iCARE: An R Package to Build and Apply Absolute Risk Models

BY PAIGE MAAS

National Cancer Institute, Rockville MD, U.S.A.

email: paige.maas@mail.nih.gov

BY WILLIAM WHEELER

Information Management Services, Rockville MD, U.S.A.

email: wheelerb@imsweb.com

BY MARK BROOK

The Institute of Cancer Research, London, United Kingdom

email: mark.brook@icr.ac.uk

BY DAVID CHECK

National Cancer Institute, Rockville MD, U.S.A.

email: david.check@mail.nih.gov

BY MONTSERRAT GARCIA-CLOSAS

National Cancer Institute, Rockville MD, U.S.A.

email: montserrat.garcia-closas@mail.nih.gov

BY NILANJAN CHATTERJEE

Johns Hopkins University, Baltimore MD, U.S.A.

email: nchatte2@jhu.edu

Abstract

This report describes an R package, called the Individualized Coherent Absolute Risk Estimation (**iCARE**) tool, which allows researchers to quickly build models for absolute risk, and apply them to estimate an individual's risk of developing disease during a specified time interval, based on a set of user defined input parameters. An attractive feature of the software is that it gives users flexibility to update models rapidly based on new knowledge of risk factors and tailor models to different populations. The tool requires three input arguments be specified: (1) a model for relative risk (2) an age-specific disease incidence rate and (3) the distribution of risk factors for the population of interest. The tool handles missing risk factor information for individuals for whom risks are to be predicted using a coherent approach where all estimates are derived from a single model after appropriate model averaging. The software allows single nucleotide polymorphisms (SNPs) to be incorporated into the model using published odds ratios and allele frequencies. We discuss the statistical framework, handling of missing data and genetic factors, and provide real data examples that demonstrate the utility of **iCARE** for building and applying absolute risk models, using breast cancer as an example.

1 Introduction

Absolute risk models estimate disease risk in an upcoming time interval based on known risk factors for healthy individuals in a population, accounting for the presence of competing outcomes, such as death from other causes [5]. Absolute risk models for cancers and other diseases have important clinical and public health applications. Assessment of absolute risk of disease is fundamental for developing health intervention strategies to optimize an individual's risks and benefits. For example, absolute risk models can be used to identify individuals who have a high risk of disease in order to target screening and disease prevention strategies [8, 9, 14, 4]. Decisions regarding the initiation of screening or preventative intervention are often made on the basis of age and family history (fh), as proxies for risk. However, there is increasing consensus in the medical community that these decisions should instead be guided directly by individualized estimates of risk, which can be obtained from absolute risk models that include a wider array of environmental and genetic risk factors. Assessment of the distribution of risks for individuals in the population allows public health researchers to weigh the risks and benefits of a given intervention, such as a screening regimen, for the entire population [6, 3, 12]. Absolute risk models can also be applied to assess the power of clinical trials by projecting the expected distribution of disease risk from the distribution of risk factors in a population [4]. At an individual level, absolute risk estimates can be used to counsel individuals on the basis of their personal risk.

As large-scale epidemiologic studies continue to discover new risk factors for many diseases, there is a growing demand to develop and apply models for absolute risk prediction that can facilitate translation of our understanding of etiology into tools for managing health at the clinical and public health levels. There currently does not exist general software for researchers to build, update, and apply absolute risk models in R, and the Individualized Coherent Absolute Risk Estimation (**iCARE**) package provides this much needed capability.

The **iCARE** package fits absolute risk models by synthesizing multiple data sources containing information on relative risks, the distribution of risk factors in the population, and age-specific incidence rates for the disease of interest and rates of competing risks. This compartmentalization allows researchers to incorporate the best available information on key model parameters, to easily update models as new information becomes available, and to tailor or extend models to particular populations. Releasing **iCARE** will reduce that start up time for researchers, help standardize the methodology, and make it easy to share absolute risk models and make associated analyses reproducible. The package also implements methods for handling missing data, which is likely to be an issue in practice, and gives special attention to the efficient incorporation of genetic factors based solely on published information.

2 iCARE Methodology: Synthesizing Data Sources

Here, we present the statistical framework underlying the **iCARE** package. We describe the data inputs that are required to use the tool, examples of appropriate sources for the data, and details regarding how the key inputs are used to estimate model parameters. Specifically, we explain the methodology used to estimate the baseline hazard function component of the model and the approach used to handle missing data in the risk factor profiles used in the estimation of individuals' risks. We describe the tool's special treatment of SNPs, which allows genetic information to be incorporated into the model based on published information.

