Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(517): 54–63. doi:10.1080/01621459.2016.1191499.

Defining Cancer Subtypes With Distinctive Etiologic Profiles: An Application to the Epidemiology of Melanoma

Audrey Mauguen^{1,*}, Emily C. Zabor¹, Nancy E. Thomas^{2,3}, Marianne Berwick³, Venkatraman E. Seshan¹, and Colin B. Begg¹

¹ Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY

² Department of Dermatology, University of North Carolina, Chapel Hill, NC

³ Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC

⁴ Department of Internal Medicine, University of New Mexico Cancer Center, University of New Mexico, Albuquerque, NM

Abstract

We showcase a novel analytic strategy to identify sub-types of cancer that possess distinctive causal factors, i.e. sub-types that are “etiologically” distinct. The method involves the integrated analysis of two types of study design: an incident series of cases with double primary cancers with detailed information on tumor characteristics that can be used to define the sub-types; a case-series of incident cases with information on known risk factors that can be used to investigate the specific risk factors that distinguish the sub-types. The methods are applied to a rich melanoma dataset with detailed information on pathologic tumor factors, and comprehensive information on known genetic and environmental risk factors for melanoma. Identification of the optimal sub-typing solution is accomplished using a novel clustering analysis that seeks to maximize a measure that characterizes the distinctiveness of the distributions of risk factors across the sub-types and that is a function of the correlations of tumor factors in the case-specific tumor pairs. This analysis is challenged by the presence of extensive missing data. If successful, studies of this nature offer the opportunity for efficient study design to identify unknown risk factors whose effects are concentrated in defined sub-types.

Keywords

case-control study; etiologic heterogeneity; k-means clustering; logistic regression; multiple imputation

1. INTRODUCTION

Epidemiologic studies investigating risk factors for cancer have traditionally been based on the anatomic site of origin of the tumor. Over several decades large numbers of studies have

*Correspondence to: Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, 2nd floor, New York, NY 10017.

been conducted in breast cancer, lung cancer and so forth. However, in recent years technical advances, in particular at the molecular level, have led to an increasing interest in examining whether a disease such as breast cancer is a unitary entity with regard to risk factors, or whether it really consists of a set of diseases with distinctive risk factors. The term etiologic heterogeneity is used widely to describe this phenomenon. As an example, there is now an extensive literature reporting on the distinctive risk profiles of breast cancer cases characterized by the presence or absence of estrogen receptors (Althuis et al. 2004). For cutaneous melanoma, a malignant cancer of the skin, the topic of our detailed analysis in this article, evidence has emerged that the effects of selected risk factors differ between tumors characterized by the presence of mutations in the *BRAF* gene versus the *NRAS* gene (Thomas et al. 2007, 2010). As knowledge of the mutational profiles of tumors emerges, research will increasingly focus on determining whether the factors that influence the risk of cancer do so selectively for sub-types of cancer characterized by somatic mutations such as these. This topic of investigation is important, not only because identification of genuinely distinct sub-types has the potential to advance our understanding of cancer biology, but also because the search for new cancer risk factors can profitably take advantage of knowledge that sub-types have distinct risk factors. The statistical power of tests for detecting a risk factor that influences an individual sub-type is greater in most configurations than tests that group all cases together as in conventional epidemiologic analyses (Begg and Zabor 2012; Yu et al. 2015).

It is widely recognized that the era of genomic medicine will usher in the routine somatic profiling of tumors on clinical samples and that these profiles will identify large numbers of somatic mutations. Thus investigation of etiologic heterogeneity will become impractical on a factor by factor basis, especially as truly distinctive sub-types may be represented by combinations of groups of factors that define the tumors. To address this issue we have sought to create an “agnostic” strategy that is capable of investigating multiple tumor characteristics simultaneously using a novel clustering approach. This seeks to identify etiologically heterogeneous sub-types, i.e. groups of tumors that are similar in their characteristics and mutational profiles and that possess distinctive risk factors from other sub-types.

The analytic strategy for investigating etiologic heterogeneity stems from work by Dubin and Pasternak (1986) who demonstrated the use of polytomous logistic regression in case-control studies, the goal being to investigate whether the relative risks for individual risk factors differ by sub-types. Later it was shown that tests for etiologic heterogeneity can be accomplished by comparing the risk factors of sub-types of cases without the need for controls (Begg and Zhang 1994). Recent methodological research on this topic has focused on taking advantage of follow-up data in cohort studies. Chatterjee (2004) proposed a hybrid approach in which distinct sub-types characterized by a set of tumor factors are modeled based on polytomous logistic regression for each distinct sub-type while the odds ratio parameters of this model are related by a loglinear model. Rosner et al. (2013) have used a similar modeling strategy in an analysis of breast cancer. Wang et al. (2015) approach the problem by developing a computationally efficient two-stage framework. Xue et al. (2013) have compared the properties of polytomous logistic regression and proportional hazards regression in the context of examining pre-specified sub-types.

Our strategy is substantively different from these other approaches both in terms of study design and statistical analysis by building upon a key insight. In previous work we have shown that by studying the associations of tumor characteristics in individuals with independently occurring double primary cancers we can investigate the inherent etiologic heterogeneity of sub-types, i.e. the true heterogeneity influenced by all risk factors, not simply known risk factors (Begg 2011). This permits a strategy whereby the cancers can be optimally sub-typed using a study of double primary malignancies, and then the risk factors explaining the heterogeneity can be investigated independently in a separate study based on analysis of single primary cancers from a conventional case-control study.

In this article we present in detail the methodological strategy for a comprehensive analysis of etiologic heterogeneity of this nature. We analyze data from a large epidemiologic study of melanoma using these new methods, a study with an unusually large collection of patients with double primaries. The novelty of our work lies in the comprehensive conceptual analytic framework employed. However, specific novel challenges arise with regard to the clustering strategy to identify the appropriate tumor sub-types, and the unusual challenges involving multiple imputation in this context. We note that our group conducted an earlier pilot study of this two-stage strategy in breast cancer (Begg et al. 2015). However, that investigation had only 47 double primary cases available, too few for the elaborate multivariate clustering strategy used in the present study. In Section 2 we describe the study and clarify the goals of the analysis. In Section 3 we describe the conceptual model for the analysis at a technical level. Section 4 presents some unusual computational challenges due to the merging of missing data imputation with clustering, and the challenges of interpreting the results. The results of the analyses are presented in Section 5.

