Journal of Biomedical Informatics 46 (2013) 174–180

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Extension of the survival dimensionality reduction algorithm to detect epistasis in competing risks models (SDR-CR)

Lorenzo Beretta *, Alessandro Santaniello

Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

ARTICLE INFO

Article history:
Received 3 November 2011
Accepted 4 November 2012
Available online 12 November 2012

Keywords:
Data mining
Epistasis
Survival analysis
Competing risks
Systemic sclerosis
Polymorphism

ABSTRACT

Background: The discovery and the description of the genetic background of common human diseases is hampered by their complexity and dynamic behavior. Appropriate bioinformatic tools are needed to account all the facets of complex diseases and to this end we recently described the survival dimensionality reduction (SDR) algorithm in the effort to model gene–gene interactions in the context of survival analysis. When one event precludes the occurrence of another event under investigation in the 'competing risk model', survival algorithms require particular adjustment to avoid the risk of reporting wrong or biased conclusions.

Methods: The SDR algorithm was modified to incorporate the cumulative incidence function as well as an adapted version of the Brier score for mutually exclusive outcomes, to better search for epistatic models in the competing risk setting. The applicability of the new SDR algorithm (SDR-CR) was evaluated using synthetic lifetime epistatic datasets with competing risks and on a dataset of scleroderma patients.

Results/conclusions: The SDR-CR algorithms retains a satisfactory power to detect the causative variants in simulated datasets under different scenarios of sample size and degrees of type I or type II censoring. In the real-world dataset, SDR-CR was capable of detecting a significant interaction between the IL-1α C-889T and the IL-1β C-511T single-nucleotide polymorphisms to predict the occurrence of restrictive lung disease vs. isolated pulmonary hypertension.

We provide an useful extension of the SDR algorithm to analyze epistatic interactions in the competing risk settings that may be of use to unveil the genetic background of complex human diseases. Availability: http://sourceforge.net/projects/sdrproject/files/.

1. Introduction

The development, the availability and the widespread diffusion of high-throughput technologies have helped, in the last decade, to lay the groundwork for the comprehension of the genetic architecture of complex human diseases. Despite these efforts, however, a large proportion of the estimated genetic variance of individuals remains unexplained [1,2]; several hypothesis have been put forward to explain the so-called 'missing heritability', including gene–gene interactions or epistasis [2–4]. It has indeed been argued that genetic complexity does not arise from the independent action of a large number of different genes but it is rather the result of extensive genetic interactions among them [5,6]. These concepts prompted the development or the application of different algorithms to detect meaningful gene–gene interactions in case-control studies; each method comes up with its own solutions as well as with its own inductive biases to the many computational challenges that the modeling of epistasis poses [7–9]. Not until the recent work by Beretta et al. [10] and by Gui et al. [11] the problem of epistasis in the context of survival analysis has specifically been approached from the machine-learning point of view. Cox regression may be suitable to detect non-linearities in presence of right-censored data, however, this method may not always be the optimal choice to this end: *a priori* knowledge of the variable relationships may be needed to properly model interactions and type I error may increase due to the inflated number of polynomial terms. The "survival" issue is of particular interest whenever the event of interest takes time to happen or does not happen at all due to a short observational time or to the loss of information during the follow-up period as in the case of recurrence of a disease, response to treatment and prognostication. We demonstrated that the novel survival dimensionality reduction (SDR) algorithm

Abbreviations: SDR, survival dimensionality reduction; SDR-CR, SDR for competing risks; KM, Kaplan–Meier; CIF, cumulative incidence function; IBS, integrated Brier score; SNP, single nucleotide polymorphism; SSc, systemic sclerosis; ILD, interstitial lung disease; PAH, pulmonary arterial hypertension.

* Corresponding author. Address: Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Via Pace 9, 20122 Milan, Italy. Fax: +39 02 5503 5289.

E-mail address: lorberimm@hotmail.com (L. Beretta).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.11.002

retains a satisfactory power to sort out a set of causative genes with mild-to-moderate epistatic effect size from a pool of candidate genes in synthetic lifetime datasets and that the algorithm was capable of identifying epistatic interactions that drive the response to anti-TNF biological agents in a population of patients with active rheumatoid arthritis [11]. One of the main advantages of SDR is its inherent non-parametric nature, as it is necessary to neither make *a priori* assumptions about the underlying interaction model nor about the shape of the underlying survival distribution. Yet, despite its advantages, SDR is not suitable in a number of clinically relevant contexts, as for instance when an event precludes the occurrence of another event under investigation. In this situation, generally referred to as 'competing risks', each individual is at risk of experiencing multiple events at any time, but cannot experience one outcome once failure has occurred due to another event [12,13] and each event is mutually exclusive to the others.

The survival model in presence of competing of risks is known as the 'competing risk model' and it requires adequate statistical techniques to make proper inferences about it. Several evidences suggest that the naïve application of traditional survival methods to the competing risk model, as for instance the non-parametric Kaplan–Meier (KM) estimator, leads to biased and inflated estimates of the probability of failure [13–15] and that the bias is greater when the hazard of the competing events is larger. More honest failure estimates can be obtained calculating the cumulative incidence function (CIF) via proper modification of the KM estimator. The SDR algorithm makes extensive use of KM failure estimates in its constitutive induction phase, where multilocus genotypes are either assigned to the 'high-risk' or to the 'low-risk' group (see Section 2). In presence of competing risks, a significant deterioration in the discriminative capability of the SDR method is therefore expected if these risks are not properly estimated or if the naïve KM is applied ignoring these risks altogether. In the present work, we'll illustrate that the modification of the SDR code to incorporate the CIF in its searching phase as well as to use the adapted version of the Brier score for competing risks [16] allows a correct and adequate detection of epistatic interactions in lifetime datasets with multiple and mutually-exclusive outcomes.

