**SUPPORTING MATERIAL**

**Random Forest method**

The study presented here was carried out by performing three sets of experiments.

- the traditional randomForest algorithm of Breiman et al., based on the rpart decision tree algorithm [18]

- the implemented randomForest algorithm of Hothorn et al. based on the cpart decision tree algorithm [19] in its classical version (Out-Of-Bag, OOB with resampling)

- the implemented randomForest algorithm of Hothorn et al based on the cpart decision tree algorithm in its unbiased version (Out-Of-Bag, OOB without resampling)

Rpart and cpart are implemented in the package randomForest [20] and party [21,22], respectively, both available in the R software for statistical computing [24].

The application of different algorithms and different criteria of OOB sampling should lead to stable results, unaffected by the method and providing convergent results.

Both *randomForest* and *cforest* enable the importance of the variables to be measured against their ability to predict the properties investigated, i.e. the correct classification of each RD Registry into its respective category. The number of candidate variables considered at each split was chosen according to the values calculated using the *tuneRF* function of the *randomForest* package. For both algorithms, the number of trees was set to 1000 and ten seed numbers were used for each mtry value considered. This approach, based on multiple runs, is important to ensure the robustness and stability of the results.

The predictive power of the models, namely their ability to correctly classify the Registries, which from time to time are randomly chosen in the OOB sample, was evaluated using true positive (TP) rate (for each Class), accuracy, and weighted average F-measure, which is a combined metric that takes into account precision and recall (or true positive rate). These values were calculated according to the following equations:

TP rate for treatment Registries (TP-TR) = $\frac{numb. of TR correctly predicted}{total number of TR}$

TP rate for public-health Registries (TP-PHR) = $\frac{numb. of PHR correctly predicted}{total number of PHR}$

TP rate for clinical/genetic research Registries (TP-CGRR) = $\frac{numb. of CGRR correctly predicted}{total number of CGRR}$

Accuracy = $\frac{numb. of registries correctly predicted}{total number of registries}$

F-measure = $\frac{2\*precision\*TP rate}{precision+TP rate}$

Weighted average F-measure =
$$\left(\frac{numb. of TR correctly predicted}{total number of registries}\*Fmeasure\_{TR}\right)+\left(\frac{numb. of PHR correctly predicted}{total number of registries}\*Fmeasure\_{PHR}\right)+\left(\frac{numb. of CGRR correctly predicted}{total number of registries}\*Fmeasure\_{CGRR}\right)$$

From the best performing models of each method, the importance of the variables was extrapolated using a permutation method (variable permutation importance), according to which, a decrease in the accuracy of the prediction of the OOB samples following the permutation of a variable, indicates that that variable is important for the correct prediction of the Class. Variable importance was calculated with the functions *importance* (*randomForest* package) and *varimp* (*party* package). Variables were considered important and informative if their importance value was above the absolute value of the lowest negative-scoring variable [23]. All variables whose importance value was negative, zero or had a small positive value that lies in the same range as the negative values, were thus excluded from further exploration.

**Criteria for the selection of the best performing models**

The *importance* (for *randomForest*) and *varimp* (for *cforest*) functions were used at each step to define a subset of informative variables used for the development of models in the subsequent step.

In the first step, models that showed the best two values for TP-TR, TP-PHR, TP-CGRR, and accuracy, were considered as the best performing (Table SM2).

With the following steps, models were selected if one of the following criteria were satisfied with respect to models of the previous steps:

- models with the highest accuracy value

- models with the highest TP rates

- models with same accuracy or TP rates as those obtained in one of the previous steps, but obtained on the basis of a reduced number of variables.

Variables importance values were calculated for each of the models obtained in the first step. Dozens of new models were thereby developed using reduced subsets of variables selected from among the 272 considered in the first step; using a reduced number of variables should lead to a decreasing error of prediction on OOB sample.

The protocol used in the following steps was similar to that used previously: ten seed numbers were used for each *mtry* value considered and previously chosen with the *tuneRF* function, as described before.

