European Radiology Experimental

ORIGINAL ARTICLE Open Access

A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data

Carlos Baeza-Delgado[1], Leonor Cerdá Alberich[1], José Miguel Carot-Sierra[2], Diana Veiga-Canuto[3], Blanca Martínez de las Heras[4], Ben Raza[1,4] and Luis Martí-Bonmatí[1*]

Abstract

Background: Estimating the required sample size is crucial when developing and validating clinical prediction models. However, there is no consensus about how to determine the sample size in such a setting. Here, the goal was to compare available methods to define a practical solution to sample size estimation for clinical predictive models, as applied to Horizon 2020 PRIMAGE as a case study.

Methods: Three different methods (Riley's; "rule of thumb" with 10 and 5 events per predictor) were employed to calculate the sample size required to develop predictive models to analyse the variation in sample size as a function of different parameters. Subsequently, the sample size for model validation was also estimated.

Results: To develop reliable predictive models, 1397 neuroblastoma patients are required, 1060 high-risk neuroblastoma patients and 1345 diffuse intrinsic pontine glioma (DIPG) patients. This sample size can be lowered by reducing the number of variables included in the model, by including direct measures of the outcome to be predicted and/or by increasing the follow-up period. For model validation, the estimated sample size resulted to be 326 patients for neuroblastoma, 246 for high-risk neuroblastoma, and 592 for DIPG.

Conclusions: Given the variability of the different sample sizes obtained, we recommend using methods based on epidemiological data and the nature of the results, as the results are tailored to the specific clinical problem. In addition, sample size can be reduced by lowering the number of parameter predictors, by including direct measures of the outcome of interest.

Keywords: Sample size calculation, Clinical predictive models, PRIMAGE, Paediatric oncology, Radiology

Key points

- Estimating the appropriate sample size in clinical prediction model development is mandatory to guarantee the robustness of the results.
- The selected method is designed to be applied to epidemiological data and based on the nature of outcomes.
- Strategies based on the selection and reduction of predictor variables are proposed to reduce sample size.
- The expected recruitment in PRIMAGE project fits the estimated sample size.

* Correspondence: marti_lui@gva.es
[1]Biomedical Imaging Research Group (GIBI230-PREBI) at La Fe Health Research Institute and the Imaging La Fe node of the Distributed Network for Biomedical Imaging (ReDIB) Unique Scientific and Technical Infrastructures (ICTS), Valencia, Spain
Full list of author information is available at the end of the article

Background

In research studies, including experimental clinical trials and observational studies, estimating the sample size is essential to ensure that the results will be conclusive and representative of the studied cohort [1]. Inappropriate size estimates generate uncertainties to provide reliable and reproducible answers to the questions the study intends to address [2]. Lower number of cases limits the capacity to detect existing differences, whereas larger sample size provides reliable results at the cost of increasing resources, expenses and the duration of the study [3].

Classic univariate research questions involve both descriptive (estimates of a population parameter or change) and analytical (association and correlation studies) statistics. Both methods apply a collection of well-described equations that enable the direct estimation of the needed sample size [4]. For these assessments, a prior estimate of the parameter to be studied and of its confidence interval, or the effect size and both the acceptable type I and type II errors, are needed to perform the calculations [3]. The sample size estimation in clinical predictive models extracted from observational data is more complex, since suitable direct equations are not readily available [5, 6].

When developing predictive models, a widely used “rule of thumb” to estimate sample size is that based on simulation studies conducted in the 1990s, stating that at least 10 events per predictor variable (EPP) must be included [7–9]. It should be noted that in these observational predictions, events refer to the number of patients in the sample with the clinical characteristic of interest (Fig. 1). Nevertheless, this rule has been widely questioned due to the context-specific nature of the EPP required, which may correspond to a number other than 10 EPP [10, 11].

Two methodologies that go beyond simple rule of thumb are considered as a baseline for sample size

