Simulation des données familiales avec le package simufam2

Emilie ADZIMASE

2025-06-17

Contents

Introduction

Le package *simufam2* permet de simuler des structures familiales pour étudier la pénétrance de maladies génétiques (notamment les cancers héréditaires), de tester des stratégies de sélection de familles ou d'individus, et de constituer des cohortes prospectives ou rétrospectives dans un cadre épidémiologique. Ce document présente quelques fonctions importantes de ce package essentiel à la simulation des données familiales.

Chargement des bibliothèques

Le package peut s'installer comme suit:

```
library(simufam2)
library(dplyr)
library(ggplot2)
```

1. Définition des paramètres de simulation

Pour simuler les données, nous avons besoin de plusieurs paramètres comme la taille de l'échantillon de famille que nous considerons, les fonctions de pénétrance pour chez les individus porteurs et non porteurs de la mutation.

1.1. Simulation de la fonction de pénétrance chez les porteurs de la mutation

Nous supposons un risque sur la vie (120 ans) de 60% comme dans le cas du cancer colorectal chez les femmes avec comme gène responsable le MH1 ou le MHS2. Nous simulons ensuite le risque de cancer cumulé à chaque age (0-120 ans) à partir d'une fonction Weibull. La fonction utilisée est `penetrance.Weibull` de notre package.

```
# Simulation de la fonction de pénétrance chez les porteurs
# Ici, on suppose un risque cumulé de 60 % à 120 ans,
# avec H1 = proportion de non-malades à t1 parmi ceux à risque
# et H2 = proportion de non-malades à t2 parmi ceux non-malades à t1 mais encore à risque à t2
risk_mute_simu <- simufam2::penetrance.Weibull(d = 3/5, H1 = 3/4, H2 = 1/3)
```

La fonction `penetrance.Weibull(d, H1, H2)` génère une courbe de pénétrance (cumulative) sur la vie (jusqu'à 120 ans), en s'appuyant sur une modélisation de type Weibull. Elle prend trois paramètres indépendants :

- **d** est la pénétrance sur la vie (120 ans)
- **H1** est la proportion d'individus non atteints à *t1* années parmi ceux à risque
- **H2** est la proportion d'individus non atteints à *t2* années parmi ceux qui étaient non atteints à *t1* mais toujours à risque à *t2*.

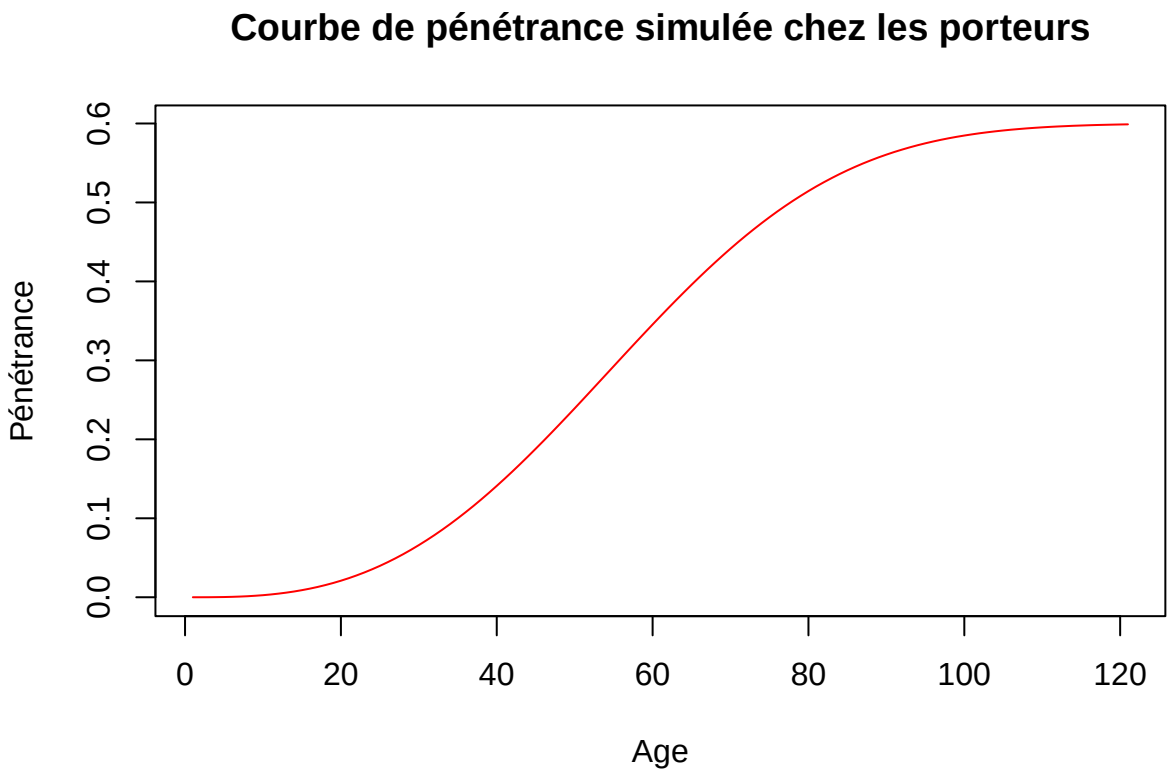
Ces trois paramètres sont indépendants les uns des autres et sont définis entre 0 et 1.

```
#Affichage des premières lignes
head(risk_mute_simu)
```

```
##            0            1            2            3            4            5
## 0.000000e+00 5.435452e-06 3.811757e-05 1.191013e-04 2.672660e-04 5.002404e-04
```