2.1 Model

The **iCARE** package fits a model for absolute risk, which assumes the age-specific incidence rates of the disease given a set of risk factors, Z, follows the Cox proportional hazard (PH) model [2] of the form

$$\mathrm{pr}(T \in [t, t + \Delta t) | T \geq t, Z) = \lambda_0(t|Z) = \lambda_0(t) \exp(\beta^T Z),$$

where T represents the event time of diagnosis for the disease of interest. The model assumes that risk factors Z act in a multiplicative fashion on the baseline hazard function, $\lambda_0(t)$. Given this model, the absolute risk of the disease for an individual who is currently at age a over the time internal $a + \tau$ is defined as [5],

$$\int_a^{a+\tau} \lambda_0(t)\exp(\beta^T Z) exp\left(-\int_a^t \left[\lambda_0(u)\exp(\beta^T Z) + m(u)\right] du\right) dt. \tag{1}$$

Formula (1) accounts for competing risks due to mortality from other causes through the age-specific mortality rate function $m(t)$. In the current implementation, for simplicity it is assumed that risk of mortality does not depend on the risk factor Z, but the method in principle can be extended to relax this assumption if covariate-specific risks of competing mortality can be estimated from external sources or models.

2.2 Data and Estimation

In order to fit the above model and apply it for absolute risk estimation, users must provide three main data sources:

- a model for the relative risk (or hazard ratio) parameters: β
- a marginal age-specific disease incidence rate: $\lambda_m(t)$
- a dataset containing risk factors for a set of representative individuals that could be used to estimate the risk factor distribution for the underlying population: Z_j for $j = 1, ..., N_{ref}$

In order to account for competing risks, an optional input with age-specific incidence rates of all-cause mortality, ideally excluding the disease of interest, $m(t)$ should also be provided.

The **iCARE** tool computes absolute risk estimates as the sum of the integrand of (1) over integer ages in the time interval of interest. The user-provided hazard ratio parameter estimates, $\hat{\beta}$, are plugged into the equation directly to carry out the computation. There are a number of ways that these input parameters may be obtained. For example, the estimates $\hat{\beta}$ may be derived from the analysis of a prospective cohort study using a multivariate PH model. Alternatively, they may be obtained from the analysis of a case-control study using a multivariate logistic regression model adjusted for fine categories of age, the parameters of which have been shown to approximate the PH model [15]. Ideally, datasets used to estimate model parameters should include information on all risk factors of interest and be large enough to provide precise estimates. When this is not available, estimates of relative risk for different risk factors could be obtained from multiple data sources (e.g. large published studies or meta-analyses). It is important that the provided estimates account for possible confounding (i.e. are adjusted for other risk factors in the model), and interactions.

The second data source needed for the model is an estimate of the overall (or marginal) age-specific disease incidence rate, defined as

$$\mathrm{pr}(T \in [t, t + \Delta t) | T \geq t) = \lambda_m(t),$$

for the population of interest. This information, for example, could be available from population-based registries, such as the United States' Surveillance Epidemiology and End Results (SEER) cancer registry maintained by the National Cancer Institute [7]. Similarly, users that wish to account for competing risks must provide the optional marginal age-specific incidence rates of all-cause mortality excluding the disease of interest

$$\mathrm{pr}(M \in [t, t + \Delta t) | M \geq t) = m(t).$$

In general it is best to incorporate rates defined for fine age categories, such as 1- or 5-year age strata, however **iCARE** can accommodate information on coarser age strata as well. For estimation, the age-specific disease incidence rates $\lambda_m(t)$ are used in combination with the third data input, a dataset of risk factors that is representative of the population of interest, to estimate the baseline hazard function, $\lambda_0(t)$.

2.3 Estimating the Baseline Hazard Function

Given the model of relative risks, $\hat{\beta}$, and marginal age-specific disease incidence rates, $\hat{\lambda}_m(t)$, we use the following relationship to derive the baseline hazard rate

$$\lambda_m(t) = \lambda_0(t) E\left[\exp(\beta^T Z | T \geq t)\right] = \lambda_0(t) \int \exp(\beta^T z) pr(z | T \geq t) dz, \qquad (2)$$

where, under the proportional hazard model,

$$pr(z | T \geq t) = \frac{\exp(-\int_0^t \lambda_0(u) exp(\beta^T z) du)}{\int \exp\left\{-\int_0^t \lambda_0(u) exp(\beta^T z) du\right\} dF(z)}$$

with $F(Z)$ denoting the distribution of the risk factors in the underlying population. If the disease can be assumed to be rare, then (2) can be approximated in closed form as

$$\lambda_m(t) \approx \int \lambda_0(t) \exp(\beta^T z) dF(z).$$

Computationally, the **iCARE** implementation starts with an initial value for $\lambda_0(t)$ based on the rare disease approximation and iterates based on formula (2) to obtain more exact estimates. This approach is closely related to an alternative formula for estimation of $\lambda_0(t)$ described by Gail et al.[5]. That approach involved an alternative maneuvering of the formula to allow estimation based on the risk factor distribution from a random sample of cases. In contrast, our estimation method relies on an available distribution of the risk factors for a general population. Thus, a model based on our proposed method (as implemented by **iCARE**) can be easily updated to reflect the risk factor distribution for different populations without requiring access to a sample of cases from each population of interest.

2.4 Specification of risk factor distribution

As detailed in Section 2.3, the risk factor distribution $F(Z)$ plays a key role in calibrating the model to the marginal disease rates in the underlying population. Thus, to carry out the calibration, the user must provide individual level data on the model risk factors for a sample that is representative of the underlying population. Ideally, this representative dataset would simply be the empirical distribution of Z, from a national survey, an epidemiologic study such as a population-based cohort, or controls from a population-based case-control study sampled from the population of interest. When empirical data are available, there are no additional modeling assumptions needed. However, if complete empirical data in all risk factors is not available, users can instead provide a representative dataset that may have been simulated under modeling assumptions appropriate to the population of interest. For example, in the application illustrated later we develop a model for absolute risk of breast cancer and incorporate a representative dataset of risk factors Z which were simulated based on data from a combination of national surveys.