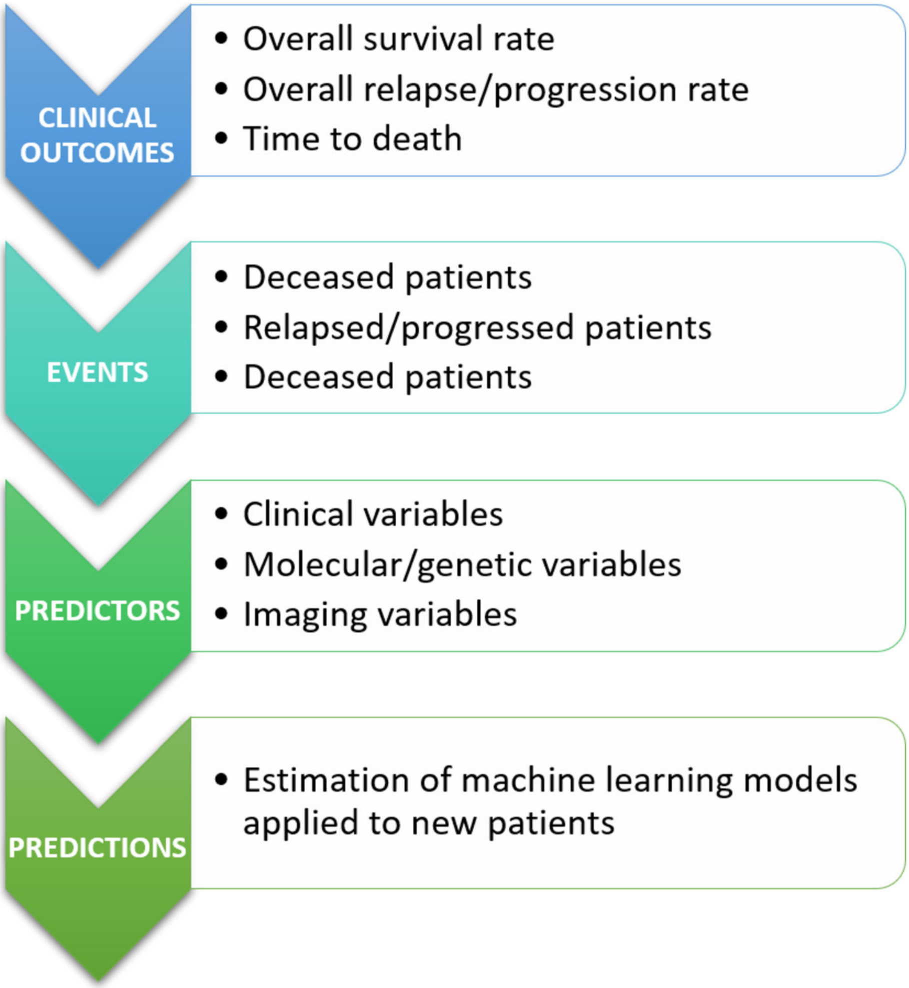

Fig. 1 Factors involved in sample size estimation and model development. Example of the clinical outcomes, events, predictors and predictions applied to neuroblastoma

estimation in this study. For logistic and Cox regression models, the 10 EPP rule of thumb can be relaxed to as low as to 5 EPP depending on the nature of the model, either logistic or Cox regression, and whether the primary predictor variable is binary or continuous [12]. A new method to calculate the sample size for parametric predictive models was proposed based on different factors, such as disease prevalence in the population, the number of predictor variables, the number of participants, and the expected fit of the regression model [13]. In this four-step method to calculate the sample size for estimation models, each step claims to meet a different criterion related to model performance. These four calculations vary depending on the type of outcome of the model (binary, continuous or time-to-event) and eventually, the largest sample size obtained is chosen. Finally, model validation has an essential role to demonstrate that an estimative algorithm is reproducible and can be consistently applied in clinical practice. In this case, there is a higher consensus that the minimum sample size for a robust validation should have at least 100 events [14].

This study aimed to use the PRIMAGE project as a use-case to apply and compare the aforementioned methods to estimate the sample size required for both model development and validation. The estimations will be performed to different scenarios regarding the clinical endpoints for neuroblastoma (NB) and diffuse intrinsic pontine glioma (DIPG) patients [15]. Secondly, the required sample size was compared with the expected recruitment within the project, and different approaches were explored to reduce the required sample size.

Methods

PRIMAGE project

PRIMAGE (PRedictive In silico Multiscale Analytics to support cancer personalised diaGnosis and prognosis, Empowered by imaging biomarkers) is a Horizon 2020 funded research project (RIA, topic SC1-DTH-07-2018), an in silico observational study for the training and validation of machine learning algorithms and multiscale prediction models [15]. This project aims to offer precise clinical assistance in the most relevant paediatric cancers: NB and DIPG. The data repository contains a high number of variables, including clinical, molecular and genetic data (above 300 different variables), as well as imaging data (more than 100 radiomic features). Throughout the project, machine learning and image processing deep learning algorithms will be used to extract pattern information from the images and link outcome results to known ground-truth diagnosis.

Sample size estimation

The methodology described by Riley [13, 16, 17] was that chosen to calculate the sample size needed to develop the computational, in silico, observational predictive model to be used in the PRIMAGE project. PRIMAGE aims to generate and validate predictive tools to diagnose and manage malignant childhood NB and DIPG tumours based on their phenotype and aggressiveness.

The sample size and EPP calculations for the different models generated, either binary or time-to-event, were implemented in R using the *pmsampsize* package [13]. For comparison, the sample size was also estimated using the 10 EPP "rule of thumb" and the updated 5 EPP rule. The calculations were applied to different scenarios for both NB and DIPG based on the clinical endpoints of interest described in the project, such as mortality risk at certain timepoint, time to death, time to relapse/progression, relapse/progression risk, event-free survival rate, and progression-free survival (PFS). In the case of NB, some of the clinical endpoints exclusively referred to the high-risk (HR) sub-group, due to their characteristics and clinical interest. A list of all these scenarios can be found in Table 1.

Epidemiological data

The clinical endpoint data required by the *pmsampsize* package [13] to perform the calculations was obtained from previous studies after a detailed review [18–21] (Table 2). In the case of the endpoints for NB, the data for the 5-year OS rate (30.7%) and time-to-death (median time-to-event 24.2 months, time of follow-up 60 months) were obtained from [18], and the data for the prevalence of relapse/progression (25.75%) was from [20]. For HR NB, the data regarding the 5-year OS rate (50%), the 5-year event-free survival rate (40.8%), the prevalence of relapse (56.1%), the median time-to-relapse (19.08 months) and the median follow-up time (72.12 months) were all collected from [19]. For the DIPG endpoints, an extensive systematic review [21] provided all the necessary data for the calculations of the different endpoints, including: the 1-year (45%) and 2-year OS rates (16.9%), the 1-year PFS rate (23.5%), the median time-to-death (11.4 months), the follow-up time for the time-to-death model (24 months), and the median time-to-progression (7.7 months) and follow-up time for the time-to-progression model (12 months).

pmsampsize settings

Among the different parameters of *pmsampsize*, shrinkage (that is the regularisation of the variability in the model's predictions to reduce overfitting) was set to the default value of 0.9 and the number of predictor variables was initially fixed to 30. This initial value was