Représentation graphique de la pénétrance simulée chez les porteurs de la mutation

```
plot(risk_mute_simu, col = "red",
     xlab = "Age",
     ylab = "Pénétrance",
     main = "Courbe de pénétrance simulée chez les porteurs",
     type = "l",
     cex = 0.8)
```



1.2. Simulation de la fonction de pénétrance chez les non-porteurs de la mutation

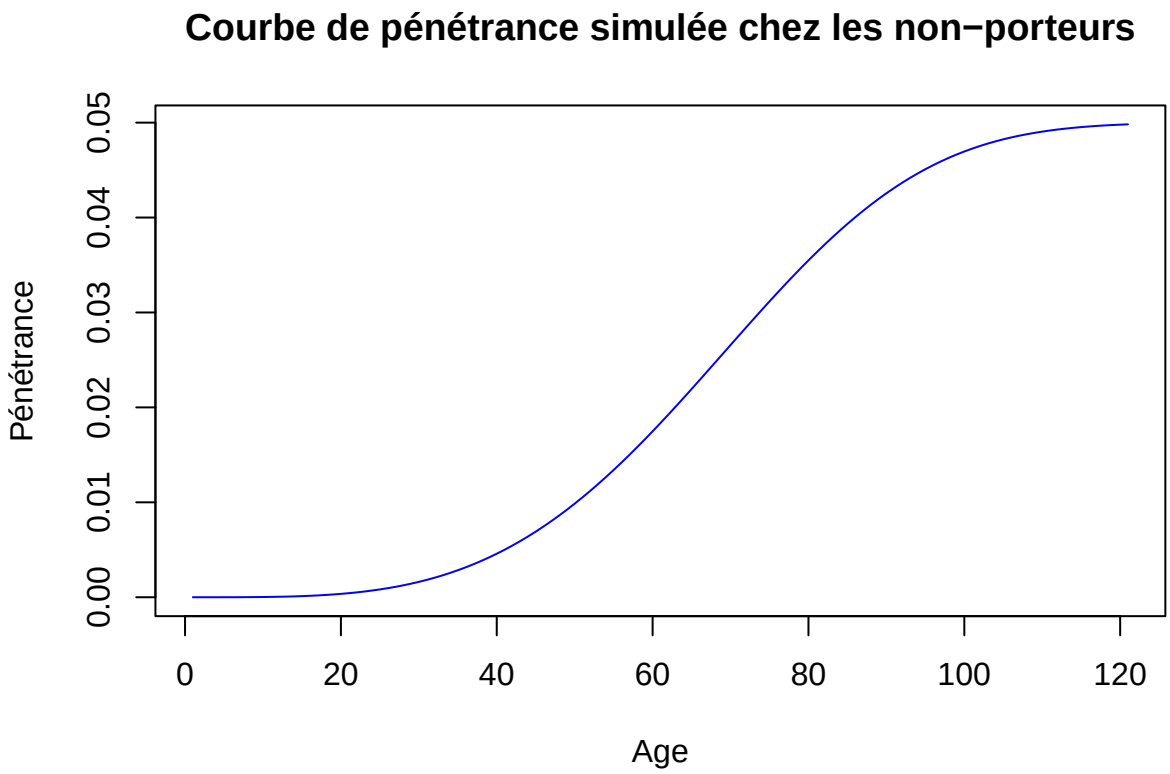
De la même manière, nous allons simuler la pénétrance chez les non porteurs de la mutation. Nous avons supposé que cette dernière est de 5%. La simulation se fera aussi avec la fonction `penetrance.Weibull` du package `simufam2`.

```
#Pénétrance chez les non-porteurs de la mutation
risk_nmute_simu <-  simufam2::penetrance.Weibull(d = 5/100, H1 = 9/10, H2 = 1/2)
#Affichage des premières lignes
head(risk_nmute_simu)
```

```
##            0           1           2           3           4           5
## 0.000000e+00 8.384672e-09 1.030317e-07 4.469689e-07 1.266050e-06 2.839096e-06
```

