DR MARINELA CAPANU AND DR COLIN BEGG

Finding mutations that matter

Dr Marinela Capanu and **Dr Colin Begg** tell us about their progress in developing new statistical approaches that, it is hoped, may help to identify key genetic mutations that lead to cancer

Can you outline the goals and aims of your research proposal involving rare genetic variants associated with cancer risks?

Recent technological progress has led to rapid identification and sequencing of large numbers of genetic variants. It is critical to identify which genetic mutations are harmful, so that one can appropriately counsel carriers of these mutations. This is a challenging task, since, typically, individual variants are frequently rare, so there is relatively little empirical evidence available about each individual mutation. The goal of this research is to develop new statistical approaches for determining which mutations are harmful by using data from genetic association studies, and to examine the technical validity of these new methods.

As we move increasingly into the genomic era, accumulated evidence about disease-causing genes is obtained from 'association' studies. Can you explain what this term entails, and the means by which such studies are conducted?

Genetic association studies involve comparison of subjects with the disease (cases) with individuals who are disease-free (controls). These groups are then compared in terms of the frequency of each genetic variant identified in a gene of interest. Ideally these are 'population-based', whereby all individuals who are diagnosed in a defined population are identified, and are approached to participate. Controls are randomly selected from the population.

In conventional statistical methods, the occurrence of only one 'event' fails to provide enough evidence to classify the mutation as either harmful, or harmless, with any degree of confidence. In what ways does your statistical model differ from this?

New mutations are continually being identified and many of them will occur very infrequently in any study (possibly in only one or two subjects). Clearly, these frequencies on their own are insufficient to provide meaningful risk predictions. However, by aggregating the results from individual rare variants on the basis of characteristics shared by groups of variants, using hierarchical modelling, we can identify groups of variants that are similar on the basis of genetic criteria – such as conservation among species – and others. If a group of this nature is identified that possesses, collectively, a high ratio of cases to controls, then we can infer that membership of a variant in the group implies high risk. In this way it is possible to obtain more accurate predictions about the impact of the individual variants on disease risk.

In what ways do you plan to study the properties of your hierarchical modelling approach to determine the circumstances in which we can be confident of future results?

Our approach involves detailed simulations. These are computer generated examples of the use of the technique. In a simulation, you know at the outset the 'true' results, i.e. which mutations are harmful and which are harmless. By studying the method in action in this way, we can figure out how frequently it will successfully classify the mutations, and we can study the extent to which the degree of confidence in our results may be overstated by the method. We can also make refinements to the method, and use the simulations to help us determine which refinements produce better results.

Can you outline any obstacles or challenges faced in implementing your proposal, and how they were overcome?

In the context of such sparse data, the obvious challenge one faces is finding a method that provides accurate risk predictions for each individual rare variant, and one that does so in a reasonable amount of time. We introduced a computationally efficient hybrid approach that involved pseudo-likelihood estimation of the relative risk parameters with Bayesian estimation of the variance components. This method was shown to be fast and straightforward to implement, and had good statistical properties.

By what means is your progress evaluated, and what do you consider to be your principal achievement thus far?