2. Methods

2.1. Review of the SDR algorithm

The SDR algorithm is detailed in [11]; it is main steps are hereafter summarized. Initially, the dataset is partitioned into k-nonoverlapping testing sets, where $k-1$ parts (training sets) are used for model construction and k testing sets are retained for model validation. In every training set, survival estimates are calculated via the KM method and multilocus cells resulting from the interaction of n biallelic genetic markers are represented into the multidimensional space. KM survival estimates are also calculated for each of these cells and their estimates are compared to those derived from the whole training set; the cells with average survival estimates lower than the training estimates are classified as 'high-risk' or as 'low-risk' otherwise. Examples from high-risk cells are pooled into one group and those from low-risk cells into another; predictions from both groups are then evaluated via the integrated Brier score (IBS) for censored data [17]. For each n-combination of variables, the k training IBS are averaged and the n-combination yielding the lowest mean IBS is selected and considered for model validation. Here, the individual subjects data from the k testing sets together with their assigned labels are merged sort to produce a larger T_M testing set and a *meta-IBS* is here computed (a validation procedure later described in [18]); the n-combination yielding the lowest *meta-IBS* is then chosen as the final model.

Whilst in the original form the SDR algorithm uses KM estimates to assign the "high-risk" or the "low-risk" labels to multilocus cells, these estimated can interchangeably substituted by the estimated cumulative hazards, 1 – KM. The possibility to adapt the SDR algorithm to the competing risks model is therefore dependent from an appropriate estimation of the outcome-specific cumulative hazards.

2.2. Notations for the competing risk model

Let $0 < t_1 < t_2 < \cdots < t$ be the ordered distinct time points at which failures *of any cause* occur and let k be the possible causes of failure, where $k > 1$.

We can calculate the survival estimates for *any* cause of failure via the Kaplan–Meier method:

$$\widehat{S}(t) = \prod_{t_j \leqslant t} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

where n_i is the number of cases "at risk" of any event prior to survival time t, and d_i, is the total number of events at time t_i.

For each time-point t_i it is possible to estimate the unconditional probability of failing from cause k, $p_k(t_i)$ as the product of the hazard for the specific cause k, $\lambda_k(t_i)$ and the probability of being event-free prior to t_i:

$$\widehat{p_k}(t_i) = \widehat{\lambda}_k(t_i)\widehat{S}(t_{i-1}), \quad \widehat{\lambda}_k(t_i) = \frac{d_{ki}}{n_i} \tag{2}$$

where d_{ki} is the number of subjects failing from cause k at t_i and $\widehat{S}(t_0) = 1$.

Finally, the cumulative incidence for the specific cause k, $I_k(t)$ can be calculated as the sum of the abovementioned terms $p_k(t_i)$ for every time-points before t:

$$\widehat{I}_k(t) = \sum_{t_i \leqslant t} \widehat{p_k}(t_i) \tag{3}$$

2.3. SDR for competing risks

The main difference between SDR and SDR for competing risks (SDR-CR) lies in the way risk estimates are calculated to assign either the "low-risk" or the "high-risk" label to multilocus cells for the competing outcomes. Predictions from pooled assignments are then evaluated with an adapted version of the Brier score that specifically accounts for competing risks.

After having parted the dataset into the desired number of training/testing sets, survival estimates are calculated into the training sets by applying Eq. (1). The complement of the estimated survival function is then computed:

$$\widehat{F}(t) = 1 - \widehat{S}(t) \tag{4}$$

Cumulative incidences for every cause of failure k, in every time-interval t_i, are then derived according to Eq. (3). Eqs. (4) and (3) are then applied to calculate the outcome-specific cumulative incidences in every multilocus cell c, here labeled as $\widehat{J}_k(t_i)$. k-Specific differences between the above-mentioned quantities are then calculated in every multilocus cell:

$$D_k(t_i)_c = \widehat{I}_k(t_i) - \widehat{J}_k(t_i) \tag{5}$$

All the $D_k(t_i)_c$ for each multilocus cells are then averaged via the geometric mean (GM); to avoid zero o negative values, these are transformed into a meaningful equivalent positive adding 1 to any $D_k(t_i)_c$ value:

$$GM_k(t)_c = \sqrt[t]{\prod_{t_i \leqslant t} [1 + D_k(t_i)_c]} \tag{6}$$

Cells with $GM_k(t)_c \leqslant 1$ are classified as "high-risk" and cells with $GM_k(t)_c > 1$ are classified as "low-risk" for the k outcome. Examples from high-risk cells for each k outcome are pooled into one group and those from low-risk cells for the same k outcome into another.