Table 1 Clinical endpoints and model type

sTumor type	Description	Type of outcome
Neuroblastoma	5-year mortality risk	Binary
Neuroblastoma	Relapse/progression risk	Binary
Neuroblastoma	Time-to-death	Time-to-event
HR-Neuroblastoma	HR 5-year mortality risk	Binary
HR-Neuroblastoma	HR 5-year relapse/progression risk	Binary
HR-Neuroblastoma	HR-time to relapse/progression	Time-to-event
DIPG	1-year mortality risk	Binary
DIPG	2-year mortality risk	Binary
DIPG	1-year progression risk	Binary
DIPG	Time-to-death	Time-to-event
DIPG	Time-to-progression	Time-to-event

List of the clinical endpoints for neuroblastoma, HR neuroblastoma and DIPG tumors for which the sample size was determined, and the type of outcome for each of these. *DIPG* diffuse intrinsic pontine glioma, *HR* high-risk

chosen as a conservative one, since in clinical predictive models including radiomic features, the number of predictor variables is usually lower, between 2 and 20 [22–26]. Moreover, the risk of overfitting or spurious discoveries would increase with the number of predictor parameter included in the models [27, 28]. Regarding the expected fit of the model, the Cox-Snell *pseudo* R^2 (R^2_{CS}) required by the equations ranged from 0 to a max(R^2_{CS}) < 1, depending on the prevalence of the outcome. To normalize the value of R^2_{CS} in order to compare between different models, Nagelkerke defined another *pseudo* R^2 ($R^2_{\mathrm{Nagelkerke}}$) [29], calculated as the ratio between R^2_{CS} and max(R^2_{CS}) (Eq. 1), such that the R^2_{CS} needed by the equations can be obtained from the $R^2_{\mathrm{Nagelkerke}}$. With respect to the $R^2_{\mathrm{Nagelkerke}}$ value, the authors suggest that in the absence of other information sample sizes should be derived assuming the R^2_{CS} value corresponds to an $R^2_{\mathrm{Nagelkerke}}$ of 0.15. However, if the predictor variable includes direct measurements or direct measures of the processes involved in the outcome, they suggest a more appropriate $R^2_{\mathrm{Nagelkerke}}$ value of 0.5 [13]. Given that we do have some information on the processes involved but we do not have direct measures, we decided to compromise and chose a $R^2_{\mathrm{Nagelkerke}}$ value of 0.3.

For the time-to-event models, the time point of interest for the prediction and the expected average follow-up time for individuals in the dataset used to develop the model was set by experienced paediatric oncologists: 24 months for the time-to-death model of NB, for the time-to-relapse/progression model of HR NB and for the DIPG time-to-death models; and 12 months for the DIPG time-to-progression model (Eq. 1: estimation of the rate of incidence (person-time)).

Table 2 Required data for sample size calculations

Model	Prevalence	Median t_{event}	$t_{follow\text{-}up}$	Rate	Time point	Follow-up
NB 5-year mortality risk	0.307	–	–	–	–	–
NB relapse/progression risk	0.258	–	–	–	–	–
NB time to death	0.307	24.2	60	0.0063	24	24
HR-NB 5-year mortality risk	0.500	–	–	–	–	–
HR-NB 5-year relapse/progression risk	0.408	-	–	–	–	–
HR-NB time to relapse/progression	0.561	19.08	72.12	0.0132	24	24
DIPG 1-year mortality risk	0.450	–	–	–	–	–
DIPG 2-year mortality risk	0.169	–	–	–	–	–
DIPG 1-year progression risk	0.235	–	–	–	–	–
DIPG time to death	0.169	11.4	24	0.0077	24	24
DIPG time to progression	0.235	7.7	12	0.0214	12	12

Data related to each clinical endpoint for which the sample size has been determined. The number of parameters and the value of *R* is the same for all 11 scenarios (30 and 0.3, respectively) and thus, only the prevalence is required for the binary models

$$R^2_{Nage/kerke} = \frac{R^2_{Cox-Snell}}{max\left(R^2_{Cox-Snell}\right)} \tag{1}$$

One of the parameters required by *pmsampsize* functions to calculate the sample size for time-to-event models is the rate of incidence or person-time rate. Briefly, the rate of incidence is the number of new events during the study follow-up, considered as those patients that present the outcome under study, relative to the total time contributed by all subjects during the observation period (Eq. 2).

$$person{-}time\ rate = \frac{new\ events\ during\ the\ study\ follow{-}up}{total\ follow{-}up\ time} \tag{2}$$

Since this data is difficult to find from previous studies, the person-time rate was estimated following the approach shown in Eqs. 3, 4, and 5 as a function of prevalence, median time-to-event and median follow-up-time. Considering that the median time-to-event is the time at which 50% of subjects become events, the sum of the total time contributed was considered as the number of events multiplied by the median time-to-event (t_{event}) plus the number of non-events multiplied by the median time of follow-up ($t_{\text{follow-up}}$: Eqs. 2 and 3).

$$person{-}time\ rate = \frac{events}{events{*}median\ t_{event} + events{*}median\ t_{follow-up}} \tag{3}$$

Due to the characteristics of the study, only the number of events and the median follow-up time could be found. As a consequence, we used Eq. 4 as an approximation to the incidence rate (person-time), where p is the proportion between the number of events (number of patients in the sample with the clinical characteristics of interest) and the total number of patients in the study (N), as stated in Eq. 5.

$$person{-}time\ rate = \frac{p}{p{*}\left(median\ t_{event}\right) + (1{-}p){*}\left(median\ t_{follow-up}\right)} \tag{4}$$

$$p = \frac{events}{N} \tag{5}$$

In the cases where it is not possible to identify the median follow-up time but the prevalence for a certain time is available, the time point for which the prevalence data is given may be considered as the follow-up time for non-events, as if all non-events had been censored at that point.