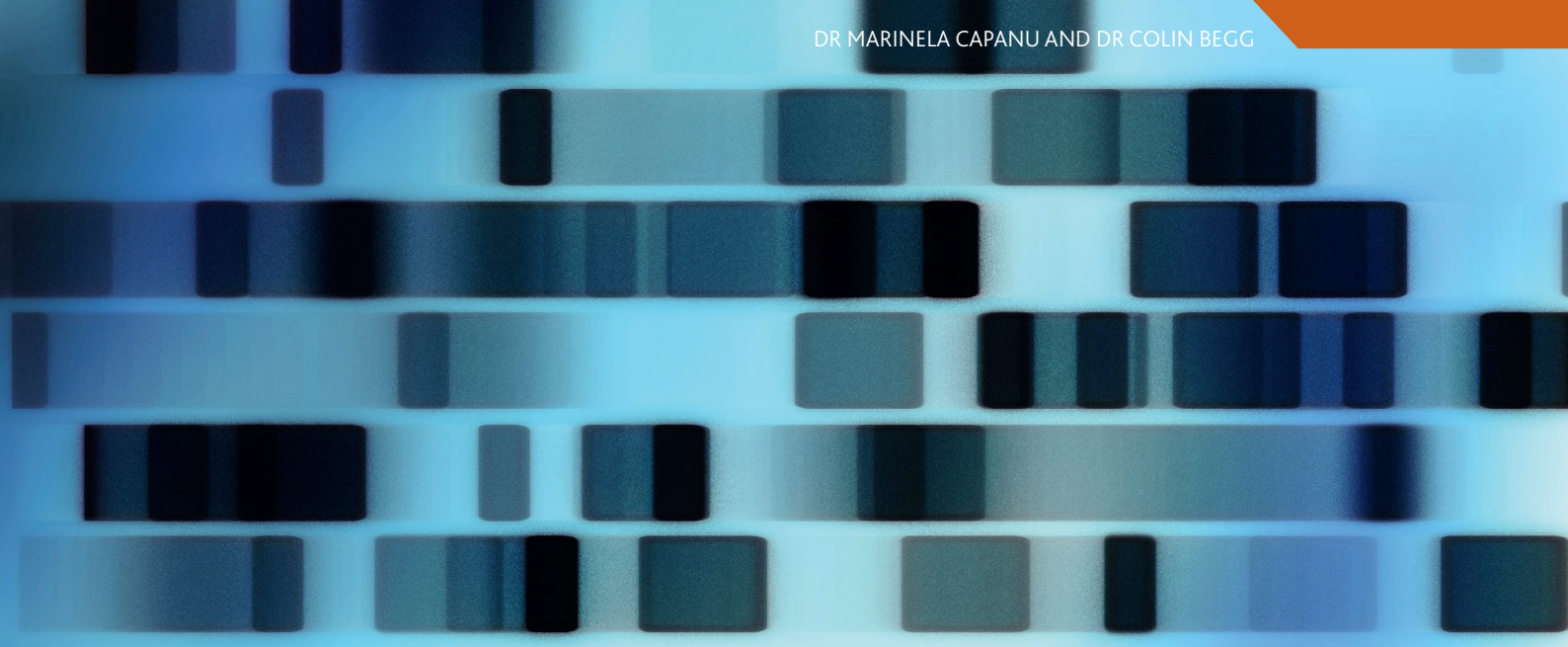
The principal achievement so far has been the development of an approach, the validity of which has been based on simulation results. Another important achievement of this study is the application of this approach to large studies involving melanoma and breast cancer.

How important do you consider the role of computer technologies in the advancement of gene-related pathological studies?

Advancement of computer technologies has played, and will continue to play, a crucial role in genetic studies. Hierarchical modelling analysis, involving hundreds or rare genetic variants, would not be feasible if not for the advancement of computational capabilities and the development of appropriate software packages. Future improvements in this field will open the door to developing even better tools to address this important issue.

How might your work act as a basis upon which further studies in this crucial area can be facilitated?

Our work establishes the potential of statistical modelling in identifying rare variants from sparse data. However, the method – as currently implemented – provides a basis only for the analysis of data from a single case-control study. The method needs to be developed further to permit aggregation of data from multiple case-control studies, and also to enable the synthesis of information from studies of various kinds, including both family-based and association studies.



Uncovering the mutations that cause cancer

Modern genomics research is opening new doors into understanding how genetic mutations can lead to cancer. **Dr Marinela Capanu**, Assistant Attending Biostatistician at Memorial Sloan-Kettering Cancer Center, is developing new statistical techniques that may be able to uncover rare variants that so far have been difficult to pin down

ALTHOUGH THERE ARE a myriad of risk factors associated with the development of cancer, there is strong evidence that genetic mutations play an important role in its pathogenesis. Progress in the relatively modern field of genomics has helped to identify some of the key genes that seem to strongly influence the likelihood of developing the disease. Although mutations occur in these genes, they appear in many different locations, and while some are deleterious (harmful), others are harmless.

In order to advance cancer research, it is crucial that the specific mutations that cause the condition are identified, and the individuals concerned given the appropriate advice. Achieving this, however, is a very challenging prospect; new mutations are regularly found, but with little evidence of their actual functions. Because the relevant mutations are usually hereditary, traditional techniques to uncover them have been based on finding multiple occurrences of the disease in families that are susceptible. Although this technique has been successful so far, it is nonetheless limited because it will only identify the variants that have a high penetrance (probability of expressing the malignant phenotype).

GROUPING BY ASSOCIATION

Modern genomic research is now leaning towards a more sophisticated approach to uncovering deleterious genetic mutations. This takes the form of association (case-control) studies, in which large groups consisting of individuals with a type of cancer (the case group) are compared with healthy individuals (the control group). All mutations are identified and statistical analysis undertaken to determine their propensity to cause disease. This technique has the benefit of being more representative of the larger population, which is diverse and contains far more genetic variants than the selected families studied traditionally. Association studies cast a far wider net and include genetic mutations with differing levels of penetrance.

> Ultimately, our research will produce a technique with good statistical properties to tackle the important scientific problem of identifying rare genetic mutations that confer disease risk

Among those undertaking such research are Dr Marinela Capanu and Dr Colin Begg, of Memorial Sloan-Kettering Cancer Center. Capanu is keen to underline the value of their research: "As more data from case-control (association) studies are accumulated, and as genome-wide genotyping methods achieve greater resolution, these studies assume a greater importance in the search for cancer genes, and for determining which variants within these genes carry risk," she explains.