2.5 Handling Missing Data in Covariate Profile

In addition to providing the three data inputs for estimating model parameters, users must provide information on risk factors for the individuals to whom the model should be applied. When there is complete information for all risk factors of interest, risk estimation is as straightforward as plugging the individuals Z into formula (1). However, in practice there may be missing data on some of the risk factors for individuals for whom we want to produce risk estimates.

One way to handle missing data on risk factors for a given individual is to use multiple imputation procedures [16]. The user would obtain estimates of absolute risk using iCARE for each of the completed-by-imputation risk factor profiles for the individual, and then average the absolute risk estimates to obtain an overall estimate of the absolute risk for that individual.

The **iCARE** tool also provides an internal option for handling missing data in the covariate profile for prediction: model-free imputation based on the referent representative dataset of risk factors provided by the user. The methodology underlying this imputation is as follows. For any subject indexed by i with a covariate profile Z_i, we define the risk score $R_i = \beta^T Z_i$, the linear predictor associated with the user specified log relative risk model. If a subject has missing values in some of the covariates, we partition $R_i = R^o_{iP} + R^u_{iP}$, where P indexes the observed pattern of missing data (i.e. which covariates are observed and which are missing) and where $R^o_{iP} = {\beta^o_P}^T Z^o_{iP}$ and $R^u_{iP} = {\beta^u_P}^T Z^u_{iP}$ denote the corresponding "observable" and "unobservable" components of the risk score. In general, this partitioning depends on which columns of the design matrix of the original model can be specified by the observed set of covariates for a given individual's risk factor profile. Given this partitioning, the absolute risk, AR, of the individual i is defined by

$$AR(R_{iP}) = \sum_{r^u_{iP}} AR(R^o_{iP}, r^u_{iP})\text{pr}(r^u_{iP}|R^o_{iP}) = \text{E}\big[AR(R^o_{iP}, r^u_{iP})|R^o_{iP}\big]. \qquad (3)$$

The absolute risk for the i-th individual is obtained by averaging over possible values for the unobserved component of the risk score given the value of the observed component of the risk score. As all the risk scores are scalar quantities, one can estimate the conditional distributions $\mathrm{pr}(r^m_{iP}|R^o_{iP})$ in a non-parametric fashion using the user-specified referent dataset.

In particular, to carry out (3) for a given covariate profile with missing data, the method finds subjects in the reference dataset that are similar on the basis of the observable component of the risk score, R^o_{iP}, and take as the risk estimate the average of the full model risk, $AR(R^o_P, r^u_P)$, for the referent subjects identified to be similar. Specifically, the observable risk scores R^o_{jP} are obtained for $j = 1, ..N$ in the referent dataset, categorized into single percentile strata, and the individual's R^o_{iP} is matched to one of the strata. The reported risk for the individual is then computed by averaging over the values of the full $AR(R^o_P, r^u_P)$ for all referent subjects in this matching stratum. This method can be viewed as a type of "hot deck" imputation based on the risk score, which is popular in survey literature.

2.6 Special treatment of SNP markers

As large genome-wide association studies continue to discover low penetrant, common SNPs associated with risk of complex chronic diseases, it is important to investigate the utility of the SNPs, in combination with other risk factors, for public health strategies of disease prevention. Evaluation of absolute risk, as opposed to relative risk, which is typically used for summarizing associations, is fundamental for these public health applications. Due to the importance of SNP markers in absolute risk models and natural assumptions specific to genetic data, the **iCARE** package provides a number of options for incorporating SNPs into the model.

Users can include individual SNPs in the model, or include a polygenic risk score (PRS), in the same way as any other risk factor as long as all input components can be identified. This allows researchers to specify interactions between SNPs and other risk factors in the model or to include PRSs with more complex weighting structures if desired. However, to include SNPs this way, a referent dataset must be provided that has the individual SNPs (or the PRSs) for all subjects. Again, researchers may create this referent distribution by creating a simulated dataset of individuals who are representative of the underlying population if necessary.

Alternatively, the **iCARE** package also provides a special approach for handling independent SNPs, which requires that the user only provide information on the odds ratio θ_k and population allele frequency f_k for each SNP to be included. **iCARE** internally creates a PRS from all provided SNPs weighted by the odds ratios,

$$PRS_i = \sum_k log(\theta_k) G_k,$$

where G_k denotes the SNP genotype status of individuals, coded as the number of non-referent alleles they carry (with respect to the referent allele for which the odds ratios are reported).