2. MOTIVATING DATASET

We used data from the Genes, Environment and Melanoma (GEM) Study, an international case-control study of multiple primary melanoma. Details of the study design were provided in an earlier article (Begg et al. 2006). The GEM study is especially insightful for our purposes since it includes a large collection of patients diagnosed with double primary melanoma. These were individuals diagnosed with a second or higher order primary melanoma during the period 2000-2003 in one of several population based registries in the USA, Canada, Italy and Australia. The comparison group was patients diagnosed with a single first primary melanoma in the same registries during the year 2000. Detailed information on all known melanoma risk factors was collected by questionnaire, including information on lifetime sun exposure, phenotypic characteristics and family history of melanoma. Germ-line mutations in a number of genes known or suspected to influence melanoma risk, including the important genes *CDKN2A* and *MC1R*, were investigated. Tumor samples were obtained for a subset of these study participants and were tested for somatic mutations in the pivotal genes *BRAF* and *NRAS*. There is emerging evidence that mutations in these genes identify sub-types in which the effects of some risk factors are different. For example Bauer et al. (2011) showed that melanomas with *BRAF* mutations tend to occur in younger people with less evidence of sun exposure, while it has also been observed that *BRAF*-associated melanomas occur at younger ages than *NRAS*-associated melanomas and that higher early-life UV exposure and number of nevi are also more

strongly associated with *BRAF* tumors (Thomas et al. 2007; Wu et al. 2014). Evaluation of the relevance of *BRAF/NRAS* mutations to etiologic heterogeneity is an important focus of our analysis, but it is hampered by the fact that tumor samples were genotyped from a relatively small subset of the GEM participants. Imputation of the missing mutations is crucial to allow us to take full advantage of the dataset.

3. STATISTICAL FRAMEWORK

The analytic strategy involves two distinct, structurally independent stages. In Stage 1 we use the patients with double primary melanomas to identify the set of tumor sub-types with the clearest evidence of etiologic heterogeneity using a novel clustering strategy. In Stage 2 we perform more conventional logistic regression analyses to correlate these sub-types with known risk factors to determine which risk factors explain the etiologic heterogeneity. This analysis involves patients with a single primary melanoma. By using double primary and single primary cases separately in this way we ensure that the two stages are both structurally and statistically independent.

3.1 Stage 1: Defining Cancer Sub-Types Using Patients with Double Primaries

Each double primary case is characterized by two sets $\{Y_{ijk}\}$ of (paired) tumor characteristics where $i = 1, \dots, n$ indexes the case, $j = 1, 2$ indexes the tumor, and $k = 1, \dots, q$ indexes the tumor characteristic. These are used to identify the mutually exclusive set of tumor sub-types that exhibit the greatest degree of etiologic heterogeneity.

Our work builds upon a key concept explained in an earlier article (Begg 2011). The concept is that any tumor characteristic that distinguishes sub-types with different distributions of risk factors will be correlated when it is cross-classified in pairs of primary tumors from cases with double primaries. Moreover the stronger the “heterogeneity” of these risk distributions, the stronger the resulting correlation in the cross-classifications. Details of the theory supporting this concept and its simplifying assumptions are reproduced in the Appendix. For the purposes of our analytic strategy the key implication of this result is that we can take any two sub-types, A and B, and calculate the cross-classification odds ratio,

denoted $\hat{\psi}_{AB}$. This is an estimate of the strength of the etiologic heterogeneity between these two sub-types which can be used to compare different sub-typing solutions and to determine which ones appear to be most strongly distinct with regard to etiology. In order to take into account the prevalences of the sub-types, we have used as the distinguishing measure

$\hat{\pi}_A \hat{\pi}_B \log(\hat{\psi}_{AB})$, where $\hat{\pi}_A$ and $\hat{\pi}_B$ represent the prevalences of the sub-types in the data. In this way we give less weight to sub-typing solutions with infrequent sub-types. When we are considering S mutually exclusive sub-types the overall measure of etiologic heterogeneity becomes:-

$$H = \sum_{s=1}^S \sum_{w>s}^S \hat{\pi}_s \hat{\pi}_w \log(\hat{\psi}_{sw}).$$

We emphasize that this measure is a function solely of the tumor factors, yet it optimizes the etiologic heterogeneity influenced by all risk factors, both known and unknown. As a result, even though we have information on known risk factors for double primary cases we do not use these risk factors to either define the sub-types in our Stage 1 analysis or to assist in multiple imputation for missing data (Section 4.1).

Our clustering strategy seeks to identify a parsimonious set of sub-types, defined by the tumor characteristics, that maximizes H . Since each case has 2 tumors we must define how each of these $2n$ tumors is allocated into its sub-type. We emphasize that the clustering involves all $2n$ tumors but the H measure of heterogeneity that we seek to maximize captures information on the co-occurrences of the sub-types in the pairs of tumors from the n patients. However, we cannot conduct an unconstrained maximization for the following reasons. First, even when the number of characteristics is modest and discrete as in our study (see Section 4.1), the number of potential permutations of sub-types is vast and so it is computationally infeasible to evaluate all of them. Second and most importantly, if there are many tumor characteristics then inevitably each tumor will have a unique combination. In these circumstances the set of all possible sub-typing solutions is invariant to the pairings and thus cannot meaningfully extract the crucial information from the pairings. For these reasons we must constrain the maximization in some principled fashion. Our strategy is based on the principle that solutions in which there is little evidence of geometrical distinctiveness of the sub-types are less plausible than those in which there is clear evidence of topographical separation. For these reasons we use k-means clustering as a convenient, computationally feasible strategy to identify candidate solutions that demonstrate clear topographical separation of the sub-types. k-means clustering (Hartigan 1979) is a well-known method that classifies multivariate data into a predefined number of homogeneous groups based on minimizing the within-cluster Euclidean distance relative to the between cluster differences. It is an iterative process that usually identifies various local minima. This strategy is made more complex because of the extensive missing data, and we outline in detail the computational strategy in Section 4. An analogous strategy was used in an earlier study of breast cancer (Begg et al. 2013). However, it did not involve double primaries and did not require missing data imputation, features that change substantially the challenges involved and the strategies used to address them.

3.2 Stage 2: Defining the Risk Factors Differentially Affecting Sub-Types

Each single primary case is characterized by a set of tumors characteristics $\{Y_{ijk}\}$, $i = n + 1, \dots, n + m$. As before, $k = 1, \dots, q$ indexes the tumor characteristics. Note that in contrast to the double primary cases used in Stage 1 the tumor indicator j is always equal to 1. The tumor characteristics are used to classify each tumor into one of the S sub-types, based on the optimal solution defined in Stage 1. In addition, single primary cases are each characterized by a set of risk factors $\{X_{il}\}$, where $l = 1, \dots, r$ indexes the risk factor. The data are used to determine which risk factors have distinctive distributions across the different sub-types. This allows us to identify the tumor factors that are most influential in defining the sub-types.