SDR-CR predictions are evaluated via a version of the Brier score for censored data modified to account multiple and mutually-exclusive outcomes, $BS_{CR}(t)_k$ [17]. The $BS_{CR}(t)_k$ for $0 < t_i < t$ and outcome k is defined as:

$$BS_{CR}(t)_k = \frac{1}{n}\sum_{i=1}^{n}[I(t_i \leqslant t \quad \delta_i = k) - \widehat{I}(t|X_i)]^2 \omega(t; t_i; \delta_i; \widehat{S}; X_i)$$

where $\widehat{S}(t)$ is the Kaplan–Meier estimate calculated according to Eq. (1) which is based on the observations $(t_i, 1-\delta_i)$ and I stands for the indicator function and:

$$\omega(t; t_i; \delta_i; \widehat{S}; X_i) = \frac{I(t_i \leqslant \delta_i)}{\widehat{S}(t_i - |X_i)} + \frac{I(t_i > t)}{\widehat{S}(t|X_i)}$$

$BS_{CR}(t)_k$ depends on time t, hence it makes sense to use the integrated Brier score for competing risks (IBS_{CRk}) as an overall measure for the prediction of the model at all times:

$$IBS_{CR_k} = t^{-1}\int_0^t BS_{CR}(t)_k dt$$

The lower the IBS_{CRk} the less inaccurate or, conversely, the more precise the prediction for the k cause-specific outcome is. Thus, for k-specific outcomes, SDR-CR yields k IBS_{CRk} values, to describe the best interactions that best explain the k competing causes of failure.

2.4. Data simulation and power calculation

The general process of data simulation is hereafter described and detailed in the end of this subsection; the process of competing risk simulation in the context of survival-time analysis is akin to [19].

a. The cumulative incidence functions and the cumulative specific hazards (CSH) for two mutually-exclusive outcomes are defined; this way the number (proportion) of individuals failing from each competing event at a pre-specified number of time-points is set.
b. CSH are generated in three ways to simulate different kind of scenarios: (1) hazards are kept constants for both outcomes (e.g. follow an exponential-shaped survival distribution); (2) the hazard for one outcome is kept constant while the other varies in time (e.g. follows a different survival class distribution); (3) the hazard of both outcomes varies in time.
c. A two-factor epistatic models is generated for each competing event and their distribution of multilocus cells "at risk" and "not at risk", are kept constant over time and fitted to the number of individuals failing at the pre-specified time-points.
d. A finite sample of individuals (0.4% and 0.6%) is randomly drawn from the original population. A 30% or 50% of the sampled individuals is then randomly censored. Sampling and censoring is repeated 100 times to eventually derive the success rate of the search algorithm

In detail, we firstly made reference to the logistic-exponential equation:

$$S(t) = \frac{1 + (e^{\lambda\theta} - 1)^k}{1 + (e^{\lambda(t+\theta)} - 1)^k} \quad t \geqslant 0;\ \lambda > 0;\ k > 0,\ \theta \geqslant 0$$

and to the corresponding hazard function [20]:

$$h(t) = \frac{\lambda k e^{\lambda t}(e^{\lambda t} - 1)^{k-1}}{1 + (e^{\lambda t} - 1)^k} \quad t \geqslant 0$$

here, $S(t)$ is the logistic-exponential survival distribution, t is the survival time, λ is a positive scale parameter and κ is a positive shape parameter and θ is $a \geqslant 0$ parameter that shifts the distribution to the left. From $S(t)$ we can derive the cumulative incidence function $F(t)$ as $1 - S(t)$.

We then considered the existence of two competing risks, each characterized by its own $S(t)$ and $F(t)$, termed $S_1(t)$ and $S_2(t)$, and $F_1(t)$ and $F_2(t)$, respectively. From these distributions we can derive the cumulative incidence function in the whole population where the two risks compete:

$$F_p(t) = F_1(t) + F_2(t)$$

For $S_1(t)$ and $S_2(t)$ a finite number of solution exist so that $F_P(t) \leqslant 1$, any other hypothetical scenario that violates this assumption also violates the conditional independence of outcomes [12,13] and thus the model is misspecified by the user. The adherence to the abovementioned requirement was ensured during simulation via a try-and-error procedure.

For simulation purposes, the survival time t was set to five time units ($t = t_5$) and λ_1, κ_1 and θ_1 as well as λ_2, κ_2 and θ_2 were adjusted so that $F_P(t_5)$ was equal to an arbitrary value of 0.6, where $F_1(t_5) = F_2(t_5) = 0.3$, respectively. The shape of the underlying survival distribution for each of the competing event was set as follows: Simulation (1) $S_1(t)$ = Exponential (EXP), $S_2(t)$ = EXP; Simulation (2) $S_1(t)$ = EXP, $S_2(t)$ = Bathtub-Shaped Failure Rate (BT); Simulation (3) $S_1(t)$ = BT, $S_2(t)$ = Increasing Failure Rate (IFR). Example plots depicting the cause-specific hazard functions, the cause-specific cumulative hazards and the CIF when $F_1(t_5) = F_2(t_5) = 0.3$ are depicted in Fig. 1.

According to Culverhouse et al. [21], we then generated two different epistatic models (e.g. one model per competing risk), each composed of two biallelic SNPs, A_1–B_1 and A_2–B_2 in Hardy–Weinberg equilibrium (HWE) with $qA_1 = qA_2 = qB_1 = qB_2 = 0.2$ so that $K_1(t_5) = KA_1 = KB_1 = F_1(t_5)$ and that $K_2(t_5) = K_{A2} = K_{B2} = F_2(t_5)$, where KA_1–KB_1 and K_{A2}–K_{B2} are the marginal penetrances for SNP A_1–B_1 and SNP A_2–B_2. In these models the proportion of phenotypic variance attributable to genetic variation, that is the broad-sense heritability (H^2) can easily be calculated. We set the multilocus penetrances so that the penetrances of the epistatic models fit the cumulative prevalence of events at the survival time, $F_1(t_5) = F_2(t_5)$; we therefore define the H^2 of these models as the a *cumulative estimate* of H^2 at t_5, $H^2(t_5)$. The cause-specific $H^2(t_5)$ was set to 0.075, 0.10, 0.125 or 0.15 to simulate populations where the genetic contribution is low-to-mild, that is we expect that the genetic interaction of two causative SNPs explains from 7.5% to 15% of the cumulated events per outcome.