**TABLES**

**Table SM1.** Question 16 is a representative example for multiple choice questions. The question was split into several questions and each corresponding answer was labeled as 1 or 0 if selected or not selected, respectively.

|  |
| --- |
| **Question 16 (Q16): “What kind of data are collected?”** |
| **Answers reported in the survey** | **New question ID** | **Answer labels** |
| Anagraphical data | Q16\_1 | Selected=1Not selected=0 |
| Diagnosis | Q16\_2 |
| Anthropometric information | Q16\_3 |
| Socio-demographic Information | Q16\_4 |
| Genetic data | Q16\_5 |
| Clinical data | Q16\_6 |
| Medications, devices and health services | Q16\_7 |
| Patient-reported outcomes | Q16\_8 |
| Family history | Q16\_9 |
| Birth and reproductive history | Q16\_10 |
| Clinical research participation and bio-specimen donation | Q16\_11 |
| Patient’s preferences for communication | Q16\_12 |

**Table SM2.** Best performing models obtained with *randomForest* and *cforest* in the first step of experiments. *Mtry* indicates the number of variables randomly selected at each node, among the 272 available; TP-TR, TP-PHR, and TP-CGRR is the True Positive rate for the classes of Treatment, Public Health, and Clinical/Genetic Research Registries, respectively. Models were ordered according to the decreasing value of the weighted average F-measure (F-measure).

|  |
| --- |
| **randomForest** |
| **Model** | **Variables** | **Seed** | **Mtry** | **TP-TR** | **TP-PHR** | **TP-CGRR** | **Accuracy** | **F-measure** |
| Rf1 | 272 | 7 | 13 | 0.753 | 0.596 | 0.709 | 0.699 | 0.6986 |
| Rf2 | 272 | 8 | 16 | 0.704 | 0.577 | 0.756 | 0.694 | 0.6926 |
| Rf3 | 272 | 1 | 13 | 0.728 | 0.558 | 0.744 | 0.694 | 0.6922 |
| Rf4 | 272 | 4 | 12 | 0.716 | 0.577 | 0.733 | 0.690 | 0.6891 |
| Rf5 | 272 | 5 | 8 | 0.728 | 0.577 | 0.721 | 0.690 | 0.6886 |
| Rf6 | 272 | 70 | 41 | 0.741 | 0.577 | 0.698 | 0.685 | 0.6849 |
| **Cforest\_classic** |
| **Model** | **Variables** | **Seed** | **Mtry** | **TP-TR** | **TP-PHR** | **TP-CGRR** | **Accuracy** | **F-measure** |
| Cf1 | 272 | 9 | 16 | 0.617 | 0.500 | 0.802 | 0.662 | 0.6584 |
| Cf2 | 272 | 20 | 16 | 0.605 | 0.538 | 0.779 | 0.658 | 0.6562 |
| Cf3 | 272 | 100 | 16 | 0.630 | 0.500 | 0.733 | 0.639 | 0.6383 |
| Cf4 | 272 | 6 | 16 | 0.556 | 0.519 | 0.791 | 0.639 | 0.6358 |
| Cf5 | 272 | 30 | 8 | 0.543 | 0.500 | 0.814 | 0.639 | 0.6344 |
| **Cforest\_unbiased** |
| **Model** | **Variables** | **Seed** | **Mtry** | **TP-TR** | **TP-PHR** | **TP-CGRR** | **Accuracy** | **F-measure** |
| Cf1ub | 272 | 6 | 32 | 0.642 | 0.423 | 0.744 | 0.630 | 0.6252 |
| Cf2ub | 272 | 5 | 16 | 0.593 | 0.404 | 0.791 | 0.626 | 0.6176 |
| Cf3ub | 272 | 2 | 16 | 0.556 | 0.442 | 0.779 | 0.616 | 0.6109 |

**Table SM3.** List of the most informative variables belonging to the overall set defined by the subsets onto which the best performing models obtained from random forest approach were developed. The variables belonging to the intersection set defined by the models, are reported in bold.

