Limitations and opportunities for identifying outcome prognostic factors in the context of small samples

Lisa Mounier, Alexandre Civet, David Pau, Julien Dupin, Cyril Esnault

Roche, France

BACKGROUND AND OBJECTIVE

Identification of prognostic factors is essential for advances in clinical research but rare diseases or genomics mutations studies face the limits of small numbers of patients, for which current methods coming from biostatistics and machine learning may be challenged. Indeed, small populations having high variability, that constraints analyses in terms of robustness and estimates with a possible production of uncertain results.

GOAL: This work aims to address the need for a synthesis around limitations and opportunities for up-to-date statistical and machine learning methods in the context of small samples.

METHODS

A literature review based on relevant keywords in the context of small sample size was performed on Google Scholar and PubMed, such as " small data", "small sample", "small population" merged with "issues" ,"bias" ,"limits" and "problems" related or not to the identification of prognostic factors. The criteria used for selecting methodological papers included the date of publication and number of citations. Then, the first selection allowed to select others papers based on associated references or on new key-words identified in previously identified articles. These principle was repeated several times to refine the search.

RESULTS

As of today, the following limits in low sample size for prognostic identification were identified in relevant literature which corresponds to 15 articles retained out of 22 identified. Theses limits are described below in the tables :

1) Overfitting leading to false-positive results and instability.

« Overffiting » is a production of an analysis that corresponds too closely of a dataset and may fail to generalize in an other dataset. An overfitted model is characterized by a high number of parameters relative to what the data needs.

Opportunity : To reduce overfitting and increase confidence in results

Type of methods	Examples of methods
Stricter variables selection protocols, which means no mixing of data to determine and assess prognostic factors	❑ Train/test split ❑ Complete Cross-validation ❑ Nested Cross-validation (most advisable)
Methods adding external knowledge to increase confidence in results	❑ Cross-referencing of data sources ❑ Meta-analyses by combining estimates from regression (based on the literature) with coefficients of estimated regression on the studied dataset ❑ Addition of prior information through Bayesian methods ❑ Clinical knowledge from experts/clinicians
Ensemble methods based on the aggregation of different results to obtain a more stable and accurate result by a voting system.	❑ At the feature-scale with principles of Bagging and Boosting, algorithms such as Random-Forest, SVM-RFE, XGBoost… ❑ At the method scale with the principle of stacking to obtain the best model
Regularized regressions which explicitly penalize overly complex models and test the model ability to generalize	❑ Lasso ❑ Ridge ❑ Elastic Net
Noise injection methods where adding a noise has a regularization effect to the training set	❑ Jittter
Sampling methods to gain stability for identified prognostic factors	❑ Sampling with replacement such as Bootstrap
Dataset transformation methods which allows to avoid the primary amount of data	❑ Data augmentation methods ❑ Virtual samples

2) High-dimensionality of the sample

The high-dimensionality is when your number of features is too large according to the number of patients, that questions usual statistical models and machine learning algorithms.

Opportunity : To deal with high-dimensionality of a small sample

Type of methods	Examples of methods
Dataset transformation methods which allows to avoid the primary amount of data	❑ Data augmentation methods ❑ Virtual samples
Feature transformation methods combining supervised and unsupervised methods to reduce data dimensions	❑ Supervised methods : Linear Discriminant Analysis, Partial Least Square Discriminant Analysis ❑ Unsupervised methods : Factor Analysis, Clustering of variables, Non-negative matrix factorization
Methods based on sparsity to reduce number of variables in the most reliable way	❑ Lasso, Ridge regressions ❑ Adaptative Lasso ❑ Group Lasso ❑ Sparse Group Lasso ❑ Group PLS ❑ Sparse Group PLS
Rule of thumb to have a sufficient number of events in data	❑ One in ten rule : the rule states that one predictive variable can be studied for every ten events
Feature selection families to reduce dimensions: ❑ Filter : univariate or multivariate analysis based on criteria ❑ Wrapper : research of optimal subset by combining iterative search and an algorithm to assess performance ❑ Embedded : variable selection process contained in methods	❑ Filter : statistical criteria (e.g. p-value), distance and information measures (e.g. Gain Ratio) ❑ Wrapper : sequential search (e.g. stepwise methods, best first strategies method), exponential search (exhaustiv methods…) and random search (e.g. genetic algorithms) ❑ Embedded : regularized methods (e.g. Lasso, Ridge) and machine learning algorithms (Decision Tree, Random forest…)

3) Lack of accuracy with large confidence intervals

Accuracy in statistics usually refers to the extent to which the results of a test are similar after several tests. In the case of small samples, there are large variations due to the lack of data and data heterogeneity.

Opportunity : To deal with accuracy

Type of methods	Examples of methods
Methods to increase precision of results based on external knowledge	❑ Meta-analyses by combining estimates from regression (based on the literature) with coefficients of estimated regression on the studied dataset ❑ Addition of prior information through Bayesian methods
Estimation methods to refine confidence intervals	❑ Boostrap confidence interval estimates
Epidemiological rule of thumb to ensure there are enough events in data	❑ "One in ten rule"

CONCLUSION

Many methods adapted to the search for prognostic factors attempt to deal with problems related to small sample sizes. Faced with this great diversity, further research is needed to better guide the statistician.

REFERENCES

1. Ludmila I. Kuncheva, Juan J. Rodríguez,On feature selection protocols for very low-sample-size data,Pattern Recognition,Volume 81, 2018,Pages 660-673,ISSN 0031-3203,https://doi.org/10.1016/j.patcog.2018.03.012.
2. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11): e0224365. https://doi.org/10.1371/journal.pone.0224365
3. Li, Y., Li, T. & Liu, H. Recent advances in feature selection and its applications. *Knowl Inf Syst* **53**, 551–577 (2017). https://doi.org/10.1007/s10115-017-1059-8
4. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000 Apr;19(8):1059-1079. DOI: 10.1002/(sici)1097-0258(20000430)19:8<1059::aid-sim412>3.0.co;2-0. PMID: 10790680.
5. Daniel McNeish (2016) On Using Bayesian Methods to Address Small Sample Problems, Structural Equation Modeling: A Multidisciplinary Journal, 23:5, 750-773, DOI: 10.1080/10705511.2016.1186549
6. Hua, J., Lowey, J., Xiong, Z. *et al.* Noise-injected neural networks show promise for use on small-sample expression data. *BMC Bioinformatics* **7**, 274 (2006). https://doi.org/10.1186/1471-2105-7-274
7. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003 May;56(5):441-7. doi: 10.1016/s0895-4356(03)00047-7. PMID: 12812818.
8. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health. 2020 Feb 16;8(1):e000262. doi: 10.1136/fmch-2019-000262. PMID: 32148735; PMCID: PMC7032893.
9. Jia, W., Sun, M., Lian, J. *et al.* Feature dimensionality reduction: a review. *Complex Intell. Syst.* **8**, 2663–2693 (2022). https://doi.org/10.1007/s40747-021-00637-x