To calculate the number of events per time-point we proceeded as follows: let $f_{A1B1}(t_i)$ and $f_{A2B2}(t_i)$ the cumulative multilocus penetrances for the two epistatic models, where $0 < t_i \leqslant t$ and $A, B = 0, 1$ or 2 according to the number of mutant alleles. Time-point cumulative multilocus penetrances are proportionally derived from $F_1(t_5)$ and $F_2(t_5)$:

$$f_{A1B1}(t_i) = f_{A1B1}(t_5) * F_1(t_i)/F_1(t_5)$$
$$f_{A2B2}(t_i) = f_{A2B2}(t_5) * F_2(t_i)/F_2(t_5)$$

From these values time-point cumulative estimates of H^2 or $H^2(t_i)$ can easily be calculated and are reported in detail in Appendix A.

Time-point multilocus penetrances $f^*_{A1B1}(t_i)$ and $f^*_{A2B2}(t_i)$ were then derived from cumulative penetrances:

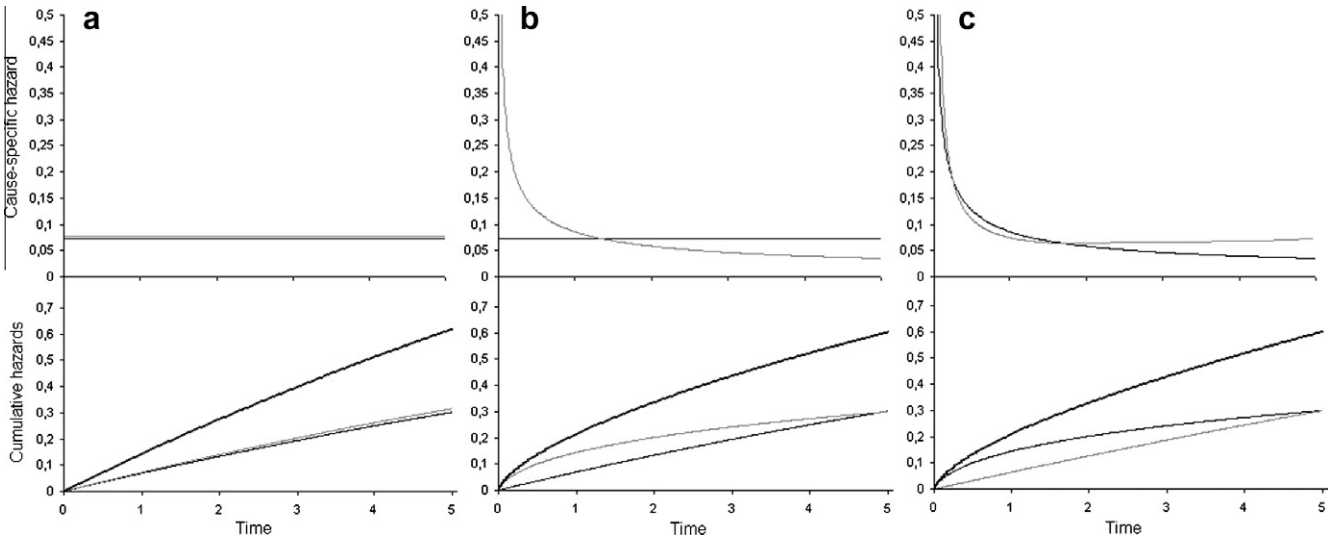

Fig. 1. Cause-specific hazards and cumulative hazards used for simulation. Cause-specific hazards (top panels) and cause-specific cumulative hazards (bottom panel) in the populations used for simulation; each outcome parameter is represented by a thin gray or black line, the cumulative incidence functions by a thick black line. Shape parameters for the logistic-exponential equation and the corresponding hazard functions [19] were as follows: (a) $\lambda = 0.072$, $\kappa = 1$ (gray and black lines); (b) $\lambda = 0.072$, $\kappa = 1$ (black line) and (c) $\lambda = 0.04$, $\kappa = 0.5$, $\theta = 0.023$ (gray line); $\lambda = 0.04$, $\kappa = 0.5$, $\theta = 0.023$ (black line) and $\lambda = 1.301$, $\kappa = 0.08$, $\theta = 5.011$ (gray line). The corresponding survival distributions were: (a) exponential (EXP) and EXP; (b) EXP and bathtub-shaped failure rate (BT); (c) BT and Increasing Failure Rate (IFR).

$$f^*_{A1B1}(t_i) = f_{A1B1}(t_i) - f_{A1B1}(t_{i-1})$$
$$f^*_{A2B2}(t_i) = f_{A2B2}(t_i) - f_{A2B2}(t_{i-1})$$

Once $f^*_{A1B1}(t_i)$ and $f^*_{A2B2}(t_i)$ for $0 < t_i \leqslant t$ are derived, they are used to build a population of 100,000 individuals which has a natural type I censoring rate equal to $F_P(t_5)$ (e.g. 0.6 in our simulation), we thus obtained four populations where the first competing risk was related to the epistatic interaction between SNP A_1–B_1 and the second competing risk was related to the epistatic interaction between SNP A_2–B_2. To each population, 21 unrelated SNPs in HWE, with MAF ranging from 0.1 to 0.5 were added (for a total of 25 SNPs per dataset). From the simulated populations we finally randomly draw 100 samples of $n = 400$ or $n = 600$ instances where 30% or 50% of the n instances were further randomly censored (type II censoring). A total of 1600 datasets per simulation were thus generated; overall 1600 datasets * 3 simulations (according to the different survival distribution used) = 4800 datasets were built.