In general, iCARE assumes this PRS to be distributed independently of all other covariates. However, if a family history variable is included in the model, then the method allows a simple adjustment for the expected correlation between PRS and family history. The adjustment method assumes the latter is coded as a binary indicator of the presence or the absence of disease among first-degree relatives. In particular, when the model risk factors include family history, **iCARE** provides the option to adjust the log odds ratio associated with family history using the formula

$$\beta^A_{Fh} = \beta_{Fh} - 0.5 \sum_k 2 \left\{ log(\theta_k) \right\}^2 \times 2 f_k (1 - f_k)$$

with θ_k denoting the disease odds ratio of the SNPs, unadjusted for family history. This adjustment reflects the fact that, with the addition of SNPs into the model, the effect of family history is attenuated by a magnitude that is proportional to the degree of heritability explained by the SNPs. This treatment should be applied only when the provided β_{Fh} represents the effect of a binary variable for first-degree family history, unadjusted for the SNPs.

Users may provide relative risk estimates for family history that are already adjusted for the SNPs in the model, and if so they should simply not select the option for the family history adjustment.

One important way in which this approach treats SNPs differently involves the referent dataset of risk factors. Recall that this dataset might come from a national survey; however, a national survey is unlikely to have genotyped individuals, particularly for the exact set of SNPs to be included in the model. Recognizing this, for user convenience the referent dataset need only include non-genetic risk factors and **iCARE** will simulate SNP genotype values based on the provided allele frequencies for the population. SNPs are multiply imputed with user-specified `n.imps` determining the number of imputations for each subject in the referent dataset. The method assumes that the SNPs are independent and that the genotype distributions follow Hardy-Weinberg Equilibrium. Specifically, the joint distribution of SNP genotypes and other risk factors (X) are assumed to follow the decomposition

$$\text{pr}(g_1, \ldots, g_k, X) = \text{pr}(g_1, \ldots, g_k | fh) \times \text{pr}(X).$$

If family history of the disease is included in the model as a binary risk factor indicating the presence or absence of any first-degree relative with disease history, assuming that the disease is rare, we approximate

$$\text{pr}(g_1, \ldots, g_k | fh = 0) \approx \text{pr}(g_1, \ldots, g_k) = \text{pr}(g_1) \times \ldots \times \text{pr}(g_k).$$

The distribution of SNP genotypes among subjects with family history is approximated as

$$\text{pr}(g_1, \ldots, g_k | fh = 1) \approx \text{pr}(g_1 | fh = 1) \times \ldots \times \text{pr}(g_k | fh = 1), \text{ where}$$
$$\text{pr}(g_k | fh = 1) = \frac{\theta_k^{0.5} \text{pr}(g_k)}{\sum_k \theta_k^{0.5} \text{pr}(g_k)}.$$

The above approximation is derived under the assumption of rare disease and multiplicative effect of SNPs on the risk of the disease. If family history is not indicated to be in the model, and is thus not provided for each referent dataset subject, we impute the SNPs based on the unconditional distribution for independent SNPs in Hardy-Weinberg equilibrium.

It is possible that SNP information may also be missing in the covariate profiles for whom the model will be applied to estimate risk. In this case, SNPs are treated the same as all other risk factors and handled according to the methodology given in Section 2.5. Again, this approach is equivalent to averaging over the possible values of the missing SNPs according to the population distribution, taking advantage of any known SNPs in the genotype profile.

3 Using the iCARE package

In this section, we demonstrate how to use **iCARE** to build and apply two absolute risk models for breast cancer: one with SNPs only, and one with risk factors and SNPs.

The main function in **iCARE** is `compute.absolute.risk`. The input arguments to this function are named with the prefix "`model.`" or "`apply.`" according to whether they are used primarily for model building or application respectively.

To begin, the R package and the example dataset `breast_cancer` should be loaded:

```
R> library("iCARE")

R> data("breast_cancer", package="iCARE")
```

Example 1: SNP-only Model

To specify a SNP-only model, we must input the marginal age-specific disease incidence rates of breast cancer and the SNP information matrix, `snp.info`, that has three columns named: `snp.name`, `snp.odds.ratio`, and `snp.freq`. Marginal age-specific incidence rates of competing risks are optional, and in this example we include them.

```
R> bc_15_snps <- breast_cancer$bc_15_snps

R> bc_inc      <- breast_cancer$bc_inc

R> mort_inc    <- breast_cancer$mort_inc
```

Here, `bc_15_snps` contains published information on 15 SNPs identified to be associated with breast cancer risk by a recent genome-wide association study [11]. `bc_inc` contains age-specific incidence rates of breast cancer from SEER, and `mort_inc` has age-specific incidence rates of all-cause mortality from the WONDER mortality database [13]. In fitting a SNP-only model, the referent dataset need not be provided as **iCARE** will impute the referent SNP distribution. The function call below builds an absolute risk model based on 15 SNPs for breast cancer and applies the model to estimate risk of breast cancer in the interval from age 50 to age 80:

```
R> res_snps_miss = compute.absolute.risk(model.snp.info =  bc_15_snps,
                                model.disease.incidence.rates =  bc_inc,
                              model.competing.incidence.rates =  mort_inc,
                                              apply.age.start =  50,
                                      apply.age.interval.length =  30
                                            return.refs.risk =   T)
```

Note, for this SNP-only model, we exercised the option of not providing any new profiles for estimation (i.e. no `apply.snp.profile` input). In this case, **iCARE** simulates N=10,000 SNP profiles internally for the referent dataset and reports as the risk estimate the average of the risks estimated from the profiles: 0.09583. We can access the estimated risks for the (simulated) referent profiles and obtain summary information by calling

```
R> summary(res_snps_miss$refs.risk),
```

which yields the following output:

```
Risk_Estimate
Min.   :0.07474
1st Qu.:0.09196
Median :0.09573
Mean   :0.09583
3rd Qu.:0.09957
Max.   :0.12008
```

From this, we learn that on average women of age 50 have a 9.6% chance of being diagnosed with breast cancer before age 80, and that the 15-SNP model stratifies breast cancer risk from a minimum risk of 7.5% to a maximum risk of 12.0% in the interval 50-80.