To identify the individual risk factors that have distinctive influences on the sub-types we perform conventional logistic regression analyses. We used case-only logistic regressions comparing each sub-type to a reference sub-type. This permits general statistical testing to determine if the distribution of any individual risk factor is heterogeneous across the sub-types after adjusting for other risk factors. For these analyses we used individual pair-wise logistic regressions, rather than a single polytomous logistic regression. These two strategies typically produce almost identical results (Begg and Gray, 1984).

4. TECHNICAL CHALLENGES

The GEM study has detailed information on all known, observable melanoma risk factors in addition to DNA for the evaluation of genetic risk factors. Efforts were made to collect tumor tissues from a subset of participants and these were analyzed for somatic mutations in *BRAF* and *NRAS* (Thomas et al. 2015). Also, various ancillary studies were accomplished involving germ-line genotyping of single-nucleotide polymorphisms in candidate genes (Gibbs et al. 2015a). Our current analysis endeavors to present a global evaluation of etiologic heterogeneity and as such we are interested in using information on all tumor factors and all known risk factors. However, because of the selective collection of tumor tissues, the occurrence of laboratory failures and for various other reasons there are considerable quantities of missing data. In particular, for the 903 double primary cases used in Stage 1 of our analysis, genotyping of both tumors for *BRAF* and *NRAS* mutations was only completed for 116 tumor pairs (see Table 1). Consequently, missing data imputation is essential for a meaningful analysis of all factors, and this leads to analytic complexities. In this Section we provide details of how we addressed these and other challenges.

4.1 Stage 1 Analyses

The analysis was based on 6 selected tumor characteristics: anatomic site of the melanoma; mutation status; presence of solar elastosis; presence of neval remnants; histology; presence of infiltrating lymphocytes. The extent of missing data for each of these is described in Table 1. We see that anatomic site of occurrence of the tumor(s) is the only factor available for both tumors for all 903 cases. *BRAF/NRAS* mutational status is available for 138 and 348 of the first and second primaries, respectively, and is available for both members of the pair for 116 cases. The other four tumor factors have intermediate amounts of missing data. We assumed that the missing data are missing at random, i.e. the missing status of a variable can be explained by the other measured factors. We imputed the missing data using a fully conditional specification, also known as Multiple Imputation with Chain Equation (van Buuren 2007). The principle is to use information from other variables iteratively when imputing missing values. As a starting value all missing values are imputed independently. Then all values that were originally missing are imputed one by one using a regression model from the imputed complete dataset. These updated values are then used to repeat the imputation process.

A key feature of our analytic strategy is the use of the heterogeneity measure H . This is crucially dependent on the correlations between tumor factors in the double primaries and so it is vital that our imputation strategy maintains the fundamental multiple correlation

framework in the data. To accomplish this, our imputation model conditioned on all tumor factors in both tumors for the case being imputed in the multiple imputation procedure. Thus, for example, to impute, Y_{i11} we used a conditional model of the form $P(Y_{i11} | Y_{i21}, (Y_{i12}, Y_{i22}), \dots, (Y_{i1q}, Y_{i2q}), \Theta)$, where Θ represents the parameters of the model. We created 25 imputed datasets, using 20 iterations per imputed dataset, as recommended by van Buuren (2011). The success of this strategy can be gauged from the results in Table 1 where we present for each tumor factor the cross-classification odds ratios and their corresponding marginal H values, first based on the available complete data for that factor and then based on the average odds ratio from the 25 imputed datasets. As can be seen in the table the complete data and imputed estimates of both the odds ratios and the corresponding H values are individually comparable for each tumor factor.

It is generally agreed that the missing at random assumption cannot be tested from the observed data (van Buuren 2011; Potthoff et al. 2006; Jaeger 2006) although imputations accommodating non-random missingness are possible based on pre-specified assumptions (van Buuren 2001). In essence we assume that there are no additional unobserved factors that influence selection beyond those used in the imputation model. In the GEM study we attempted to ascertain tumor tissue from all participants in the selected geographical regions that participated in the efforts to obtain tumor tissue. In general we failed to ascertain tissue from hospitals where the local staff were unwilling to make the effort to provide the samples, a selection mechanism that is likely to be effectively random. The most plausible challenge to this assumption would be that tumors were more likely to be ascertained if they were larger, since for small tumors there is less likely to be excess tumor tissue available for a research study of this nature. To address this we examined the Breslow depths (available from the pathology reports) of the ascertained tumors of the double primary cases (486 tumors, see Table 1) versus missing tumors (1320 tumors). The relative frequencies of Breslow depths, classified as <1mm, 1-2mm, 2-4mm and >4mm, are 70%, 18%, 9% and 3% for the former and 72%, 17%, 8% and 3% for the latter, demonstrating essentially no selection bias on the basis of tumor size.

The purpose of k-means clustering is to allocate the tumors into a pre-specified number of S clusters (sub-types) while minimizing the within-cluster sum of squares. Iteratively, each data point representing the set of tumor factors $(Y_{ij1}, \dots, Y_{ijq})$ for $i = 1, \dots, n$ and $j = 1, 2$ is allocated to the cluster with the cluster mean closest to $(Y_{ij1}, \dots, Y_{ijq})$. Once all the data points are allocated the cluster means are updated and the data points are allocated again using the new means. This iterative process is continued until the within-cluster sum of squares is minimized. This will often lead to a local minimum rather than a global solution. As a result, changing the initialization of the clustering can lead to different solutions. As explained in Section 3.1 our strategy is to calculate the etiologic heterogeneity optimization criterion H for local k-means solutions and choose the one that maximizes H . Since we used multiple imputation to generate 25 distinct datasets the clustering was performed separately on each of these distinct datasets. We generated 400 clustering solutions for each imputed dataset and thus 10,000 clusterings overall.