SDR-CR with 5-fold cross-validation was run on each dataset under the null hypothesis of no association between the causal pairs and the cause-specific outcome; to this end a 100-fold permutation test was used to set the nominal type I error rates to 0.05. From each replication set of 100 datasets/model/degree-of-censorship/n of instances, we then calculated the power of SDR-CR, defined as the proportion of simulated samples/set of which the first causal pair (SNP A_1–B_1) was selected as the best model with a significance < 0.05 for the first outcome and, at the same time, the second causal pair (SNP A_2–B_2) was selected as the best model with a significance < 0.05 for the second outcome.

Finally, to better gauge the magnitude of a correct handling of competing risks over the naïve application of the KM algorithm when competing risks are ignored, we also run a 5-fold experiment with the conventional SDR algorithm on the simulated datasets, treating the alternative outcomes as censored cases and setting the nominal error rates as above.

Datasets were built using the 2way_EpiComp_Generator tool written in python and available at, http://sourceforge.net/projects/sdrproject/files/. Details about the penetrance functions used for simulation are reported in Appendix A; a step-by-step example of the processes involved in the generation of the synthetic datasets via the simulation tool is described in Appendix B.

All the calculations were made using the SDR_V2.0b algorithm written in python with C extensions [22] and available at the Sourceforge site as outlined beforehand.

2.5. Application of the SDR-CR algorithm in systemic sclerosis (SSc) lung dataset

To illustrate the applicability of SDR-CR in a real-world setting we used a dataset of 210 SSc patients referring to our outpatient clinic. All the patients from referral undergo a thorough evaluation with a twice-a-year execution of pulmonary function testings (PFTs) and ecocardiography and, when required, high-resolution computed tomography (HRCT) or right-heart catheterization (RHC). It is thus possible to retrospectively collect observational data about lung involvement in our case-series of SSc patients. SSc-related lung involvement [23] may either present as: (a) interstitial lung disease (ILD), defined as a forced vital capacity (FVC) on PFTs < 70% of predicted values + typical appearances on HRCT extending at least up to the pulmonary venous confluence [24] and involving at least 5% of the parenchyma; or (b) pulmonary arterial hypertension (PAH), defined as an increased mean pulmonary artery pressure (mPAP) above 25 mmHg at rest on RHC with a pulmonary wedge pressure < 15 mmHg and no signs of ILD as defined above [23,25]. Thus, ILD and PAH constitute the typical competitive risk setting, being both cause of failure mutually exclusive. In our population we considered the occurrence of either outcome within 10 years from referral.

A large number of our patients underwent a program of DNA extraction and genotyping as a part of a European genetic program [26]; genotyping for a number of single-nucleotide polymorphisms (SNPs) within genes for cytokines with pro-inflammatory, profibrotic and regulatory functions on the immune system, as described elsewhere [27], is available. Overall observational and genotypic data for 17 SNPs are available for 210 patients; genotyping details are described in Beretta et al. [27] the list of the studied SNPs is as follows: IL-1α C-889T, IL-1β C-511T, IL-1β C + 3962T, IL-1R Cpst1970T, IL-1Ra Cmspal11100T, IL-2 G-330T, IL-2 G + 160T, IL-4Rα A + 1902G, IL-6 C-174G, IL-6 Ant565G, IL-10 A-1082G, IL-12 A-1188C, TGF-β1 T/C codon 10, TGF-β1 G/C codon 25, IFNγ AUTR5644T, TNFα A-308G, TNFα A-238G.

We run the SDR-CR algorithm on the SSc-lung dataset (up to 3-way interactions) with a 10-fold cross-validation searching strategy and 1000-fold permutation testing as described above. SDR-CR was compared to plain SDR with the same settings where the competing outcomes were alternatively censored.

3. Results

3.1. Simulation study

The power for the SDR-CR algorithm to identify both the causative pairs of SNPs in the simulated datasets with nominal type error I rate = 0.05, is reported in Table 1. The shape of the underlying cause-specific incidence function apparently does not affect the power of the SDR-algorithm, even if a slight increase in the detection rate can be observed in models involving a BT distribution (simulation 2 and 3). This is most likely due to sharper increase in the hazards for the BT-shaped risk (see Fig. 1), and consequently in a higher genetic contribution (e.g. cumulative heritability) into early time-points. As expected, the power of the algorithm is impaired by the increase in type II censor rate and is recovered when the sample size increases.

When competing risks are ignored and treated as censored cases, a marked reduction in the power to contemporaneously detect both the causative pairs of SNP can be observed, with a reduction as high as 60% in models with low $H^2(5)$. This reduction is less pronounced models with high $H^2(5)$, with a loss of power no higher than 20%. Table 2 summarizes the power for the conventional SDR method and the relative decrease of power with respect to the SDR-CR algorithm, as described in Table 1.

