# Mining dosimetry data: Sun exposure behaviors in hereditary melanoma participants

Tracy C. Petrie[1], Tammy K. Stump[2,3], Lisa G. Aspinwall[2,3], Pamela Cassidy[1], Jennifer M. Taber, Steve Jacques[1], Paul Tanner[2], Richard McKenzie[4], Ben Liley[4], Sancy Leachman[1]

1. Oregon Health & Science University, Portland, Oregon, United States
2. The University of Utah, Salt Lake City, Utah, United States
3. Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, United States
4. National Institute of Water & Atmospheric Research (NIWA), Lauder, Central Otago, New Zealand

**Abstract**. UVR dosimetry data has been collected as part of the BRIGHT project. The context of the data is explained and various challenges with the data are described along with strategies for dealing with these challenges. Application of a Bayesian Network for mining the disparate types of collected data, including the dosimetry data, is proposed.

## Introduction

Of all cancers, melanoma has perhaps the best characterised environmental etiologic contributor, ultraviolet radiation (UVR). Approximately 10% of melanoma patients have a hereditary pattern of inheritance and approximately 25-40% of these hereditary melanoma cases are associated with carrying a cyclin-dependent kinase inhibitor 2A (CDKN2A) mutation. CDKN2A is also referred to as *p16* as it will be through the rest of this paper. Like another well-known oncogene, *p53*, *p16* is a tumour suppressor and evidence suggests a mitigating role in melanoma, among other cancers. A mutation in this gene subsequently reduces the body's defence against melanoma.

In the United Kingdom, this high-risk population has an approximate 58% lifetime risk of developing melanoma by age 80, whereas mutation carriers living in Australia have a 91% lifetime risk. This increased gene penetrance of *p16* in a geographical area with higher UVR exposure suggests that hereditary melanoma patients would benefit substantially by optimizing photoprotection regimens. Published data on an initial cohort of hereditary melanoma patients suggests that provision of genetic test results improved compliance with photoprotection and screening behaviours. However, these data were dependent on delayed retrospective self-reporting by the participants. To increase our confidence in this causal relationship, we sought to develop objective measures of photoexposure, including the use of a UVR dosimeter, and then correlate this with subjective reporting. These objective measures should provide a better assessment of actual behavioural changes motivated by genetic test reporting.

## BRIGHT

The Behavior, Risk Information, Genealogy and Health Trial, or BRIGHT Project is a prospective longtitudinal study of, among other things, changes in photoprotection behavior following melanoma genetic counseling and test reporting using a variety of objective and subjective measurements. The subjective measurements include a new self-reporting metric of photoprotection (the Protection Adjusted Length of Exposure or PALE) and a record of sunburns. The objective measurements include skin color differences measured with a Konica Minolta CM-700d spectrophotometer and UVR exposure measured with Scienterra UVR dosimeters.

Several other questions regarding the psychological impact of genetic test results are included in the study, but only the photoprotection aspects are considered here. Two fundamental questions are at the heart of this component of the study: 1) does counseling for high-risk patients change their photoprotection behavior and 2) does a positive-for-mutation *p16* test result change their photoprotection behavior significantly more than counseling alone?

## UVR Dosimetry Measurements

For the UVR dosimetry component of the study, a baseline set of measurements is acquired one month prior to genetic test reporting and then repeated during a one month follow-up after reporting to assess short-term changes. A further one-month sample is acquired one year later – during the same calendar month as the initial follow-up sample to minimize seasonal effects – to assess long-term compliance. Patients are selected from families known to carry the *p16* mutation and a novel control group is selected from families known *not* to carry the *p16* mutation but who have similar risk to *p16*+ patients. None of the patients have or have had melanoma.

At this point in the study, drawing conclusions from the dosimetry data would be premature due to some discovered difficulties in interpreting the data. For this discussion, cases where patients did not strictly adhere to the dosimeter wearing protocol will be called *protocol adherence negative* or PA- and correctly wearing the dosimeter will be called PA+. As an example, the data on many patients show several days where no UVR was observed by the dosimeter. The initial hypothesis would be to assume the patients were PA-. It might instead be that those patients practiced avoidance as their primary form of photoprotection. To then presume PA- would lead to a faulty conclusion. Similarly, there are many days where the dosimeter records a small number of very short periods (under a minute) of UVR. Many possible explanations can be hypothesized for these results. They could be small static shocks, an indoors patient who passes near a window, an outdoors patient who wears the badge under a long sleeve who scratches their head for a moment, etc. If these readings are in fact random noise that is expected in these devices, the data should be excluded (in a documented way) from calculations of compliance. If these readings are not normal device noise, they would support compliance to some degree. On some days, it is very clear that the patient only wore the device when, for instance, the got home from work. This kind of behavior does not support a clear conclusion because the patient was indeed PA+ for part of the day, but not for the whole day. Therefore, more specific

conclusions about amount of exposure during peak periods vs. non-peak periods become problematic.