Any given solution is a mapping of each of the $U=144$ distinct combinations of the 6 tumor characteristics into one of S sub-types. [We note that the number of possible arrangements of

the 144 combinations into 4 sub-types is 2.1×10^{85} , and thus complete enumeration of all combinations is computationally infeasible.] Identical solutions were grouped into *candidate* aggregate solutions. Specifically, we identified candidate solutions that were observed to occur at least 10 times. We note that since the imputed datasets are different, the frequencies of the 144 combinations are different and it is very unlikely to find exactly the same solution in different imputed datasets. We thus need to reconcile similar solutions obtained from different datasets. For each candidate solution, *similar* solutions were identified on the basis of a measure of concordance. The concordance, C , was defined as follows. Let the mapping of each of the 144 distinct combinations of tumor characteristics to the candidate solution be defined by the vector of sub-type indicators $(s_1, s_2, \dots, s_U)$ and let the numbers of tumors allocated to each combination in the imputed dataset from which the solution was generated be $(t_1, t_2, \dots, t_U)$, where $\sum_{u=1}^U t_u = 2n$. Consider any other solution defined by $(s_1^*, s_2^*, \dots, s_U^*)$ with corresponding frequencies $(t_1^*, t_2^*, \dots, t_U^*)$. Note that $(t_1, t_2, \dots, t_U) \neq (t_1^*, t_2^*, \dots, t_U^*)$ if the solutions arose from different imputed datasets. Then the similarity of the two solutions can be characterized by the proportion of tumors allocated to concordant sub-types, defined by

$$C = \sum_{u=1}^U (t_u + t_u^*) I(s_u = s_u^*) / 4n.$$

All solutions that possessed a concordance $C \geq 0.95$ with the candidate solution were included in the candidate group of similar solutions. Using all the similar solutions in the candidate group, we then redefined the candidate solution by a voting method in which each combination of factors was assigned to the sub-type that occurred most frequently among the similar solutions. The H measure of the resulting solution corresponds to the mean of the H measures among all the similar solutions in the group. The optimal solution is the redefined candidate solution with the highest H value. The purpose of this redefining procedure is to ensure that the final solution is representative of the fundamental dataset and is not overly influenced by a single unusual dataset among all the datasets created by the multiple imputation procedure. The entire process is displayed schematically in Figure 1.

To determine the number of sub-types, we conducted analyses with two sub-types, three sub-types, four sub-types, five sub-types and six sub-types. We then compared the increments in the optimal H values. We stopped when the increment in H was not large enough to justify the addition of another cluster.

4.2 Stage 2 Analyses

This stage of the analysis was based solely on study participants with a single primary melanoma, and involved first allocating each tumor to one of the optimal sub-types identified by the clustering analysis described in Section 4.1 above. Logistic regressions were then used to identify individual risk factors whose distributions differed substantially across the sub-types in order to determine the features that explain the etiologic heterogeneity of the sub-types. Again there were substantial quantities of missing data for the tumor factors. For the risk factors the frequency of missing data ranged from no missing

data for age, gender and center to 8% missing for sun exposure. We again employed multiple imputation with 25 imputed datasets using the multiple imputation with chain equation procedure. Final estimates of the regression coefficients and their standard-errors were obtained by combining the estimates from the 25 datasets using Rubin's approach (Rubin 1996). The final p-values were obtained from combining likelihood ratio statistics from the 25 datasets (Marshall 2009).

5. ANALYSIS OF MELANOMA DATA

We analyzed a total of 903 double primary cases in Stage 1 and 2103 single primary cases in Stage 2. For 14% of the double primaries the two tumors were diagnosed simultaneously, and overall in 81% of these cases the tumors were diagnosed within 10 years of each other, an interval of time used conventionally by epidemiologists as a period in which the incidence rate is approximately constant. We note that, for synchronous cases, the labeling of one tumor as the 1st primary is arbitrary and that our heterogeneity criterion H is not invariant to this selection. The labeling was accomplished by the data managers who recorded data for the GEM Study.

5.1 Stage 1: Identification of Optimal Sub-Types

We elected to limit our analysis to 6 tumor characteristics (from a total of 13). These were selected based either on previous evidence from the literature that the tumor factor is convincingly related to risk factors or because the unadjusted cross-classification odds ratios indicated empirically that the factor was likely to be influential in our clustering analysis. Table 1 provides these marginal H values for each factor, showing that anatomic site appears to be the factor that most strongly influences etiologic heterogeneity, with solar elastosis also exhibiting notably more heterogeneity than the other tumor factors. We performed analyses with 2, 3, 4, and 5 sub-types. The optimal H values obtained were, respectively, 0.29, 0.46, 0.58 and 0.62. We also attempted an analysis with 6 sub-types but only one solution met the criteria outlined in Section 4.1.2 and its H value did not improve on the 5 sub-type solution ($H=0.57$). On this basis we chose to investigate further the results with four sub-types, as the addition of the fifth sub-type increased the H value only modestly.

Results of the clustering analyses are summarized in Table 2. The top three solutions are presented. For each solution the allocations of the 144 distinct combinations to the 4 sub-types are displayed in Supplementary Table S1. This shows that these allocations almost perfectly correspond to the simple verbal descriptors provided in Table 2. Thus for Solution #1 the allocations match almost exactly the descriptions in the table:- A: Legs; B: Trunk/ Arms with Solar Elastosis; C: Trunk/Arms without Solar Elastosis; D: Head/Neck. The only exception is the specific combination of head/neck tumors with no solar elastosis, neval remnants, *BRAF* mutations, TILs and "other" histology, which is classified into sub-type C instead of sub-type D. Note also from Supplementary Table S1 that some combinations of factors were never observed, either in the complete dataset or any of the imputed datasets, and so these have been assumed to correspond to the sub-types defined by the verbal descriptions. The second best solution had an H value of 0.56, only slightly lower than the 0.58 of the optimal solution. This solution also classifies melanomas of the head/neck and

melanomas of the leg into distinct sub-types, but the remaining site (trunk/arms) is classified into two sub-types on the basis of neval remnants rather than solar elastosis (Table 2 and Supplementary Table S1). The 3rd solution involves mutational status but has a considerably lower value of the heterogeneity measure H .

The H values for all of the individual k-means solutions that met our criterion of occurring in at least 10 clustering runs are plotted against the k-means distance measure in Figure 2. All similar solutions that contributed to the top aggregate solution are plotted in red while the second and third solutions are plotted in green and blue, respectively. Interestingly the top solution (red) appears to comprise a set of sub-types that have lower separation in terms of the k-means distance metric on the horizontal axis than the other top two solutions (in green and blue). This is probably due to the fact that the top solution involves solar elastosis, a tumor factor that is more strongly correlated with the dominant factor (anatomic site) than either neval remnants or mutational status, the factors that drive the clustering for the second and third solutions. The orange dots correspond to the remaining k-means local maxima that were not merged with one of the top three candidate solutions. Figure 2 also displays (in grey) 50000 solutions that were generated by randomly allocating factor combinations to one of 4 sub-types. This shows the range of degrees of topographical separation expected on the horizontal axis and the range of values of the H measure on the vertical axis that are likely to occur simply by chance. The graph shows clearly that the top three aggregated solutions (red, green, blue) display measures of heterogeneity that exceed values that could have occurred by chance, although the remaining solutions (yellow) are less clearly stochastically higher than the random solutions on this axis. Regarding the reproducibility of the chosen solution (red) it is clear from the variations in H within each solution that there is little to choose between this solution and the second best (green) solution. Finally the black dots represent all of the individual k-means solutions where the corresponding H values are computed with the tumors randomly paired, i.e. such that there is no heterogeneity signal. The grey dots are stochastically larger than the black dots because in the grey dot solutions the pairings are retained.