If we wished to predict breast cancer risk for three specific women whom we had genotyped, we might call

```
R> new_snp_prof <- breast_cancer$new_snp_prof

R> res_snps_dat <- compute.absolute.risk(model.snp.info =  bc_15_snps,
                              model.disease.incidence.rates =  bc_inc,
                            model.competing.incidence.rates =  mort_inc,
                                            apply.age.start =  50,
                                  apply.age.interval.length =  30,
                                         apply.snp.profile =  new_snp_prof,
                                          return.refs.risk =   T)
```

Now our output `res_snps_dat$risk` contains the risk estimates for the three women whose genotype profiles we provided. Additionally, `res_snps_dat$refs.risk` contains the risk estimates for the referent dataset (again N=10,000 simulated internally) because we requested that those risks also be reported. These results allow us to create a useful plot, like Figure 1, showing the distribution of risks in our referent dataset and to add the risks of the three women to see where they fall on the population distribution, with the code

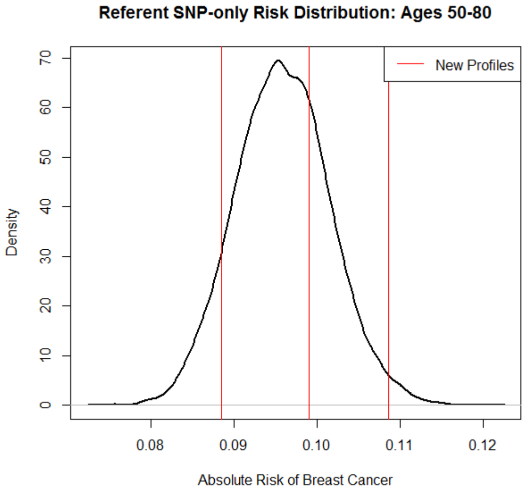
```
R> plot(density(res_snps_dat$refs.risk), xlab="Absolute Risk of Breast Cancer",
main="Referent SNP-only Risk Distribution:  Ages 50-80")
R> abline(v=res_snps_dat$risk, col="red")
R> legend("topright", legend="New Profiles", col="red", lwd=1)
```

Note, in this example the first genotype profile was missing two SNP values, demonstrating **iCARE**'s ability to produce risk estimates when there is missing data in the profile, with no inconvenience to the user.

Example 2: Breast Cancer Risk Model with Risk Factors and SNPs

The process of building and applying a breast cancer risk model with risk factors and SNPs follows much the same approach as in the SNP-only model in Example 1, however we must

Figure 1: Estimated Risk for Three Women on Population Distribution of Risk in 50-80



specify a few additional arguments.

```
v1 = list(); v1$name = "famhist"; v1$type = "continuous"

v2 = list(); v2$name = "parity";  v2$type = "factor"   ; v2$levels = c(0,1,2,3,4)

bc_model_cov_info <- list(v1, v2)

bc_model_log_or   <- breast_cancer$bc_model_log_or

ref_cov_dat        <- breast_cancer$ref_cov_dat
```

Having prepared the data sources, we can now run

```
R> res_covs_snps$details = compute.absolute.risk(
                                   model.formula = caco ~ famhist + as.factor(parity),
                                  model.cov.info = bc_model_cov_info,
                                  model.snp.info = bc_15_snps,
                                    model.log.RR = bc_model_log_or,
                               model.ref.dataset = ref_cov_dat,
                   model.disease.incidence.rates = bc_inc,
                 model.competing.incidence.rates = mort_inc,
                               model.bin.fh.name = "famhist",
                                 apply.age.start = 50,
                       apply.age.interval.length = 30,
                               apply.cov.profile = new_cov_prof,
                               apply.snp.profile = new_snp_prof,
                                return.refs.risk = T)
```

With the exception of `model.bin.fh.name`, which is always optional, all arguments listed in `green` should either be included or excluded in the function call as a set. This is to say that if one is included, then all should be included.

This fits an absolute risk model with risk factors family history and parity (i.e. number of children) additively with the 15 SNPs associated with breast cancer. In a model that includes risk factors, such as this one, we must supply the model formula, the risk factor information, the log odds ratios for the risk factors, and a referent dataset of risk factors to build the model. The `model.cov.info` input tells the function that family history can be treated as a continuous variable (though it only has levels 0 and 1) and that parity should be treated as a factor variables with levels 0,1,2,3, and 4 indicating the number of children for a given subject. Here, the `bc_model.log.or` input contains the log odds ratios for family history and parity, from a logistic regression model adjusted for cohort and fine categories of age in the Breast and Prostate Cancer Cohort Consortium [1, 10]. The `ref_cov_dat` dataset was created by simulation from the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES), which are representative of the US population. We indicate `model.bin.fh.name = "famhist"` to allow the software to properly attenuate the log odds ratio for family history to account for the addition of the 15 SNPs.