Sample size variability

The effect of the number of predictor variables on the sample size was studied by executing the *pmsampsize* functions, varying the number of variables from 5 to 30 at intervals of 5, and leaving constant all other conditions of the equations. The variability in sample size as a function of the $R^2_{\text{Nagelkerke}}$ was assessed by establishing a value of 0.15 and 0.5 as indicated previously, and to 0.8 in accordance with a hypothetical situation in which the expected fit of the model would be higher. For the time-to-event models, the effect of the ratio between the time point of prediction and the expected time of follow-up was analysed by varying this ratio, such that the higher the ratio the longer the follow-up time relative to the time point, with ratio values between 1 and 4.

Sample size for model validation

The sample size for model validation was calculated from equation 4, using 100 as a minimum and 200 as a desirable number of events [14].

Results

Sample size determination

The sample size for each different endpoint was determined by applying the *pmsampsize* algorithms to data described in Methods. Accordingly, the sample size needed to develop robust clinical predictive models ranged from 1111 to 1397 NB patients, from 1043 to 1060 HR NB patients, and from 1043 to 1345 DIPG patients (Table 3). When more than one endpoint prediction was under study, and therefore more than one sample size was required, the largest estimated sample size should be chosen, such that the definite sample size was selected as the upper limit of the different ranges: 1397 for NB, 1060 for HR NB, and 1345 for DIPG tumours.

In order to compare the sample size obtained with other accepted methodologies, sample sizes were also calculated following the 10 EPP "rule of thumb" [7–9] and the 5 EPP estimation [12]. In the first case the sample sizes obtained were smaller, ranging from 978–1166 for the neuroblastoma, 536–736 for the HR neuroblastoma, and 668–1776 for the DIPG models. With the 5 EPP estimations the sample sizes were half those calculated with the 10 EPP rule, 490–584 for neuroblastoma, 268–369 for HR neuroblastoma, and 334–889 for DIPG. Finally, the EPP for the sample size estimated using Riley's methodology was also obtained from *pmsampsize*, and the number of events per variable were > 10 in more than half of the scenarios analysed and ≥ 7 in the rest of cases.

Sample size variability

To address the possibility of reducing the sample size while maintaining statistical power, additional calculations were performed in which conditions of the *pmsampsize* equations were varied.

The variation in the number of predictor variables showed a direct proportional behaviour between sample size and number of variables in the range 10–30

Table 3 Results of sample size calculations

Model	Riley's sample size	10 EPP	5 EPP	Riley's EPP
NB 5-year mortality risk	1111	978	490	11.37
NB relapse/progression risk	1168	1166	584	10.03
NB time to death	1397	978	490	7.00
HR-NB 5-year mortality risk	1043	600	300	17.38
HR-NB 5-year relapse/progression risk	1060	736	369	14.42
HR-NB time to relapse/progression	1060	536	268	11.23
DIPG 1-year mortality risk	1043	668	334	15.65
DIPG 2-year mortality risk	1345	1776	889	7.58
DIPG 1-year progression risk	1208	1278	639	9.46
DIPG time to death	1273	1776	889	7.87
DIPG time to progression	1130	1278	639	9.67

Sample size estimated by Riley's methodology, the 10 EPP "rule of thumb" and 5 EPP. The number of events per predictor (EPP) variable derived from the sample size obtained with Riley's methodology was also calculated. *EPP* event per predictor parameter