FINDING THE NEEDLE IN THE HAYSTACK

Many of the previous large-scale epidemiological case-control studies on cancer have uncovered genetic mutations in some subjects. For instance, these have been found in the genes CDKN2A (melanoma) and BRCA1 and BRCA2 (breast cancer). However, some of these variants have only been observed in very few participants, while for many variants only a single participant (out of hundreds) was found to possess the mutation. Capanu highlights the difficulties this presents: "The immediate evidence available for evaluating variant-specific risks consists merely of the relative case-control frequencies of the few subjects in the study that harbour the variant of interest," she says. This would normally provide insignificant statistical power to classify the mutation as deleterious, neutral or beneficial. Clearly, finding stronger evidence of the function of such uncommon variants is required, or else little meaning can be ascribed to them.

Capanu's team have adopted a more efficient technique to overcome this limitation. Their research is based on hierarchical statistical modelling, which works through grouping individual rare mutations with similar characteristics ('bioinformatic' predictors), and calculating whether they have a higher frequency in positive cases than in healthy controls. The variants can then be categorised into high and low risk groups on the basis of their aggregated case-control ratios. With such data, this method allows researchers to analyse multiple groupings, while also allowing for the effects of overlapping groups and other risk factors for disease. As

ESTIMATING CANCER RISKS OF RARE GENETIC VARIANTS

OBJECTIVES

Recent technological progress has led to rapid identification and sequencing of large numbers of genetic variants. The goal of this research is to develop new statistical approaches for determining which mutations are harmful using data from genetic association studies, and to examine the technical validity of these new methods.

COLLABORATORS

Dr Marinela Capanu Co-PI
Dr Colin Begg Co-PI

CONTACT

Marinela Capanu, PhD

Assistant Attending Biostatistician
Memorial Sloan-Kettering Cancer Center
Department of Epidemiology and Biostatistics
307 East 63rd Street, 3rd Floor
New York, NY 10065, USA

T +1 (646) 735-8120
F +1 (646) 735-0012
E CapanuM@mskcc.org

www.mskcc.org/mskcc/html/60449.cfm

DR CAPANU has been as Assistant Attending Biostatistician at Memorial Sloan-Kettering Cancer Center since completing her Doctorate at the University of Florida in 2005. Her current research interests are in the development of hierarchical models for epidemiologic studies to identify genetic variants that increase the risk of cancer. Capanu is involved in the WECARE (Women's Environment Cancer and Radiation Exposure) Study which investigates the joint associations of radiation exposure and genetic variation in the ATM and BRCA1/2 genes in the etiology of breast cancer. She is also engaged in clinical collaborations with the Gastrointestinal Oncology Service, providing assistance on the design and analysis of prospective and retrospective studies.

DR COLIN BEGG is a biostatistician with a long track record of research on statistical methods applicable in medical research. Since 1989 he has served as Chair of the Department of Epidemiology and Biostatistics. He is also Head of the Prevention Control and Population Research Program at Memorial Sloan-Kettering Cancer Center.

Memorial Sloan-Kettering Cancer Center

Capanu stresses, the value of such a technique is significant: "Hierarchical modelling provides the mathematical structure for accomplishing this grouping of variants within a single statistical analysis," she asserts.

COMBINING OLD TECHNIQUES TO CREATE NEW ONES

With the use of computer simulations, the team has been validating their statistical method and fine-tuning it accordingly to increase accuracy. Faced with the challenge of finding a technique that was suitable for use with such sparse data, they decided that a Bayesian analysis using Gibbs sampling and the pseudo-likelihood method were most appropriate for their needs. However, choosing the appropriate estimation method from these presented another challenge. This is because an analysis using Gibbs sampling can take a long time even on a fast computer, while the pseudo-likelihood method uses mathematical assumptions that may not be entirely valid for such sparse data. In the end, they settled on a hybrid of the two which was able to combine the benefits of both, but without their limitations.

DEVELOPING TOOLS THAT ARE FIT FOR PURPOSE

The work of Capanu and Begg is innovative, but remains in its early stages. At this point, they have found only modest associations between the current bioinformatic predictors and the risk of disease. Capanu is realistic but hopeful for the future of such research: "Future improvement of bioinformatic tools to predict functional relevance will enhance the ability of this hierarchical modelling approach to predict the risk conferred by individual rare variants," she says, before elaborating: "Ultimately, our research will produce a technique with good statistical properties to tackle the important scientific problem of identifying rare genetic mutations that confer disease risk". Such progress is impossible without the concerted efforts of a multidisciplinary team consisting of biostatisticians, epidemiologists and cancer biologists, along with specialists from other relevant fields. Their hard work could well pave the way to a new era of cancer genomics that promises not to leave behind those who have inherited rare variants of important genes.