5.2 Stage 2: Relation of Risk Factors to Sub-Types

Results of the logistic regressions relating the classical risk factors for melanoma to the sub-types are presented in Table 3. Sub-type A (melanoma of the leg) is the reference category in this presentation. The results show that age and gender are very strongly and highly significantly related to these sub-types and that mole count and sun exposure also appear to be significantly influential. In particular, there is a very strong female predominance in melanomas of the leg, melanomas of the trunk/arms with solar elastosis and melanomas of the head/neck occur in a significantly older population and are related to sun exposure, and melanomas of the trunk/arms, especially those without solar elastosis, occur in people with a greater number of moles. We also conducted exploratory analyses of the many single-nucleotide polymorphisms (SNPs) in candidate genes that have been evaluated in the GEM study. Results for an analysis involving 50 of these SNPs in 25 candidate genes in addition to indicators for the presence of germ-line mutations in *CDKN2A* and *MC1R* are provided in Supplementary Table S2. The results identify a strong positive association of the *IRF4*

rs12203592 variant allele with sub-types B (trunk/arms with solar elastosis) and D (head/neck), though the p-value does not survive a correction for multiple testing.

6. DISCUSSION

We have outlined a prototype staged analysis of etiologic heterogeneity, involving a novel clustering analysis of double primary cancers to identify the set of tumor sub-types with the strongest evidence for mutual etiologic heterogeneity, followed by a case-only analysis of individuals with single primary melanoma to examine which known risk factors appear to segregate the sub-types. This strategy builds on various conceptual premises. The first is that examining the association of tumor sub-types in independent double primary cancers provides unique insight into the inherent etiologic heterogeneity of the sub-types, since the patterns of occurrence of double primaries are influenced by all risk factors, not just known risk factors (Begg 2011). The second premise is that we can effectively identify the etiologically heterogeneous sub-types using clustering techniques. Finally we make use of the fact that information about the influence of individual risk factors on etiologic heterogeneity is contained solely in incident cases, and thus population controls are not required to identify risk factors that exert different effects on different disease sub-types. Healthy population controls would be required, however, for estimating the relative risks of individual risk factors on specific sub-types.

This kind of strategy presents several challenges and limitations. An obvious challenge is the need for large samples of double primaries that are annotated with the key tumor characteristics. To be fully effective this will require tumor samples from both primaries that are thoroughly evaluated for somatic tumor characteristics. This kind of resource is generally not available at present but will become available gradually as clinical molecular profiling of tumors becomes more commonplace. Our melanoma dataset is unusual in having a relatively large number of double primaries, but testing for somatic mutations was available for only the small fraction for which tissue samples from both tumors were available, necessitating multiple imputation to take full advantage of the resource. Furthermore, the tumor profiling was limited to two important genes, *BRAF* and *NRAS*.

Our analyses of the GEM study produced some clear findings. First, it demonstrated that anatomical site of occurrence of melanomas is the strongest known tumor characteristic that distinguishes etiology. Other factors such as *BRAF/NRAS* mutational status, presence of solar elastosis and presence of neval remnants also appear to be relevant axes that define etiologic heterogeneity, but these factors are notably less important than anatomic site. To some extent our findings reflect results already well known to experts in melanoma epidemiology, such as the fact that melanomas of the leg tend to occur preferentially in young women, melanomas of the head and neck tend to occur in older men, and there are preponderances of *BRAF* and *NRAS* mutations in melanomas of the trunk (Cho et al. 2005, Neilson et al. 2012, Thomas et al. 2015). Nonetheless, it is reassuring that our novel strategy of deciphering heterogeneity from double primaries identifies trends that are already clearly known, providing support for the validity of this approach. The analysis also identified the *IRF4* SNP rs12203592 as being strongly related both to melanomas of the head/neck and melanomas of the trunk or arms with solar elastosis. This result was identified in an earlier

report using this dataset in which melanomas characterized by solar elastosis were contrasted with melanomas characterized by neval remnants (Gibbs et al. 2015b). It is reassuring that our agnostic strategy found sub-types that reflect this important genetic effect.

Our decision to use k-means clustering solutions as a strategy for identifying candidate solutions is clearly arbitrary. As explained in Section 3.1 it is essential to constrain the maximization of H in some principled way, and we believe that use of a distance measure is a rational way to identify candidate solutions since this leads to solutions where the sub-types are plausibly biologically distinct. However, there is no particular reason to use Euclidean distance rather than other “distance” criteria. Finding a more satisfactory strategy for maximizing H is a topic for future research.

Clustering analyses are notorious for producing clusters that are not reproducible, i.e. where an entirely different set of clusters emerges from a study of a replicate dataset, or where the clustering solution is highly dependent on the tumors included in the analysis. This is especially problematic when the clustering is performed on large numbers of features, as in studies using genomic micro-arrays, for example. Our dataset consisted of 12 tumor features and we elected to restrict the analysis to the 6 that demonstrated strong marginal evidence of etiologic heterogeneity at the outset. The solution with the largest heterogeneity measure is essentially driven by the two tumor characteristics with the largest marginal effects on heterogeneity, anatomic site and solar elastosis. However, it is clear from Figure 2 that the heterogeneity (y-axis) in this solution (red dots) is barely any different from the solution that replaces solar elastosis with neval remnants (green dots) and not so convincingly better than the solution that replaces solar elastosis with mutational status (blue dots). Since solar elastosis, neval remnants and mutational status are all very biologically different and largely uncorrelated entities, the analysis is far from conclusive in its selection of the optimal sub-types. Indeed the distinctiveness of these three candidate factors suggests that there may really be more than four etiologically distinct sub-types and that our dataset is simply not large enough to identify the more refined sub-types. Clustering analyses are inherently descriptive, and they do not possess the reproducibility properties of conventional statistical methods. Consequently we feel we would be overstating the strength of the evidence if we were to suggest that melanoma etiologic heterogeneity is either characterized by anatomic site and solar elastosis alone, or that it is restricted to 4 sub-types. However, we do feel that the analysis provides strong evidence for the following conclusions: anatomic site is the strongest known driver of etiologic heterogeneity; tumors of the trunk and arms comprise more than one sub-type with distinctive etiologic pathways; risk factors that strongly distinguish etiologically distinct sub-types include age, gender, and mole count.

