has previously been described in detail [5]. In short, the study cohort was made up by 41 breast cancer patients given post-mastectomy radiotherapy using a three field technique. The patients were scored for subcutaneous fibrosis in each of the treatment fields. Furthermore, detailed dosimetric recordings were available for each field. Thus, the data were well-suited for dose–response assessments. The original study assessed 7 SNPs of which 6 were combined into a multiple SNP model as described in the introduction. Dose–response curves were established for patients with a low and a high number of risk alleles respectively. Statistical significance was determined by comparing the ED_{50} values for these two dose–response curves [5].

Instead of the actual SNP genotypes, we randomly assigned 'genotypes' to the 41 patients for 7 fictitious SNPs that had the same relative distribution as the SNPs in the original dataset. Subsequently, we selected 'risk alleles' for 5–6 of these 'SNPs' and established a multiple SNP model that was tested exactly as in the original study [5]. This procedure was repeated 10 times hence producing 10 independent multiple SNP models based on randomly assigned 'SNPs'.

Results

In 8 out of 10 times, a significant result was found for the model. Fig. 2 shows the result of one of the 10 simulated SNP experiments. Fig. 3 provides a waterfall plot showing the *p*-values obtained for the models.

Discussion

Even though our study was based on randomly assigned genotypes, significant results were obtained for the multiple SNP models in the majority of the cases. Thus, the calculations clearly demonstrate that the process of actively fitting the model to the dataset is indeed per se capable of producing nominally significant results. The scientific method usually involves two separate steps. *First*, a hypothesis is formulated. *Then*, a set of observational data is used to test whether the hypothesis is likely to be true. In the case of the multiple SNP models, this process is short-circuited: at the starting point, the hypothesis is only vaguely formulated when the SNPs to be investigated are selected. In the next step, the observational dataset is opened up and based on these data, the predictive model is finished up as the SNPs to be included are selected and the risk alleles are defined. In that way, the model is specifically adapted to fit that particular dataset. Thus, the model somehow represents an imprint of the dataset rather than a real scientific hypothesis. The comparison between the model and the dataset therefore departs fundamentally from the scheme of scientific hypothesis testing. The usual definition of the *p*-value is the probability of getting the observed result or a more extreme result due to random fluctuations alone. It is indeed true that the probability of getting the distribution of patients observed for the SNP experiments due to coincidence alone is rather small. Nevertheless, when we did the initial 'sneak look' and prearranged the input variables (risk alleles) to make them fit the outcome parameter (radiosensitivity); the comparison with a random segregation becomes irrelevant. We literally paved the way for the association when we selected the risk alleles and standard statistical tests are therefore no longer applicable.

Analytical methods have been developed that enables a simultaneous test of numerous genetic markers. These methods are referred to as 'penalized multi marker regression models' [23]. Examples of such models are 'the lasso' and 'the elastic net'. These methods are specifically designed to counteract over-fitting (hence the term 'penalized') thereby keeping the false discovery

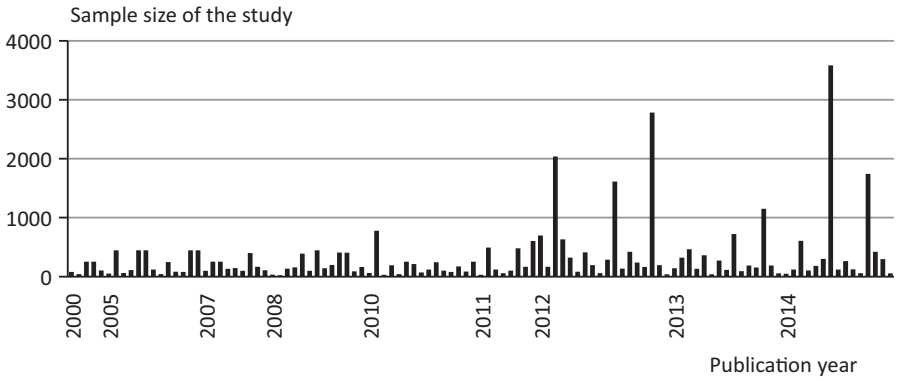

Fig. 1. Overview of 111 published SNP studies addressing normal tissue radiosensitivity showing the sample size of each study.