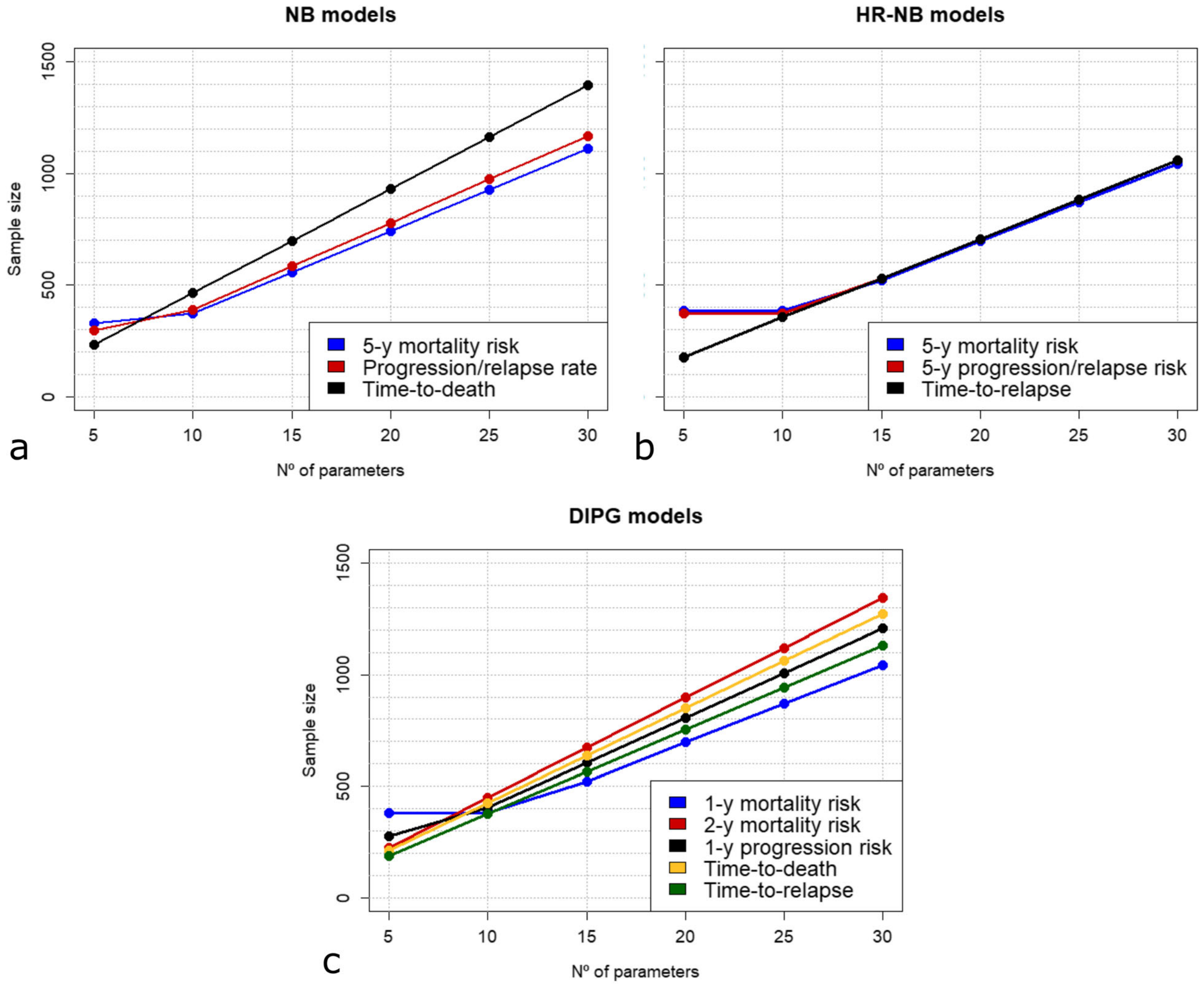

Fig. 2 Impact of the number of predictor variables on the sample size. Variability in the sample size relative to the number of predictor variables included in model development for neuroblastoma (**A**), HR neuroblastoma (**B**) and DIPG (**C**). *DIPG* diffuse intrinsic pontine glioma, *HR* high-risk, *NB* neuroblastoma

predictor variables for all the 11 scenarios analysed, as well as for 6 scenarios in the 5–30 range (Fig. 2). For example, the sample size can be reduced by half in all 11 scenarios analysed if the number of variables is reduced to a half, from 30 to 15 predictors: 556, 584, and 699 patients in the 5-year mortality risk, relapse risk and time-to-death models for NB; 522, 530, and 530 in the 5-year mortality risk, 5-year progression/relapse risk and time-to-relapse models for HR NB; and 522, 673, 604, 637, and 565 patients for the 1-year mortality risk, 2-year mortality risk, 1-year progression risk, time-to-death and time-to-relapse in the models for DIPG.

Regarding the variability of sample size as a function of the $R^2_{\text{Nagelkerke}}$, results show that including direct or indirect measures of the processes involved in the outcome to be predicted ($R^2_{\text{Nagelkerke}} = 0.5$) as opposed to not doing so ($R^2_{\text{Nagelkerke}} = 0.15$) strongly reduced the required sample size by an average of 71.2% (Table 4). This reduction in sample size was slightly lower when $R^2_{\text{Nagelkerke}}$ values of 0.5 and 0.8 were compared. Regarding the number of EPP variables, very high values (maximum 37.45) were found when the $R^2_{\text{Nagelkerke}}$ was 0.15, far above that of the classic 10 EPP. By contrast, when the $\mathrm{R}^2_{\text{Nagelkerke}}$ was set to 0.8 the EPP value dropped to as low as 3.75.

In addition, for time to event models, increasing the ratio between the time of follow-up and the time point of interest also leads to a lower sample size, with a reduction of between 13.8 and 23% when comparing a ratio of 1 and 2 (Fig. 3), although this reduction diminished as the ratio increased.

Sample size requirements for model validation

Finally, the minimum sample size required to validate the predictive models, considered as 100 events, was 326 patients for the NB models, 246 for the HR NB models, and 592 for the DIPG models, with a desirable size (200 events) of 652, 491 and 1184 patients, respectively (Table 5).

Discussion

We have explored a practical solution to estimate the sample size necessary to develop robust clinical predictive models [13] to the specific case of the observational PRIMAGE project [5]. Unlike other estimation methods, such as the 10 EPP rule [7–9], this solution provides a set of algorithms to calculate the sample size required to construct and validate robust parametric predictive models based on model quality criteria, and type of clinical outcome.

The sample size obtained with the proposed methodology was compared to its analogous estimation predicted with other more basic rules, having significant discrepancies. In addition, when comparing the number of EPPs obtained when using this method with respect to the 10 and 5 EPP rules, the number of EPPs rise above 17 in some scenarios (17.38 EPPs for the HR-NB 5-year mortality risk model) but fall to as low as to 7 in others (NB 5-year mortality risk). This confirms that 10 and 5 EPP rules may not be generally applicable since the number of EPPs for sample size estimations might depend on the context of the study, the prevalence of the outcome, the quality of the predictor variables chosen and the type of model to be developed [10, 11].