In addition to summarizing and plotting the risk estimates, **iCARE** includes an option to view more detailed output, by calling

```
R> print(res_covs_snps$details),
```

which reports the interval start and end ages over which absolute risk was computed, the entire covariate profile to which the model was applied (SNPs and risk factors if applicable), and the resulting risk estimate.

	Int_Start	Int_End	Risk_Est	rs12405132	rs12048493	rs72755295	rs6796502	rs13162653	rs2012709	rs7707921	rs9257408	rs4593472	rs13365225	rs13267382	rs11627032	chr17:29230520:D	rs745570	rs6507583	famhist	parity
P1	50	80	0.09434	NA	NA	0	0	0	2	1	1	0	0	1	2	1	0	0	0.00	2.00
P2	50	80	0.08072	0	0	1	0	1	1	2	1	1	1	1	1	0	0	0	0.00	4.00
P3	50	80	0.07232	2	0	0	0	1	1	0	0	1	0	0	1	1	1	0	0.00	2.00

In this case, both the profiles P1 and P3 had the same levels of the risk factors family history and parity, however we estimate that P1 has a 9.4% chance of breast cancer in the age interval 50 to 80, which is higher than P3's chance of 7.2%, due to the fact that the two have very different genotype profiles. This detailed output is also helpful for visually reminding users of whether they had any missing data in the covariate profiles used for estimation.

Additional Options

iCARE provides several advanced options as well. For example, `model.ref.dataset.weights` allows the user to optionally specify a vector of weights for each row in the referent dataset. Whenever any averaging is performed over the referent dataset, such as in the case of missing covariates for prediction, a weighted average is applied using the provided sampling weights. Additionally, **iCARE** allows the time intervals over which risk is to be computed to differ for each subject; this flexibility is useful, for example, in estimating 5-year risks for healthy individuals starting from their current ages.

Using the `compute.absolute.risk.split.interval` function a user can also specify that the absolute risk interval be computed in two parts, using two different sets of parameters. This allows the proportional hazards assumption to be relaxed to some extent, by allowing the relationship between risk factors and the outcome to vary over time. For example, it is well documented that the relationships between certain risk factors, such as body mass index, and breast cancer are different among premenopausal and postmenopausal women. Using `compute.absolute.risk.split.interval`, users can specify a different set of relationships by inputting `model.log.odds.ratio` and `model.log.odds.ratio.2` for use prior to and after a cutpoint of age 50, the median age at menopause. This more advanced function is also helpful in the context where the distribution of risk factors varies with age.

In addition to returning risk estimates for the specified profiles, the `iCARE` functions can optionally return the absolute risks for the referent dataset as well if `return.refs.risk=T`. The relative risk scores, or $\beta^T Z_i$, for the covariate profiles can be obtained by requesting `return.lp=T`. For individuals where there is missing data in covariate profile Z, the reported linear predictor is the average of the full linear predictors of all referent subjects in the matching strata according to the approach described in Section 2.5.

4 Conclusion

The **iCARE** package is a new tool for building and applying absolute risk models by synthesizing data sources on key model parameters. The tool standardizes methodology and gives researchers the ability to easily update and share absolute risk models, and to evaluate the public health implications of etiologic findings by translating relative risks onto the absolute risk scale. The package incorporates calibration to population-based age-specific disease rates and handling of missing data by leveraging a referent dataset of risk factors for the population of interest. Through this handling of missing data and the ability to incorporate SNP information based on published estimates, the tool gives researchers the ability to easily handle analytic issues that are likely to arise in practice when building absolute risk models for health contexts. In this paper we have described the methodology underlying this new tool and illustrated its use with examples by building absolute risk models for breast cancer.

Acknowledgements

We acknowledge funding from Breast Cancer Now and NHS funding to the Royal Marsden and Institute Cancer Research NIHR Biomedical Research Centre. The research of Maas, Check and Chatterjee was supported by the intramural fund of the National Cancer Institute.

References

[1] Daniele Campa, Rudolf Kaaks, Loc Le Marchand, Christopher A. Haiman, Ruth C. Travis, Christine D. Berg, Julie E. Buring, Stephen J. Chanock, W. Ryan Diver, Lucie Dostal, Agnes Fournier, Susan E. Hankinson, Brian E. Henderson, Robert N. Hoover, Claudine Isaacs, Mattias Johansson, Laurence N. Kolonel, Peter Kraft, I-Min Lee, Catherine A. McCarty, Kim Overvad, Salvatore Panico, Petra H.M. Peeters, Elio Riboli, Maria Jos Sanchez, Fredrick R. Schumacher, Guri Skeie, Daniel O. Stram, Michael J. Thun, Dimitrios Trichopoulos, Shumin Zhang, Regina G. Ziegler, David J. Hunter, Sara Lindstrm, and Federico Canzian. Interactions between genetic variants and breast can-

cer risk factors in the breast and prostate cancer cohort consortium. *Journal of the National Cancer Institute*, 103(16):1252–1263, 2011.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):pp. 187–220, 1972.