Représentation graphique de la pénétrance simulée chez les non-porteurs de la mutation

```
plot(risk_nmute_simu, col = "blue",
     xlab = "Age",
     ylab = "Pénétrance",
     main = "Courbe de pénétrance simulée chez les non-porteurs",
     type = "l",
     cex = 0.8)
```



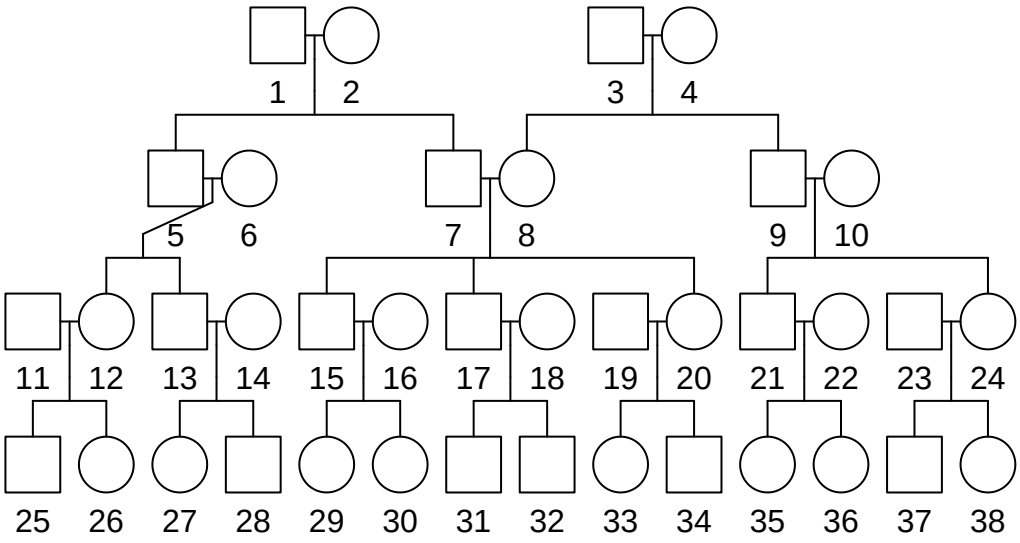
2. Simulation des données de familles

Nous allons simuler un échantillon de 100 000 familles avec un risque chez les porteurs de mutation à 60% et 5% chez les non porteurs. Nous supposons aussi une fréquence allélique de 0.001 pour cette simulation.

Chaque famille est générée avec une structure généalogique standard que nous propose Bonaïti et al 2011 https://pmc.ncbi.nlm.nih.gov/articles/PMC3025788/. L'individu id 17 est le cas index.

```
ped2 <- kinship2::pedigree(id = ped$id,
                           dadid = ped$dadid,
                           momid = ped$momid,
                           sex = ped$sex)

plot(ped2)
```




```
#Simulation des familles
fam <-  simufam2::simul_families(fA = 0.001,
                                 n = 10e4,
                                 risk_mute = risk_mute_simu,
                                 risk_nmute = risk_nmute_simu)

#Affichage des premières lignes
print(fam, width = Inf)
```

```
## # A tibble: 3,800,000 x 13
##    id_family genotype    id dadid momid sex    generation kin_id17 cod
##        <int>    <dbl> <dbl> <dbl> <dbl> <chr>       <dbl>    <dbl> <chr>
##  1         1        0     1     0     0 Male           -2     0.25 -2_0.25
##  2         1        0     2     0     0 Female         -2     0.25 -2_0.25
##  3         1        0     3     0     0 Male           -2     0.25 -2_0.25
##  4         1        1     4     0     0 Female         -2     0.25 -2_0.25
##  5         1        0     5     1     2 Male           -1     0.25 -1_0.25
##  6         1        0     6     0     0 Female         -1     0    -1_0
##  7         1        0     7     1     2 Male           -1     0.5  -1_0.5
##  8         1        1     8     3     4 Female         -1     0.5  -1_0.5
##  9         1        0     9     3     4 Male           -1     0.25 -1_0.25
## 10         1        0    10     0     0 Female         -1     0    -1_0
##    age_ddn tested age_cancer Phenotype
##      <dbl>  <dbl>      <dbl>     <dbl>
##  1      74      0         NA         0
##  2      61      0         NA         0
```

```
##  3      82      0         NA         0
##  4      74      0         67         1
##  5      68      1         NA         0
##  6      70      0         NA         0
##  7      75      0         NA         0
##  8      64      0         58         1
##  9       7      0         NA         0
## 10      75      0         NA         0
## # i 3,799,990 more rows
```