To explore possible solutions for cases where the required sample size exceeds the available sample, the size variation relative to certain parameters used in the calculations was analysed (e.g. the number of predictor variables, estimated $R^2_{\text{Nagelkerke}}$ and follow-up/time point ratio). The most feasible option to decrease the required

Table 4 Variability of sample size with $R^2_{\text{Nagelkerke}}$

Model	$R^2_{\text{Nagelkerke}} = 0.15$		$R^2_{\text{Nagelkerke}} = 0.5$		$R^2_{\text{Nagelkerke}} = 0.8$	
	Sample size	**EPP**	**Sample size**	**EPP**	**Sample size**	**EPP**
NB 5-year mortality risk	2383	24.39	666	6.82	550	5.63
NB relapse/progression risk	2495	21.42	704	6.04	590	5.06
NB time to death	2951	14.79	858	4.30	749	3.75
HR-NB 5-year mortality risk	2247	37.45	620	10.33	501	8.35
HR-NB 5-year relapse/progression risk	2280	31.01	631	8.58	513	6.98
HR-NB time to relapse/progression	2280	24.15	631	6.68	513	5.43
DIPG 1-year mortality risk	2247	33.70	620	9.30	501	7.52
DIPG 2-year mortality risk	2848	16.04	824	4.64	713	4.02
DIPG 1-year progression risk	2575	20.17	731	5.73	618	4.84
DIPG time to death	2705	16.72	775	4.79	663	4.10
DIPG time to progression	2419	20.69	679	5.81	563	4.82

Sample size calculations with different values of $R^2_{\text{Nagelkerke}}$ (0.15, 0.5, and 0.8). The number of events per predictor variable was obtained with the *pmsampsize* package

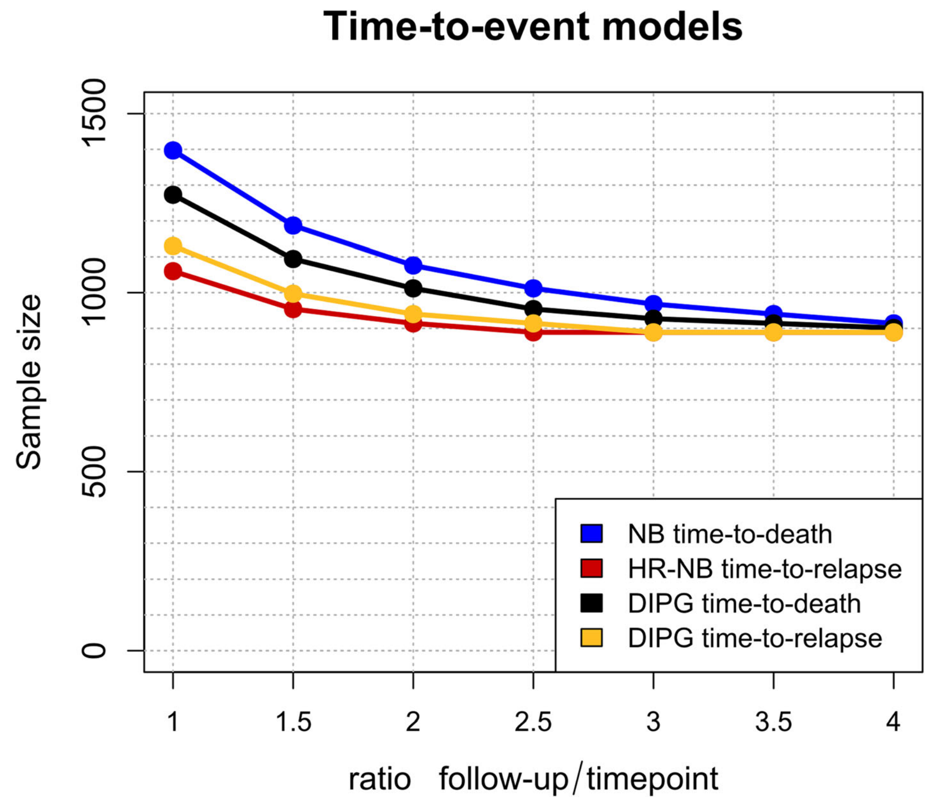

Fig. 3 How the follow-up/timepoint ratio affects the sample size. Analysis of the variation in sample size for the time-to-event models relative to the ratio between the expected average follow-up of the dataset and the time points of interest for the predictions. *DIPG* diffuse intrinsic pontine glioma, *HR* high-risk, *NB* neuroblastoma

sample size is to reduce the number of predictor variables included in the model as the number of predictors and the sample size have a directly proportional relationship in the range of 5 to 30 predictor variables. One proposed strategy is to decrease the potential predictive variables to be included in the models. For this purpose, we propose to carry out an exhaustive manual selection process of the variables to be collected in the design phase of the study. To this end, it is of great important to have the opinion of experts in the field of interest, as well as to carry out a rigorous analysis of the related literature, thus selecting the candidate predictor variables considered most important in function of the outcome to be predicted [30, 31].

Other possible approaches are to include measures of the outcome to predict which would result in an increase of the $R^2_{\mathrm{Nagelkerke}}$ value from 0.3 to 0.5, reducing the sample size by an average of 39.8 ± 0.7% (mean ± standard deviation).

Other mitigation strategies may include subject-wise cross-validation [32], resampling techniques [33], or data augmentation methods for medical images [34], and even exploring data imputation solutions for clinical data as suggested by Pezoulas et al. [35]. In some cases, the required sample size is not achievable even after applying sample reduction strategies. This is a study limitation, and researchers should be careful with the degree of evidence of the results.

Table 5 Sample sizes for external validation

Model	100 events	200 events
NB 5-year mortality risk	326	652
NB relapse/progression risk	288	776
NB time to death	326	652
HR-NB 5-year mortality risk	200	400
HR-NB 5-year relapse/progression risk	246	491
HR-NB time to relapse/progression	179	357
DIPG 1-year mortality risk	223	445
DIPG 2-year mortality risk	592	1184
DIPG 1-year progression risk	426	852
DIPG time to death	592	1184
DIPG time to progression	426	852

The sample size for the external validation of the models has been calculated considering a minimum effective sample size of 100 events and a desirable situation of 200 events

Considering that the expected sample size in the PRIMAGE project is more than 2900 NB cases, of which at least 1500 are HR NB, the expected cohort is therefore appropriate to develop reliable models with up to 30 predictive variables. However, the number of DIPG patients expected in PRIMAGE project (n = 700) falls below the sample size estimated with the default parameters for the *pmsampsize* equations (1345 cases). A downsizing strategy should be considered by applying feature reduction/selection methods and reducing the number of predictive parameters. In this way, 673 DIPG cases would be required when the number of predictive variables included in the prediction models is set to 15.