Our approach to evaluating etiologic heterogeneity is novel both with regard to study design, i.e. the central use of cases with double primaries, and with regard to statistical analysis. The methodology requires further research. For example, we seek to allocate tumors into mutually exclusive sub-types, but determining the correct number of sub-types is a difficult challenge, as indeed it is for clustering methods more generally. A popular strategy is to use the gap statistic (Tibshirani et al. 2001). However this method and more traditional options such as the one proposed by Calinski and Harabasz (1974) are based on increments in the

probability distributions of sums of squares distance measures when you add an extra cluster. Our clustering strategy differs radically from this framework in that it involves maximizing the H measure among candidate k-means solutions, a measure of pair-wise correlation not distance. Creation of a formal strategy for testing whether a new cluster adds significantly to the H measure is a topic for future research. We elected to simply use judgment on the basis of the trend in optimal H values to pick the most appropriate number of sub-types, though we recognize that there may truly be more than 4 sub-types. Furthermore, the sub-types may overlap in ways that are hard to meaningfully disentangle through the use of a construct that demands mutual exclusivity. Others who have proposed methodology for this problem have approached it by modeling the tumor factors in a regression model rather than seeking mutually exclusive sub-types (Chatterjee 2004; Wang et al. 2015; Rosner et al. 2013). We need to further understand the strengths and limitations of these contrasting strategies. In short, our presentation represents a suggested framework for approaching this important and challenging area of epidemiology rather than a definitive methodology.

In summary, we have demonstrated how to employ a mix of traditional statistical methods in an unusual study design to draw important inferences about the epidemiology of melanoma. Although the method requires a key assumption that cannot be considered to be strictly correct, the assumption that occurrences of two independent cancers in a single subject are experimental replicates, this assumption provides a conceptual framework for analysis that greatly simplifies the analytic strategy and the interpretation of the results. The major conclusion, that melanoma consists of (at least) four sub-types, primarily distinguished by the anatomic site of the primary melanoma, has strong face validity, in that there is ample evidence in the literature of the distinctive patterns of risk factors in these anatomic sub-types. Complex statistical analyses like this one are, in our opinion, appropriately tackled using a judicious blend of methodological rigor and pragmatic assumptions to simplify the mathematical framework for analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The research was supported by the National Cancer Institute, awards CA163251, CA008748, CA112243 and CA112524. We are also grateful to Kathleen Conway, Sharon Edmiston, Eloise Parrish, Audrey Alexander, and Honglin Hao who performed the *BRAF* and *NRAS* genotyping, to Lynn From and Klaus Busam who performed the pathology review, and to the GEM team of investigators for access to the data.

APPENDIX

Expected Sub-Type Concordances in Double Primary Malignancies

Our analysis is based on the premise that in patients with double primary melanoma the odds ratio from the cross-classification of the sub-types of the first and second primaries is directly related to the degree of etiologic heterogeneity of the sub-types. In the following we

summarize the technical justification that was previously published in the Statistical Appendix of Begg (2011).

Consider two sub-types, A and B. Consider the patients in the study with double primary melanoma where each tumor is either of sub-type A or sub-type B. Let the cross-classification frequencies be F_{AA}, F_{AB}, F_{BA} and F_{BB} with corresponding odds ratio estimate $\hat{\psi}_{AB} = F_{AA}F_{BB}/F_{AB}F_{BA}$. Thus $\psi_{AB} = E_{AA}E_{BB}/E_{AB}E_{BA}$, where $E_{AA} = E(F_{AA})$. Cases of melanoma recruited to a conventional population-based case-control study emerge from a population of people at risk of melanoma who are alive but have not previously experienced the disease. Typically the recruitment period will be a defined, relatively short period, say 1 to 3 years, as was the case for the GEM study. During this recruitment period every individual has a personal “risk” of being diagnosed with melanoma and thus eligible for the study. This risk is influenced by all known melanoma risk factors such as the person's age, gender, history of sun exposure, genetic predisposition, in addition to all unknown influences on melanoma risk. Let the risk be r_i for the i^{th} individual in the population. Further, let $r_i = r_{Ai} + r_{Bi}$, where r_{Ai} and r_{Bi} are the risks for sub-types A and B. Then the expected frequencies of double cancers are given by

$E_{AA} = \sum r_{Ai}^2$, $E_{BB} = \sum r_{Bi}^2$, $E_{AB} = E_{BA} = \sum r_{Ai}r_{Bi}$. It follows that:-

$$\psi = \frac{\sum r_{Ai}^2 \sum r_{Bi}^2}{(\sum r_{Ai}r_{Bi})^2} = \frac{(1+K_A^2)(1+K_B^2)}{(1+K_{AB})^2}, \quad (\text{A.1})$$

where K_A and K_B are the coefficients of variation of the population distribution of risks of type A and type B cancers, respectively, and K_{AB} is the coefficient of covariation, i.e.

$K_{AB} = \mu_A^{-1} \mu_B^{-1} \left(\sum r_{Ai}r_{Bi} - \mu_A\mu_B \right)$, where μ_A and μ_B are the mean risks of the sub-types. This standardized covariance of the risk profiles represents the degree of etiologic heterogeneity. Highly correlated risk distributions of type A and type B cancers in the population implies that the two sub-types have very similar etiologies, while low or negative correlation implies distinctive risk distributions, i.e. distributions that represent etiological heterogeneity. From (A.1) we see that the cross-classification odds ratio of double primaries is inversely related to K_{AB} , i.e. the *lower* the covariation in the risk profiles the *higher* the odds ratio.

Clearly the preceding is an idealized representation of the sampling of double primaries. Suppose that the i^{th} subject in the population experienced two type-A tumors. The theory assumes that the independent risk of each of the tumors is r_{Ai} . It is based on the premise that the risk is identical for each tumor. However, if the second tumor occurred a number of years later than the first tumor and risk increases with age then the risk affecting the second tumor must be higher. Likewise, additional environmental exposures in the intervening period, or even treatment for the first primary, could impact risk. Also, the sampling scheme in GEM did not mandate that both tumors in the pair occurred in the case accrual study period. GEM cases were eligible if a second primary melanoma occurred in the accrual period (2000-2003) regardless of when the previous tumor was diagnosed. As a consequence the

result in (A.1) is approximate, and serves as a heuristic justification for using the cross-classification odds ratio as a metric for distinguishing or ranking different sub-typing options on the basis of etiologic heterogeneity.