In an attempt to reduce the ambiguities in a principled fashion and to increase the confidence in any conclusions drawn from the data already collected, two efforts are actively being pursued. The first is to collect a comprehensive set of hypotheses that could explain the data patterns we observe. A controlled experiment will be run with a small number of dosimeters to recreate these potential causes and compare them to the actual data collected from the dosimeters. The result of this experiment should give us a basis for classifying the data collected from patients in 2012 and 2013.

The second effort to strengthen the conclusions drawn is to use a bracketed approach to the analysis. In this approach, the data are first assessed in a liberal fashion where the patient is assumed to have been PA+ except in cases where the interpretation of PA- is beyond doubt. In parallel with this method the data are also assessed in a very conservative fashion and suspicious data, where the patient may have been PA-, is eliminated from consideration. Thus given a patient who was photoprotective-compliant, the liberal interpretation would tend to support an accurate interpretation while the conservative interpretation would under-report their compliance. In the case where a patient was not particularly photoprotective-compliant, the liberal scenario would over-report compliance and the conservative scenario would tend towards an accurate interpretation. Picking an interpretation in the middle of the bracket then avoids over-reporting or under-reporting photoprotective-compliance. A further refinement of this method is to weight the results of the two interpretations by the patient's own self-reported photoprotection-compliance.

## Bayesian Network Models

The effort to answer the main photoprotective behavioral questions will require the use of multi-level model approaches. Given the variety of measurements and the inclusion of age, gender, and education attributes of the patients, further interesting insights may be mined from the data using graphical models. In particular, a Bayesian Network which has the property of introducing causal relationships between random variables may yield new insights or reinforce existing conclusions.

A Bayesian Network (BN) is a graphical model constructed as a directed acyclical graph (DAG). Each node in the graph represents a random variable with a probability distribution (or conditional probability distribution if it is the child of one or more parents) and each edge defines a conditional relationship between a child and parent node. The model itself is a factorization of a joint probability distribution across the random variables. Such a factorization encodes variable independence and conditional independencies. From such a graph, different forms of inference can be applied regarding the probable values a variable could have given knowledge of the values of any other variables in the graph. In addition to inference, the conditional probability distributions associated with each random variable can often be learned from data.

As a simple example, consider a graph with three nodes: two nodes represent the independent throw of two die and one node represents the sum of the dice. With a sufficient number of examples, the probabilities associated with each node (or random variable) can be learned. Now if the outcome of one of the die is known, the probability distribution of the sum node is marginalized over that node and changes significantly. This is called 'causal reasoning'. If the outcome of the sum is instead known, then the probability distribution of the two die nodes changes accordingly. This is called 'inferential reasoning'. A third type of reasoning, called 'intercausal reasoning' requires a more complex example but allows us to infer the outcome of the parent of a particular node given some special conditions and the knowledge of a sibling node. An excellent reference on graphical models can be found in Koller and Friedman (2009).

In the case of the BRIGHT study, a BN is being constructed to apply some of the reasoning forms to understand how gender or age differences, for example, affect compliance. Elicited expert knowledge is being used to develop the structure of the graph (the nodes and their causal relationships). The data gathered in the study will be used to populate the conditional probability distributions using established training algorithms (either Maximum Likelihood Expectation or Bayesian Learning). The chief potential obstacle, common to all machine learning efforts, is the limited amount of data. Because this is such a common problem, various strategies exist to compensate for smaller data sets. In the case of a BN, some paths of intercausal reasoning can be sacrificed to strengthen causal and inferential reasoning on key relationships. Specifically in the BRIGHT BN, each signifiant variable (e.g. age, education, gender) which is causally linked to the behavior-change variable can be learned independently rather than simultaneously.

## Discussion

The BRIGHT project hopes to answer key questions about how behavior changes with knowledge of hereditary risk alone and with *p16+* risk. It will succeed in this effort through a combination of objective and improved subjective measurements using a novel control group. The difficulties to be overcome relate to the amount of data collected and the noise in the UV dosimetry data. Strategies have been identified to attempt to manage these difficulties. A possible addition to the BRIGHT project, a Bayesian Network, may supplement the knowledge gained from the variety of data collected.

## References

Aspinwall, L. G., Taber, J. M.,Leaf, S. L., Kohlmann, W., & Leachman, S. A. (2013). Melanoma genetic counseling and test reporting improve screening adherence among unaffected carriers 2 years later. Cancer Epidemiology, Biomarkers & Prevention, 22, 1687-1697. Electronic publication date, August 15, 2013.

Koller, D., Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. The MIT Press.