It should also be highlighted that the most important step in clinical prediction models is the validation phase, in which the true fit of the model and its applicability to daily clinical practice is assessed to ensure reproducibility. Using the lower limit of 100 events, the minimum sample size obtained for the external validation of the PRIMAGE models was 326 patients for NB and 592 for DIPG, which is an achievable number in the case of NB but somewhat more challenging for DIPG given its lower incidence.

Regarding the possible biases, the *pmsampsize* formulas were developed considering only linear regression models for continuous outcomes, logistic regression for binary outcomes, or proportional hazards regression models for time-to-event data. These three different algorithms are parametric, suitable to obtain predictive models when the relationships between the different variables in the dataset are known and well-defined. However, when the relationships between variables are not direct, it seems more appropriate to apply non-parametric models that can efficiently exploit the more complex relationships between the variables, such as the k-nearest neighbours, support vector machines or decision tree algorithms. Therefore, the quality of the variables, the selection of the most appropriate algorithm for the data model, and the process of hyperparameter tuning are essential to obtain robust predictive models.

In summary, we have applied a recently devised method to determine sample sizes for clinical predictive model development and validation to the use-case of the observational PRIMAGE project, providing an overview of different sample size reduction approaches. This methodology is based on the epidemiological data and the nature of the outcome, tailoring the obtained sample size to the specific medical problem of interest. A common research framework for sample size estimation methodologies for the development and validation of clinical predictive models should be defined by the clinical research community.

Abbreviations

DIPG: Diffuse intrinsic pontine glioma; EPP: Events per predictor; HR: High-risk; NB: Neuroblastoma; OS: Overall survival; PFS: Progression-free survival

Acknowledgements

We thank Ana García Pérez and the team from Viceversa Traduccions for the language corrections and helpful comments to improve the manuscript.

Authors' contributions

CBD, LCA, LMB, and JMCS contributed to the conception and design of the work. BMH and BR searched for and provided epidemiological data. CBD performed sample size estimations and the analysis of sample size variability and was the major contributor in writing the manuscript. CBD, LCA, LMB, and DVC participate in the interpretation and discussion of results. LCA contributed to writing the manuscript. LCA and LMB substantially revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is funded by the HORIZON2020 PRIMAGE project (RIA, topic SC1-DTH 07-2018), from the EU Framework Programme for Research and Innovation of the European Commission.

Availability of data and materials

Not applicable

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Biomedical Imaging Research Group (GIBI230-PREBI) at La Fe Health Research Institute and the Imaging La Fe node of the Distributed Network for Biomedical Imaging (ReDIB) Unique Scientific and Technical Infrastructures (ICTS), Valencia, Spain. [2]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain. [3]Radiology Department, Hospital Universitario y Politécnico La Fe, Valencia, Spain. [4]Pediatric Oncology Department, Hospital Universitario y Politécnico La Fe, Valencia, Spain.

Received: 28 June 2021 Accepted: 12 April 2022
Published online: 01 June 2022

References

1. Eng J (2003) Sample size estimation: how many individuals should be studied? Radiology 227:309–313. https://doi.org/10.1148/radiol.2272012051
2. Nayak BK (2010) Understanding the relevance of sample size calculation. Indian J Ophthalmol 58:469–470. https://doi.org/10.4103/0301-4738.71673
3. Das S, Mitra K, Mandal M (2016) Sample size calculation: basic principles. Indian J Anaesth 60:652–656. https://doi.org/10.4103/0019-5049.190621
4. Cohen J (1977) Statistical power analysis for the behavioral sciences. Academic Press, New York
5. Eng J (2004) Sample size estimation: a glimpse beyond simple formulas. Radiology 230:606–612. https://doi.org/10.1148/RADIOL.2303030297
6. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, Garcia-Pedrero A, Ramirez SC, Kong D, Moody AR, Tyrrell PN (2019) Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J 70:344–353. https://doi.org/10.1016/j.carj.2019.06.002
7. Concato J, Peduzzi P, Holford TR, Feinstein AR (1995) Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. J Clin Epidemiol 48:1495–1501. https://doi.org/10.1016/0895-4356(95)00510-2

8. Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis II. J Clin Epidemiol 48:1503–1510. https://doi.org/10.1016/0895-4356(95)00048-8
9. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49:1373–1379. https://doi.org/10.1016/j.amepre.2003.12.002
10. Ogundimu EO, Altman DG, Collins GS (2016) Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 76:175–182. https://doi.org/10.1016/j.jclinepi.2016.02.031
11. Austin PC, Steyerberg EW (2017) Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 26:796–808. https://doi.org/10.1177/0962280214558972
12. Vittinghoff E, McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and cox regression. Am J Epidemiol 165:710–718. https://doi.org/10.1093/aje/kwk052
13. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M (2020) Calculating the sample size required for developing a clinical prediction model. BMJ 368:1–12. https://doi.org/10.1136/bmj.m441
14. Collins GS, Ogundimu EO, Altman DG (2015) Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 35:214–226. https://doi.org/10.1002/sim.6787
15. Martí-Bonmatí L, Alberich-Bayarri Á, Ladenstein R, Blanquer I, Segrelles JD, Cerdá-Alberich L, Gkontra P, Hero B, García-Aznar JM, Keim D, Jentner W, Seymour K, Jiménez-Pastor A, González-Valverde I, Martínez de las Heras B, Essiaf S, Walker D, Rochette M, Bubak M, Mestres J, Viceconti M, Martí-Besa G, Cañete A, Richmond P, Wertheim KY, Gubala T, Kasztelnik M, Meizner J, Nowakowski P, Gilpérez S, Suárez A, Aznar M, Restante G, Neri E (2020) PRIMAGE project: predictive in silico multiscale analytics to support childhood cancer personalised evaluation empowered by imaging biomarkers. Eur Radiol Exp 4:22. https://doi.org/10.1186/s41747-020-00150-9
16. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, Collins GS (2019) Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. Stat Med 38:1262–1275. https://doi.org/10.1002/sim.7993
17. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, Collins GS (2019) Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 38:1276–1296. https://doi.org/10.1002/sim.7992
18. Al-Tonbary Y, Badr M, Mansour A et al (2015) Clinico-epidemiology of neuroblastoma in north east Egypt: a 5-year multicenter study. Oncol Lett 10:1054–1062. https://doi.org/10.3892/ol.2015.3335
19. Simon T, Berthold F, Borkhardt A, Kremens B, de Carolis B, Hero B (2011) Treatment and outcomes of patients with relapsed, high-risk neuroblastoma: results of German Trials. Pediatr Blood Cancer 56:578–583. https://doi.org/10.1002/pbc.22693
20. London WB, Castel V, Monclair T, Ambros PF, Pearson ADJ, Cohn SL, Berthold F, Nakagawara A, Ladenstein RL, Iehara T, Matthay KK (2011) Clinical and biologic features predictive of survival after relapse of neuroblastoma: a report from the International Neuroblastoma Risk Group Project. J Clin Oncol 29:3286–3292. https://doi.org/10.1200/JCO.2010.34.3392
21. Gallitto M, Lazarev S, Wasserman I, Stafford JM, Wolden SL, Terezakis SA, Bindra RS, Bakst RL (2019) Role of radiation therapy in the management of diffuse intrinsic pontine glioma: a systematic review. Adv Radiat Oncol 4:520–531. https://doi.org/10.1016/j.adro.2019.03.009
22. Delzell DAP, Magnuson S, Peter T, Smith M, Smith BJ (2019) Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. Front Oncol 9:1–8. https://doi.org/10.3389/fonc.2019.01393
23. Corso F, Tini G, Lo Presti G, Garau N, de Angelis SP, Bellerba F, Rinaldi L, Botta F, Rizzo S, Origgi D, Paganelli C, Cremonesi M, Rampinelli C, Bellomi M, Mazzarella L, Pelicci PG, Gandini S, Raimondi S (2021) The challenge of choosing the best classification method in radiomic analyses: recommendations and applications to lung cancer CT images. Cancers (Basel) 13. https://doi.org/10.3390/cancers13123088
24. Shiri I, Sorouri M, Geramifar P, Nazari M, Abdollahi M, Salimi Y, Khosravi B, Askari D, Aghaghazvini L, Hajianfar G, Kasaeian A, Abdollahi H, Arabi H, Rahmim A, Radmard AR, Zaidi H (2021) Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients. Comput Biol Med 132:104304. https://doi.org/10.1016/j.compbiomed.2021.104304
25. Chang R, Qi S, Yue Y, Zhang X, Song J, Qian W (2021) Predictive radiomic models for the chemotherapy response in non-small-cell lung cancer based on computerized-tomography images. Front Oncol 11:1–13. https://doi.org/10.3389/fonc.2021.646190
26. Shin J, Lim JS, Huh YM, Kim JH, Hyung WJ, Chung JJ, Han K, Kim S (2021) A radiomics-based model for predicting prognosis of locally advanced gastric cancer in the preoperative setting. Sci Rep 11:1–12. https://doi.org/10.1038/s41598-021-81408-z
27. Liu R, Gillies DF (2016) Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognit 53:73–86. https://doi.org/10.1016/j.patcog.2015.11.015
28. Fan J, Zhou WX (2016) Guarding against spurious discoveries in high dimensions. J Mach Learn Res 17:1–34. https://doi.org/10.5555/2946645.3053485
29. Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. Biometrika 78:691–692. https://doi.org/10.1093/biomet/78.3.691
30. Box GEP, Hunter JS, Hunter WG (2005) Statistics for experimenters: design, innovation, and discovery, 2nd edn. Wiley-Interscience
31. Chatfield C (1995) Problem solving. A statistician's guide, 2nd ed. Chapman & Hall
32. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP (2017) The need to approximate the use-case in clinical machine learning. Gigascience 6:1–9. https://doi.org/10.1093/GIGASCIENCE/GIX019
33. White D, Lawson RS (2015) A Poisson resampling method for simulating reduced counts in nuclear medicine images. Phys Med Biol 60:N167–N176. https://doi.org/10.1088/0031-9155/60/9/N167
34. Shin H-C, Tenenholtz NA, Rogers JK, et al (2018) Medical image synthesis for data augmentation and anonymization using generative adversarial networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11037 LNCS:1–11. https://doi.org/10.1007/978-3-030-00536-8_1
35. Pezoulas VC, Grigoriadis GI, Gkois G, Tachos NS, Smole T, Bosnić Z, Pičulin M, Olivotto I, Barlocco F, Robnik-Šikonja M, Jakovljevic DG, Goules A, Tzioufas AG, Fotiadis DI (2021) A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains. Comput Biol Med 134:104520. https://doi.org/10.1016/j.compbiomed.2021.104520

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.