The result in (A.1) justifies the use of $\hat{\psi}$ as a measure of etiologic heterogeneity. However, on its own it does not account for the prevalences of the two sub-types. From a practical perspective we seek sub-types that occur frequently rather than rarely. Thus in our clustering algorithm we have elected to weigh the measure by $\hat{\pi}_A \hat{\pi}_B$ to down-weight the chance of selecting sub-types that occur very rarely. Thus, for a general set of S sub-types we use the

$$\text{measure:- } H = \sum_{s=1}^S \sum_{w>s}^S \hat{\pi}_s \hat{\pi}_w \log(\hat{\psi}_{sw}).$$

REFERENCES

- Althuis MD, Fergenbaum JH, Garcia-Closas M, Brinton LA, Madigan MP, Sherman ME. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer Epidemiology, Biomarkers and Prevention*. 2004; 13:1558–1568.
- Bauer J, Büttner P, Murali R, Okamoto I, Kolaitis NA, Landi MT, Scolyer RA, Bastian BC. BRAF mutations in cutaneous melanoma are independently associated with age, anatomic site of the primary tumor, and the degree of solar elastosis at the primary tumor site. *Pigment Cell and Melanoma Research*. 2011; 24(2):345–51. [PubMed: 21324100]
- Begg CB, Gray RJ. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 1984. 1984; 71:11–18.
- Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiology, Biomarkers and Prevention*. 1994; 3:173–175.
- Begg CB, Hummer AJ, Mujumdar U, Armstrong BK, Kricker A, Marrett LD, Millikan RC, Gruber SB, Culver HA, Zanetti R, Gallagher RP, Dwyer T, Rebbeck TR, Busam K, From L, Berwick M. A design for cancer case-control studies using only incident cases: experience with the GEM study of melanoma. *International Journal of Epidemiology*. 2006; 35:756–764. [PubMed: 16556646]
- Begg CB. A strategy for distinguishing optimal cancer subtypes. *International Journal of Cancer*. 2011; 129:931–937. [PubMed: 20949563]
- Begg CB, Zabor EC. Detecting and exploiting etiologic heterogeneity in epidemiologic studies. *American Journal of Epidemiology*. 2012; 176:512–518. [PubMed: 22922440]
- Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Statistics in Medicine*. 2013; 32:5039–5052. [PubMed: 23857589]
- Begg CB, Orlov I, Zabor EC, Arora A, Sharma A, Seshan VE, Bernstein JL. Identifying etiologically distinct sub-types of cancer: a demonstration project involving breast cancer. *Cancer Medicine*. 2015; 4:1432–1439. [PubMed: 25974664]
- Calinski T, Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*. 1974; 3:1–27.
- Chatterjee N. A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*. 2004; 99:127–138.
- Cho E, Rosner BA, Colditz GA. Risk factors for melanoma by body site. *Cancer Epidemiology, Biomarkers and Prevention*. 2005; 14:1241–1244.
- Dubin N, Pasternak BS. Risk assessment for case-control subgroups by polychotomous logistic regression. *American Journal of Epidemiology*. 1986; 123:1101–1117. [PubMed: 3706280]
- Hartigan JA, Wong MA. A K-means clustering algorithm. *Applied Statistics*. 1979; 28:100–108.
- Gibbs DC, Orlov I, Kanetsky PA, Luo L, Kricker A, Armstrong BK, Anton-Culver H, Gruber SB, Marrett LD, Gallagher RP, Zanetti R, Rosso S, Dwyer T, Sharma A, La Pilla E, From L, Busam KJ, Cust AE, Ollila DW, Begg CB, Berwick M, Thomas NE, on behalf of the GEM Study Group.

- Inherited genetic variants associated with occurrence of multiple primary melanoma. *Cancer Epidemiology, Biomarkers and Prevention*. 2015a; 24:992–7.
- Gibbs DC, Orlov I, Bramson JI, Kanetsky PA, Luo L, Krickler A, Armstrong BK, Anton-Culver H, Gruber SB, Marrett LD, Gallagher RP, Zanetti R, Rosso S, Dwyer T, Sharma A, La Pilla E, From L, Busam KJ, Cust AE, Ollila DW, Begg CB, Berwick M, Thomas NE, on behalf of the GEM Study Group. Association of interferon regulatory factor-4 (IRF4) polymorphism rs12203592 with divergent melanoma pathways. *Journal of the National Cancer Institute*. 2015b in press.
- Jaeger, M. On testing the missing at random assumption.. In: Fürnkranz, J.Scheffer, T., Spiliopoulou, M., editors. *Machine Learning: ECML 2006. Lecture Notes in Computer Science*. Springer; Berlin Heidelberg: 2006. p. 671-78.
- Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modeling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology*. 2009; 9:57. [PubMed: 19638200]
- Nielsen K, Masback A, Olsson H, Ingvar C. A prospective, population-based study of 40,000 women regarding host factors, UV exposure and sunbed use in relation to risk and anatomic site of cutaneous melanoma. *International Journal of Cancer*. 2012; 131:706–715. [PubMed: 21898390]
- Potthoff RF, Tudor GE, Tudor KS, Pieper, Hasselblad Vic. Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*. Jun 1; 2006 15(3): 213–34. doi:10.1191/0962280206sm448oa. [PubMed: 16768297]
- Rosner B, Glynn RJ, Tamimi RM, Chen WY, Colditz GA, Willett WC, Hankinson SE. Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *American Journal of Epidemiology*. 2013; 178:296–308. [PubMed: 23645624]
- Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996; 91:473–489.
- Thomas NE, Edmiston SN, Alexander A, Millikan RC, Groben PA, Hao H, Tolbert D, Berwick M, Busam K, Begg CB, Mattingly D, Ollila DW, Tse CK, Hummer A, Lee-Taylor J, Conway K. Number of nevi and early-life ambient UV exposure are associated with BRAF-mutant melanoma. *Cancer Epidemiology, Biomarkers and Prevention*. 2007; 16(5):991–997.
- Thomas NE, Kanetsky PA, Edmiston SN, Alexander A, Begg CB, Groben PA, Hao H, Busam K, Ollila DW, Berwick M, Conway K. Relationship between germline MC1R variants and BRAF-mutant melanoma in a North Carolina population-based study. *Journal of Investigative Dermatology*. 2010; 130(5):1463–5. [PubMed: 20043015]
- Thomas NE, Edmiston SN, Alexander A, Groben PA, Parrish E, Krickler A, Armstrong BK, Anton-Culver H, Gruber SB, From L, Busam KJ, Hao H, Orlov I, Kanetsky PA, Luo L, Reiner AS, Paine S, Frank JS, Bramson JI, Marrett LD, Gallagher RP, Zanetti R, Rosso S, Dwyer T, Cust AE, Ollila DW, Begg CB, Berwick M, Conway K, for The GEM Study Group. Association between NRAS and BRAF mutational status and melanoma-specific survival among patients with higher risk primary melanoma: A Population-Based Study. *JAMA Oncology*. 2015; 1(3):359–368. [PubMed: 26146664]
- Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society: Series B*. 2001; 63:411–423.
- van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007; 16:219–242. [PubMed: 17621469]
- van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011; 45(3)
- Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiologic heterogeneity across disease sub-types classified by multiple biomarkers. *American Journal of Epidemiology*. 2015; 182:263–270. [PubMed: 26116215]
- Wu S, Kuo H, Li WQ, Canales AL, Han J, Qureshi AA. Association between BRAFV600E and NRASQ61R mutations and clinicopathologic characteristics, risk factors and clinical outcome of primary invasive cutaneous melanoma. *Cancer Causes and Control*. 2014; 25(10):1379–1386. [PubMed: 25048604]
- Xue X, Kim MY, Gaudet MM, Park Y, Heo M, Hollenbeck AR, Strickler HD, Gunter MJ. A comparison of the polytomous logistic regression and joint Cox proportional hazards models for