3.2. SSc-lung dataset

The main characteristics of our SSc population were as follows: 191 patients were females (91.1%) and 46 patients (21.9%) had the diffuse cutaneous subset of the disease; 205 patients (97.6%) tested positive for antinuclear antibodies (ANAs) and specifically 84 (40%) were positive for anti-centromere antibodies, whilst 89 (42.4%) tested positive for anti-topoisomerase I antibodies. Overall type I censoring was equal to 36 cases (17.1%), 20 subjects (9.5%) experienced PAH during the course of the follow-up, whilst 47 (22.4%) developed ILD; in the dataset type II censoring accounted for 107 (50.1%) of cases.

SDR-CR identified the IL-10 A-1082G as the best predictor for PAH, however, this association was not significant after permutation testing ($p = 0.602$); conversely, a significant interaction between IL-1α C-889T and the IL-1β C-511T SNP was found to be associated with the occurrence of ILD, with a IBS = 0.1596 and a permutation p value equal to 0.029. As outlined in Fig. 2, panel a, the interaction between these SNPs shows the typical non-linear or epistatic pattern. Fig. 2, panel b depicts the cumulative incidences of ILD for the pooled high- and low-risk cells.

The naïve application of the SDR algorithm to the SSc-lung dataset, treating either outcomes as censored cases, failed to find any significant association with PAH or ILD.

Table 1
Power for the survival dimensionality reduction algorithm for competing risks (SDR-CR) in simulated datasets.

$H^2(5)$	C (%)	$n = 400$			$n = 600$		
		EXP–EXP	EXP–BT	BT–IFR	EXP–EXP	EXP–BT	BT–IFR
0.075	30	52	56	58	61	62	60
	50	36	41	37	46	49	48
0.1	30	73	75	74	81	82	81
	50	58	62	60	67	71	66
0.125	30	93	95	94	95	99	96
	50	84	83	84	91	93	92
0.15	30	99	100	100	100	100	100
	50	99	100	99	100	100	100

Power for the SDR-CR algorithm under different scenarios in synthetic epistatic lifetime datasets with competing risks; power calculated after 5-fold cross-validation and 100-fold permutation test to set a nominal type I error rate = 0.05. $H^2(5)$, cumulative heritability at the survival time t, where $t = 5$; C, degree of type II censorship; n, datasets size. Shapes of the cause-specific cumulative incidence functions: EXP, exponential; BT, bath-tube failure rate; IFR, increasing failure rate.

3.3. Computational cost

To determine the computational demand of the SDR-CR algorithm we run a 5-fold cross-validation experiment on datasets with two competing causes of failure, 10 distinct time-points 10, 20, 50 or 100 SNPs and 500, 1000 or 2000 instances. On a Intel® Core™ i7 CPU Q740 @ 1.73 GHz and 4 Gb RAM, the CPU time necessary to run a 2-fold experiment is described by the following exponential equations: 500 instances experiment, CPU time = $0.0109 * (n \text{ of SNPs})^{2.0703}$; 1000 instances experiment, CPU time = $0.0141 * (n \text{ of SNPs})^{2.0703}$; 2000 instances experiment, CPU time = $0.0206 * (n \text{ of SNPs})^{2.0703}$. Thus, for instance a pairwise interaction in a 500 instances dataset with 50 SNPs would require approximately 35 s.

4. Discussion

The comprehension of complex human diseases requires the development and the availability of adequate investigational tools capable of analyzing their intricate architecture. In the study of chronic diseases, time represents a variable of primary interest to evaluate the relationship between attributes and outcomes, especially when the occurrence of a certain event is scattered in the time-course of the illness. Thus, to avoid the possibility to draw wrong conclusions when censoring is ignored, specific statistical algorithms have to be applied. SDR is an analytical approach conceived in the effort to unveil another layer of complexity of human disease: the non-linear interaction among genes in time-dependent contexts [10]. In the present work we further push forward this potential taking into account the not unusual situation where the endpoint consists of several mutually exclusive events of interest, which defines the 'competing risks model'. We demonstrated in synthetic epistatic lifetime datasets the importance to properly handle competing risks, as, indeed, when these are ignored and naively treated as censored cases, the predictive capability of SDR is greatly impaired. The better performance of SDR-CR compared to the naïve SDR in the competing risk analysis mirrors previous observations made in the context of univariate or linear interaction analysis [12,14,15].

Herein we also showed that SDR-CR can be fruitfully applied in the real-world, describing an epistatic interaction that is significantly associated with the occurrence of ILD in a population of scleroderma patients. In accordance with simulation results, this association was overlooked when competing risks were ignored. SDR-CR, similarly to other algorithms aiming at discovering epistatic relations, do not provide clues about the precise mechanism by which significant genetic variations do interact at the cellular or molecular level [28]. Therefore, it would be speculative to hypothesize the means by which the joint effect of the IL-1α C-889T and the IL-1β C-511T SNPs would promote or sustain the development of ILD in SSc subjects. Of interest, increased concentrations of IL-1β in bronchoalveolar lavage fluid from SSc patients compared to controls and a negative correlation between IL-1β and FVC, have been reported by Hussein et al. [29].

Table 2
Power for the naïve survival dimensionality reduction algorithm ignoring competing risks and relative change vs. the SDR-CR algorithm risks (SDR-CR) in simulated datasets.