- *id_family* : identifiant de la famille
- *genotype* : 1 si porteur de la mutation, 0 sinon
- *id, dadid, momid* : identifiants de l'individu, du père et de la mère
- *generation* : génération de l'individu (génération 0 est celui du cas index, id 17)
- *kin_id17* : Lien de paranté avec le cax index (id = 17)
- *age_ddn* : Age aux dernières nouvelles à partir de la distribution empirique dans la base OFELY
- *tested* : indicateurs du statut de test (individu testé ou pas)
- *age_cancer, Phenotype* : âge au cancer, phénotype (1 si l'individu a le cancer)

Temps de simulation de 100 000 familles :

```
system.time({
  families <- simufam2::simul_families(fA = 0.001,
                                       n = 1e5,
                                       risk_mute = risk_mute_simu,
                                       risk_nmute = risk_nmute_simu)
})
```

```
##    user  system elapsed
##    1.76    0.31    2.28
```

Cette commande montre un temps d'exécution de ~1.9 secondes pour la simulation de 100 000 familles, ce qui reste raisonnable pour une simulation de grande ampleur.

3. Filter les familles ayant un nombre minimal de cas de cancer donné

Dans le cadre de l'estimation de la pénétrance, il est souvent pertinent de restreindre l'analyse aux familles les plus informatives, c'est-à-dire celles comptant un certain nombre minimal de cas de cancer. La fonction `filter_byNMA()` permet de filtrer l'échantillon simulé en ne conservant que les familles comportant au moins un nombre spécifié d'individus atteints.

```
fam2 <- simufam2::filter_fam_byMNA(families = fam ,nma = 3)
```

```
## nb fam =  47508
```

4. Constitution d'une cohorte prospective

La fonction `select_prospective()` permet d'extraire une cohorte prospective à partir des familles simulées. On sélectionne ici les porteurs de la mutation et indemnes de cancer à l'inclusion *(t = 0)* et on les suit pendant 10 ans *(t = 10)*. Le but est de simuler une étude de suivi permettant d'estimer la pénétrance par méthode de Kaplan-Meier ou Cox.

```
cohort <- simufam2::select_prospective(families = families ,t = 10)
```

```
print(cohort, width = Inf)
```

```
## # A tibble: 174,383 x 17
##    id_family genotype    id dadid momid sex    generation kin_id17 cod
##        <int>    <dbl> <dbl> <dbl> <dbl> <chr>       <dbl>    <dbl> <chr>
##  1         1        1     1    17     7     8 Male            0   1      0_1
##  2         1        1     1    31    17    18 Male            1   0.5    1_0.5
##  3         2        1     1    15     7     8 Male            0   0.5    0_0.5
##  4         2        1     1    17     7     8 Male            0   1      0_1
##  5         2        1     1    25    11    12 Male            1   0.0625 1_0.0625
##  6         3        1     1    15     7     8 Male            0   0.5    0_0.5
##  7         3        1     1    17     7     8 Male            0   1      0_1
##  8         3        1     1    32    17    18 Male            1   0.5    1_0.5
##  9         4        1     1    17     7     8 Male            0   1      0_1
## 10         4        1     1    20     7     8 Female          0   0.5    0_0.5
##    age_ddn tested age_cancer Phenotype age_fsuivi age_deces event fin_suivi
##      <dbl>  <dbl>      <dbl>     <dbl>      <dbl>     <dbl> <dbl>     <dbl>
##  1      52      1         63         0         62        88     0        62
##  2      54      1         NA         0         64        93     0        64
##  3      59      1         93         0         69        72     0        69
##  4      44      1         NA         0         54        86     0        54
##  5      25      1         NA         0         35        95     0        35
##  6      43      1         84         0         53       100     0        53
##  7      70      1         NA         0         80        73     0        73
##  8       1      1         27         0         11        87     0        11
##  9      50      1         NA         0         60        91     0        60
## 10      48      1         61         0         58        94     0        58
## # i 174,373 more rows
```

- *age_fsuivi*: Age à la fin de suivi
- *age_deces*: Age de decès en utilisant les tables de mortalité de l'INSEE, 2022
- *event*: 1 si l'individu a un cancer sur la période de suivi
- *fin_suivi*: Âge minimal entre l'âge au cancer, l'âge au décès et l'âge à la fin du suivi.