evaluating multiple disease sub-types in prospective cohort studies. *Cancer Epidemiology, Biomarkers and Prevention*. 2013; 22:275–285.

Yu K, Zhang H, Wheeler W, Horne HN, Chen J, Figueroa JD. A robust association test for detecting genetic variants with heterogeneous effects. *Biostatistics*. 2015; 16:5–16. [PubMed: 25057183]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Schematic Representation of Clustering Strategy

Figure 2.
Plot of Heterogeneity Measure Versus k-Means Distance Measure for Clustering Solutions

Table 1

Analyses of Double Primaries: Marginal Correlations

Tumor Factor	# Cases with Data (n=903)			Individual Factor Comparisons	Cross-Classification Odds Ratio		H Values	
	First Primary	Second Primary	Both Tumors		Complete Data	Imputed Data	Complete Data	Imputed Data
Anatomic Site	903	903	903	Trunk/Arms vs H&N	3.4	3.4	0.41	0.41
				Legs vs H&N	18.6	18.6		
				Legs vs Trunk/Arms	4.0	4.0		
Mutation	138	348	116	<i>NRAS</i> vs WT	1.8	2.0	0.18	0.18
				<i>BRAF</i> vs WT	2.3	2.0		
				<i>BRAF</i> vs <i>NRAS</i>	3.0	4.1		
Histology	718	775	652	LMM vs Others	3.7	3.5	0.19	0.19
Nevus	585	717	534	Present vs Absent	2.4	2.2	0.19	0.19
Solar Elastosis	559	702	500	Present vs Absent	5.5	5.2	0.30	0.29
TILs *	562	672	485	Present vs Absent	2.2	2.1	0.12	0.12

* Tumor infiltrating lymphocytes

Table 2

Top Three Clustering Solutions

Solution #	Heterogeneity Measure (H)	Approximate Description
1	0.58	A: Legs B: Trunk/Arms + Solar Elastosis C: Trunk/Arms + no Solar Elastosis D: Head/Neck [*]
2	0.56	A: Legs B: Trunk/Arms + Neval Remnants C: Trunk/Arms + no Neval Remnants D: Head/Neck [†]
3	0.48	A: Legs B: Trunk/Arms + Wild Type C: Trunk/Arms + <i>BRAF/NRAS</i> D: Head/Neck [‡]

^{*} Head/Neck, *BRAF* mutated with nevus and tumor infiltrating lymphocytes are classified in cluster C.

[†] Head/Neck, *BRAF* mutated with nevus and tumor infiltrating lymphocytes are classified in cluster B.

[‡] Head/Neck, *BRAF* or *NRAS* mutated with nevus and tumor infiltrating lymphocytes are classified in cluster C.

Table 3

Logistic Regression Analysis of Sub-Types from Optimal Solution

Factors		Comparisons with Reference Sub-Type A [*] (Legs)			P-Value
		Odds Ratio (95% Confidence)			
		Sub-Type B [*] Trunk/ Arms SE	Sub-Type C [*] Trunk/ Arms no SE	Sub-Type D [*] Head/Neck	
Age	50-70 vs <50	1.4 (1.0-1.9)	0.8 (0.5-1.1)	1.6 (1.0-2.4)	<10 ⁻⁴
	>=70 vs <50	1.9 (1.3-2.8)	0.9 (0.5-1.4)	4.1 (2.4-6.7)	
Gender	Female	0.2 (0.2-0.3)	0.2 (0.1-0.2)	0.2 (0.1-0.3)	<10 ⁻⁴
Center	Rest vs Australia	1.7 (1.2-2.2)	1.4 (1.0-1.9)	1.4 (1.0-2.1)	0.01
Mole Count	5-10 vs 0-4	1.1 (0.8-1.5)	1.2 (0.8-1.8)	0.7 (0.5-1.2)	0.005
	11-25 vs 0-4	1.2 (0.8-1.6)	1.8 (1.2-2.6)	0.8 (0.5-1.2)	
	>25 vs 0-4	1.4 (0.9-2.0)	2.0 (1.3-3.1)	1.2 (0.7-2.0)	
Eye Color	Light	1.5 (1.1-2.1)	1.2 (0.9-1.8)	1.5 (1.0-2.4)	0.02
Hair Color	Light vs Dark	0.9 (0.7-1.2)	1.0 (0.7-1.4)	0.8 (0.5-1.2)	0.34
	Red vs Dark	1.0 (0.6-1.6)	0.8 (0.4-1.4)	0.9 (0.4-1.7)	
Sun Exposure (UVE dose)	2 nd quart. vs 1 st	1.5 (1.0-2.1)	0.7 (0.5-1.1)	1.1 (0.7-1.8)	<10 ⁻⁴
	3 rd quart. vs 1 st	1.2 (0.8-1.8)	0.6 (0.4-0.9)	1.4 (0.8-2.4)	
	4 th quart. vs 1 st	1.9 (1.2-3.1)	0.5 (0.3-0.9)	1.2 (0.7-2.3)	
Family Hx.	Yes vs No	1.3 (0.9-1.9)	1.5 (0.9-2.3)	1.2 (0.7-2.1)	0.79

* The relative frequencies of the 2103 cases allocated to the sub-types are: A (21%); B (40%); C (23%); D (15%). These are average frequencies over the 25 imputed datasets.