|  |  |
| --- | --- |
| **Variable ID** | **Question** |
| **q6** | How many active cases/patients are included in your Register? |
| q7 | How many new cases are entered in the last year? |
| q8 | How many rare diseases are included? |
| **q10** | Are your data used for pharmacovigilance? |
| **q11** | What is the geographical coverage of the register? |
| **q12** | What is the target population of the register? |
| **q14** | Are standardised inclusion/exclusion criteria defined for the RD cases? |
| q15\_1 | Is ORPHACODE disease coding system in use? |
| q15\_2 | Is MIM disease coding system in use? |
| **q15\_345** | Is ICD disease coding system in use? |
| **q16\_3** | Are anthropometric data collected? |
| **q16\_4** | Are socio-demographic data collected?  |
| **q16\_5** | Are genetic data collected? |
| **q16\_6** | Are clinical data collected? |
| **q16\_7** | Are Medications, devices and health services data collected? |
| **q16\_8** | Are Patient-reported outcomes (e.g. quality of life data, Health status, etc) reported? |
| **q16\_9** | Are family history data reported? |
| **q16\_10** | Are birth and reproductive history data reported? |
| **q16\_11** | Are clinical research participation and biospecimen donation data reported? |
| **q17** | Is the date of the patient death collected? |
| q19\_1 | Are Clinical Units one of data providers of the Register? |
| q19\_2 | Are Clinical Genetic Units one of data providers of the Register? |
| q19\_5 | Are Hospital Databases (Discharge registers) one of data providers of the Register? |
| **q19\_9** | Are mortality registers one of data providers of the Register? |
| q19\_12 | Are other registers one of data providers of the Register? |
| q20\_5 | Are data entered in the Register by download from other primary databases? |
| q22\_1 | Is information typed twice (manual control) in order to avoid data entry mistakes? |
| **q22\_2** | Does an internal program automatically check the type of response (automatic control) in order to avoid data entry mistakes? |
| q22\_4 | Is there a specific method to avoid data entry mistakes? |
| q24\_3 | Is the percentage of missing values and the completeness of information a quality indicator of the Register? |
| q30 | Are instructions for use of the register available? |
| q31\_5 | Is there a structured training for new users? |
| **q32** | What are the reasons for which register has been established? |
| q34 | Has the register protocol been approved by an Ethics Committee? |
| q35 | Do you ask for patient’s written and informed consent to include his/her identifiable data in the register? |
| **q36\_1** | Is the “Register’s general scope” provided on the patient information sheet? |
| **q36\_2** | Are Register’s specific research objectives provided on the patient information sheet? |
| **q36\_3** | Does the patient information sheet report that the register is part of a network? |
| q36\_4 | Does the patient information sheet report the type of coding and access to the register? |
| **q36\_7** | Is the “right to withdraw” provided on the patient information sheet? |
| q36\_8 | Does the patient information sheet report Information on data property rights? |
| **q36\_11** | Is the “possibility to be contacted for participating in clinical trials” provided on the patient information sheet? |
| q36\_12 | Does the patient information sheet report the name and contact of the data processor? |
| q37\_1 | Is the patient information sheet revised by a local ethics committee? |
| **q38** | Is the patient information sheet publicly available and easily accessible? |
| **q39\_1** | If a participant decides to withdraw from the Register, have data anonymised for future research (irreversible destruction of participant’s activity)? |
| **q39\_2** | If a participant decides to withdraw from the Register, may he/she withdraw the authorisation to future uses of the data (except already aggregated or published data)? |
| q39\_3 | If a participant decides to withdraw from the register, may he/she have data destroyed? |
| q39\_6 | If a participant decides to withdraw from the Register, withdrawal is not possible as this is a mandatory public health register |
| q42\_3 | Is one of the main function of the Governing Board “Database content, research objectives, epidemiology, biostatistics, etc.”? |
| q43 | Has the register other governing bodies? |
| **q44\_4\_1** | Has the Register governing bodies composed by internal members with the function of Communication with the funding source, health care providers, patients, etc? |
| **q45\_2** | Are Local Ethics Committee informed of the Register activities? |
| **q45\_3** | Are Hospitals/Centres of Expertise informed of the Register activities? |
| q45\_4 | Is University informed of the Register activities? |
| q45\_5 | Is Personal Data Protection Authority (or similar body) informed of the Register activities? |
| **q45\_8** | Are public health policy makers informed of the register activities? |
| q46\_3 | Does the Register communicate the activities by Newsletter? |
| q46\_6 | Does the Register communicate the activities by Communication to scientific meetings? |
| q47 | Has the register been acknowledged in any peer reviewed scientific publication? |
| q48\_1 | Is the Register collaborating or sharing data with other registries? |
| q48\_3 | Is the Register collaborating or sharing data with Centres of Expertise? |
| **q48\_4** | Is the register collaborating or sharing data with other registries, biobanks, centres of expertise? |
| q49 | Are data available outside the Register? |
| q51\_3\_2 | Are data transferrable worldwide as de-identified? |
| q53\_1 | Is the Register security ensured by Passwords periodically and systematically renewed? |
| **q54\_2** | Was the Register set up with funding coming from Regional Authority? |
| **q55\_2** | Is a Regional Authority funding the register today? |
| q55\_3 | Is the National Authority funding the register today? |
| q55\_4 | Is University/Research Institute funding the register today? |
| **q59\_4** | Is one of the main needs of the register “to motivate data providers”? |
| q60 | Do you find desirable that the European Commission (EC) draws new legislations to uniformly regulate registries across the EU? |
| **q61**/other | Do you expect your country to provide public funding for a centralised national register on RD? |
| **q64\_3** | Are “Model documents (e.g. Informed consent form)” one of the services that should be offered by a EU platform for registries? |
| q64\_5 | Is “Quality control systems, quality experts advice, etc.” one of the services that should be offered by a EU platform for Registries? |
| **q64\_7** | Is “Facilitated access to useful data sources” one of the services that should be offered by a EU platform for Registries? |