$H^2(5)$	C (%)	n = 400			n = 600		
		EXP–EXP	EXP–BT	BT–IFR	EXP–EXP	EXP–BT	BT–IFR
0.075	30	17 (−67.3%)	20 (−64.3%)	19 (−67.2%)	24 (−60.7%)	25 (−59.7%)	24 (−60%)
	50	10 (−72.2%)	13 (−68.3%)	12 (−67.6%)	19 (−58.7%)	21 (−57.1%)	19 (−60.4%)
0.1	30	41 (−43.8%)	42 (−44%)	41 (−44.6%)	48 (−40.7%)	50 (−39%)	49 (−39.5%)
	50	30 (−48.3%)	32 (−48.4%)	32 (−46.7%)	41 (−38.8%)	43 (−39.4%)	43 (−34.8%)
0.125	30	51 (−45.2%)	52 (−45.3%)	51 (−45.7%)	55 (−42.1%)	57 (−42.4%)	54 (−43.8%)
	50	42 (−50%)	44 (−47%)	43 (−48.8%)	51 (−44%)	55 (−40.9%)	52 (−43.5%)
0.15	30	66 (−33.3%)	68 (−32%)	66 (−34%)	73 (−27%)	75 (−25%)	72 (−28%)
	50	57 (−42.4%)	59 (−41%)	58 (−41.4%)	63 (−37%)	66 (−34%)	63 (−37%)

Power for the naïve SDR algorithm to detect both the couples of causative SNPS under different scenarios in synthetic epistatic lifetime datasets with competing risks and relative change in power (in brackets) compared to the SDR-CR method (see Table 1). Competing risks are handled as censored cases. For legend see Table 1.

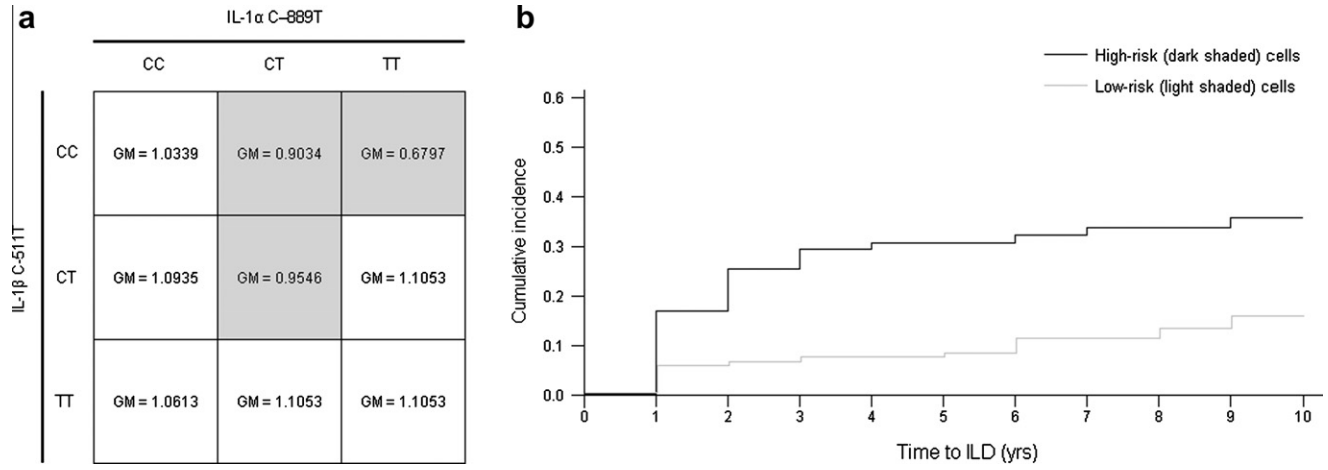

Fig. 2. Best epistatic model for interstitial lung disease (ILD) in the systemic sclerosis (SSc)-lung dataset. Results from SDR-CR analysis in the SSc-lung datasets. (a) Pattern of interaction between the IL-1α C-889T and the IL-1β C-511T single nucleotide polymorphisms (SNP) to explain the occurrence of ILD. The scattered distribution of high-risk cells (geometric mean [GM] < 1) and low-risk cells is indicative of non-linear interaction or epistasis. (b) Cumulative incidence of ILD in patients classified as high-risk and low-risk by the SDR-CR algorithm. No results are provided for the competing event (e.g. pulmonary arterial hypertension), which was not significantly associated with any SNP or SNP–SNP interaction.

To our knowledge, SDR-CR is the first data-mining method capable of handling non-linear interactions in lifetime competing risk models. From the simulation we conducted we can conclude that SDR-CR is suitable for this kind of analysis in candidate gene studies and in small-to-medium-size datasets. We indeed observed a satisfactory power also in situations with low number of instances (e.g. 200 per outcome), low heritability and high rates of lost-to follow-up observations (e.g. type II censoring) which pose a number of not easily solvable detection challenges. Conversely, we cannot make inferences about SDR-CR predictive ability in large-scale datasets; even if previous studies have shown that multifactor dimensionality reduction, the case-control inspiring counterpart of SDR [10], is relatively insensitive to background noise [30], this property could not directly translated to SDR-CR and it would require an adequate simulation study. Moreover the SDR-CR algorithm is computationally demanding and an exhaustive analysis in large-scale datasets may be unfeasible due to the large number of interactions to test and to the time that this analysis would require. The test experiment we conducted indicates, for instance, a CPU time of 9.5 h to analyze via a 5-fold cross-validation experiment a pairwise interaction in a 2000 instance dataset with 1000 SNPs and five time units.