[3] Mitchell H. Gail. The estimation and use of absolute risk for weighing the risks and benefits of selective estrogen receptor modulators for preventing breast cancer. *Annals of the New York Academy of Sciences*, 949(1):286–291, 2001.

[4] Mitchell H Gail. Personalized Estimates of Breast Cancer Risk in Clinical Practice and Public Health. *Statistics in Medicine*, 30(10):1090–1104, 2011.

[5] Mitchell H Gail, Louise A Brinton, David P Bfyar, K Donald, Sylvan B Green, Catherine Schairer, and John J Mutvihill. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *Journal Of The National Cancer Institute*, pages 1879–1886, 1989.

[6] S. M. Grundy. Primary prevention of coronary heart disease: integrating risk assessment with intervention. *Circulation*, 100(9):988–998, Aug 1999.

[7] N Howlader, AM Noone, M Krapcho, N Neyman, R Aminou, W Waldron, CL Altekruse, SF amd Kosary, J Ruhl, Z Tatalovich, H Cho, and et al. Seer cancer statistics review, 1975-2008. *National Cancer Institute*, 2011.

[8] R. Jackson. Guidelines on preventing cardiovascular disease in clinical practice. *BMJ*, 320(7236):659–661, Mar 2000.

[9] R. Jackson, C. M. Lawes, D. A. Bennett, R. J. Milne, and A. Rodgers. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *Lancet*, 365(9457):434–441, 2005.

[10] Amit D. Joshi, Sara Lindstrm, Anika Hsing, Myrto Barrdahl, Tyler J. VanderWeele, Daniele Campa, Federico Canzian, Mia M. Gaudet, Jonine D. Figueroa, Laura Baglietto, Christine D. Berg, Julie E. Buring, Stephen J. Chanock, Mara-Dolores Chirlaque,

W. Ryan Diver, Laure Dossus, Graham G. Giles, Christopher A. Haiman, Susan E. Hankinson, Brian E. Henderson, Robert N. Hoover, David J. Hunter, Claudine Isaacs, Rudolf Kaaks, Laurence N. Kolonel, Vittorio Krogh, Loic Le Marchand, I-Min Lee, Eiliv Lund, Catherine A. McCarty, Kim Overvad, Petra H. Peeters, Elio Riboli, Fredrick Schumacher, Gianluca Severi, Daniel O. Stram, Malin Sund, Michael J. Thun, Ruth C. Travis, Dimitrios Trichopoulos, Walter C. Willett, Shumin Zhang, Regina G. Ziegler, and Peter Kraft. Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the breast and prostate cancer cohort consortium. *American Journal of Epidemiology*, 180(10):1018–1027, 2014.

[11] Kyriaki Michailidou, Jonathan Beesley, Sara Lindstrom, Sander Canisius, Joe Dennis, Michael J. Lush, Mel J. Maranian, Manjeet K. Bolla, Qin Wang, Mitul Shah, Barbara J. Perkins, Kamila Czene, Mikael Eriksson, Hatef Darabi, Judith S. Brand, Stig E. Bojesen, Borge G. Nordestgaard, Henrik Flyger, Sune F. Nielsen, Nazneen Rahman, Clare Turnbull, BOCS, Olivia Fletcher, Julian Peto, Lorna Gibson, Isabel dos Santos-Silva, Jenny Chang-Claude, Dieter Flesch-Janys, Anja Rudolph, Ursula Eilber, Sabine Behrens, Heli Nevanlinna, Taru A. Muranen, Kristiina Aittomaki, Carl Blomqvist, Sofia Khan, Kirsimari Aaltonen, Habibul Ahsan, Muhammad G. Kibriya, Alice S. Whittemore, Esther M. John, Kathleen E. Malone, Marilie D. Gammon, Regina M. Santella, Giske Ursin, Enes Makalic, Daniel F. Schmidt, Graham Casey, David J. Hunter, Susan M. Gapstur, Mia M. Gaudet, W. Ryan Diver, Christopher A. Haiman, Fredrick Schumacher, Brian E. Henderson, Loic Le Marchand, Christine D. Berg, Stephen J. Chanock, Jonine Figueroa, Robert N. Hoover, Diether Lambrechts, Patrick Neven, Hans Wildiers, Erik van Limbergen, Marjanka K. Schmidt, Annegien Broeks, Senno Verhoef, Sten Cornelissen, Fergus J. Couch, Janet E. Olson, Emily Hallberg, Celine Vachon, Quinten Waisfisz, Hanne Meijers-Heijboer, Muriel A. Adank, Rob B. van der Luijt, Jingmei Li, Jianjun Liu, Keith Humphreys, Daehee Kang, Ji-Yeob Choi, Sue K. Park, Keun-Young Yoo, Keitaro Matsuo, Hidemi Ito, Hiroji Iwata, Kazuo Tajima, Pascal Guenel, Therese Truong, Claire Mulot, Marie Sanchez, Barbara Burwinkel, Frederik

Marme, Harald Surowy, Christof Sohn, Anna H. Wu, Chiu-chen Tseng, David Van Den Berg, Daniel O. Stram, Anna Gonzalez-Neira, Javier Benitez, M. Pilar Zamora, Jose Ignacio Arias Perez, Xiao-Ou Shu, Wei Lu, Yu-Tang Gao, Hui Cai, Angela Cox, Simon S. Cross, Malcolm W. R. Reed, Irene L. Andrulis, Julia A. Knight, Gord Glendon, Anna Marie Mulligan, Elinor J. Sawyer, Ian Tomlinson, Michael J. Kerin, Nicola Miller, kConFab Investigators, A. O. C. S. Group, Annika Lindblom, Sara Margolin, Soo Hwang Teo, Cheng Har Yip, Nur Aishah Mohd Taib, Gie-Hooi Tan, Maartje J. Hooning, Antoinette Hollestelle, John W. M. Martens, J. Margriet Collee, William Blot, Lisa B. Signorello, Qiuyin Cai, John L. Hopper, Melissa C. Southey, Helen Tsimiklis, Carmel Apicella, Chen-Yang Shen, Chia-Ni Hsiung, Pei-Ei Wu, Ming-Feng Hou, Vessela N. Kristensen, Silje Nord, Grethe I. Grenaker Alnaes, NBCS, Graham G. Giles, Roger L. Milne, Catriona McLean, Federico Canzian, Dimitrios Trichopoulos, Petra Peeters, Eiliv Lund, Malin Sund, Kay-Tee Khaw, Marc J. Gunter, Domenico Palli, Lotte Maxild Mortensen, Laure Dossus, Jose-Maria Huerta, Alfons Meindl, Rita K. Schmutzler, Christian Sutter, Rongxi Yang, Kenneth Muir, Artitaya Lophatananon, Sarah Stewart-Brown, Pornthep Siriwanarangsan, Mikael Hartman, Hui Miao, Kee Seng Chia, Ching Wan Chan, Peter A. Fasching, Alexander Hein, Matthias W. Beckmann, Lothar Haeberle, Hermann Brenner, Aida Karina Dieffenbach, Volker Arndt, Christa Stegmaier, Alan Ashworth, Nick Orr, Minouk J. Schoemaker, Anthony J. Swerdlow, Louise Brinton, Montserrat Garcia-Closas, Wei Zheng, Sandra L. Halverson, Martha Shrubsole, Jirong Long, Mark S. Goldberg, France Labreche, Martine Dumont, Robert Winqvist, Katri Pylkas, Arja Jukkola-Vuorinen, Mervi Grip, Hiltrud Brauch, Ute Hamann, Thomas Bruning, G. E. N. I. C. A. Network, Paolo Radice, Paolo Peterlongo, Siranoush Manoukian, Loris Bernard, Natalia V. Bogdanova, Thilo Dork, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M. Hartikainen, Peter Devilee, Robert A. E. M. Tollenaar, Caroline Seynaeve, Christi J. Van Asperen, Anna Jakubowska, Jan Lubinski, Katarzyna Jaworska, Tomasz Huzarski, Suleeporn Sangrajrang, Valerie Gaborieau, Paul Brennan, James McKay, Susan Slager, Amanda E. Toland, Christine B. Ambrosone, Drakoulis Yannoukakos, Maria Kabisch, Diana Torres, Susan L. Neuhausen, Hoda Anton-Culver, Craig Luccarini, Caroline Baynes, Shahana Ahmed, Catherine S. Healey, Daniel C.

Tessier, Daniel Vincent, Francois Bacot, Guillermo Pita, M. Rosario Alonso, Nuria Alvarez, Daniel Herrero, Jacques Simard, Paul P. D. P. Pharoah, Peter Kraft, Alison M. Dunning, Georgia Chenevix-Trench, Per Hall, and Douglas F. Easton. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet*, 47(4):373–380, Apr 2015. Letter.

[12] C. J. Murray, J. A. Lauer, R. C. Hutubessy, L. Niessen, N. Tomijima, A. Rodgers, C. M. Lawes, and D. B. Evans. Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk. *Lancet*, 361(9359):717–725, Mar 2003.

[13] National Center for Health Statistics (NCHS). Underlying cause of death 1999-2011 on cdc wonder online database, released 2014. data are from the multiple cause of death files, 1999-2011, as compiled from data provided by the 57 vital statistics jurisdictions through the vital statistics cooperative program., 2014.

[14] Paul D.P. Pharoah, Antonis C. Antoniou, Douglas F. Easton, and Bruce A.J. Ponder. Polygenes, risk prediction, and targeted prevention of breast cancer. *New England Journal of Medicine*, 358(26):2796–2803, 2008. PMID: 18579814.

[15] R L Prentice, J D Kalbfleisch, A V Peterson, N Flournoy, V T Farewell, and N E Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–54, 1978.

[16] Donald B. Rubin. *Procedures with Ignorable Nonresponse*, pages 154–201. John Wiley & Sons, Inc., 2008.