Overall, SDR-CR strengths and weaknesses are largely similar to those we previously described for SDR [10] among the former, besides power, we can list, the fully non-parametric nature of the algorithm, the small chance to describe false positive results due to the cross-validation procedure (that can also be complemented by permutation testing); among the latter, we remember the difficulty to interpreter the results at the biological level as well as the possibility that the performance of the algorithm may to some extent be dampened in presence of heterogeneity [30].

The current version of SDR-CR as well as its precursor SDR, cannot handle covariates or model interactions for quantitative trait loci. A straightforward extension to tackle these issues would be to estimate the population and the multilocus cell survival functions via the Cox regression method (for SDR) or to model the hazards of the subdistributions (for SDR-CR) according to Fine and Gray [31]. High-risk and low-risk assignments would then be performed via the usual SDR procedure or could be accomplished via a parametric estimation, similarly to the method outlined by Calle et al. [32].

5. Conclusions

Summarizing, herein we presented an extension of the SDR algorithm to analyze competing risks models (SDR-CR). Simulation studies let us think that this approach may be fruitfully used in the analysis of genetic lifetime datasets with mutually-exclusive outcomes, overcoming the limitations and the remarkable loss of power observed when data are not properly handled.

Appendix A. Supplementary material

Supplementary data associated with this article (cause-specific multilocus prevalences; generation of simulated epistatic lifetime

dataset) can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2012.11.002.

References

[1] So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 2011;35:310–7.
[2] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–53.
[3] Clarke AJ, Cooper DN. GWAS: heritability missing in action? Eur J Hum Genet 2010;18:859–61.
[4] Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 2010;11:446–50.
[5] Thornton-Wells TA, Moore JH, Haines JL. Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. BMC Bioinform 2006;7:204–21.
[6] Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56:73–82.
[7] Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet 2009;10:392–404.
[8] McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene–gene interactions: a review. Appl Bioinform 2006;5:77–88.
[9] Van Steen K. Travelling the world of gene–gene interactions. Brief Bioinform; 2011 [March 26].
[10] Beretta L, Santaniello A, van Riel PL, Coenen MJ, Scorza R. Survival dimensionality reduction (SDR): development and clinical application of an innovative approach to detect epistasis in presence of right-censored data. BMC Bioinform 2010;11:416.
[11] Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, Andrew AS. A novel survival multifactor dimensionality reduction method for detecting gene–gene interactions with application to bladder cancer prognosis. Hum Genet 2011;129:101–10.
[12] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med 2007;26:2389–430.
[13] Bakoyannis G, Touloumi G. Practical methods for competing risks data: a review. Stat Methods Med Res 2012;21:257–72.
[14] Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med 1999;18:695–706.
[15] Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res 2007;13:559–65.
[16] Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J 2011;53:88–112.
[17] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999;18:2529–45.
[18] Simon RM, Subramanian J, Li MC, Menezes S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. Brief Bioinform 2011;12:203–14.
[19] Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. Stat Med 2009;28:956–71.
[20] Leemis LM, Lam Y. The logistic-exponential survival distribution. Nav Res Log 2008;55:252–64.
[21] Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 2002;70:461–71.
[22] Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS, Smith K. Cython: the best of both worlds. Comput Sci Eng 2011;13:31–9.
[23] Wells AU, Steen V, Valentini G. Pulmonary complications: one of the most challenging complications of systemic sclerosis. Rheumatology 2009;48(iii):40–4 [Oxford].
[24] Goh NS, Desai SR, Veeraraghavan S, Hansell DM, Copley SJ, Maher TM, et al. Interstitial lung disease in systemic sclerosis: a simple staging system. Am J Respir Crit Care Med 2008;177:1248–54.
[25] Galiè N, Hoeper MM, Humbert M, Torbicki A, Vachiery JL, Barbera JA, et al. Guidelines for the diagnosis and treatment of pulmonary hypertension: the task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS), endorsed by the International Society of heart and Lung Transplantation (ISHLT). Eur Heart J 2009;30:2493–537.
[26] Martin JE, Broen JC, Carmona FD, Teruel M, Simeon CP, Vonk MC, et al. Identification of CSK as a systemic sclerosis genetic risk factor through genome wide association study follow-up. Hum Mol Genet; 2012 March 22 [Epub ahead of print].
[27] Beretta L, Cappiello F, Moore JH, Barili M, Greene CS, Scorza R. Ability of epistatic interactions of cytokine single-nucleotide polymorphisms to predict susceptibility to disease subsets in systemic sclerosis patients. Arthritis Rheum 2008;59:974–83.
[28] Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 2005;27:637–46.
[29] Hussein MR, Hassan HI, Hofny ER, Elkholy M, Fatehy NA, Abd Elmoniem AE. Alterations of mononuclear inflammatory cells, CD4/CD8 + T cells, interleukin 1beta, and tumour necrosis factor alpha in the bronchoalveolar lavage fluid, peripheral blood, and skin of patients with systemic sclerosis. J Clin Pathol 2005;58:178–84.
[30] Edwards TL, Lewis K, Velez DR, Dudek S, Ritchie MD. Exploring the performance of multifactor dimensionality reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models. Hum Hered 2009;67:183–92.
[31] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. JASA 1999;94:496–509.
[32] Calle ML, Urrea V, Vellalta G, Malats N, Steen KV. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. Stat Med 2008;27:6532–46.