**Appendix A Performance results of linear regression metamodels in comparison with various other metamodeling techniques.** A generalized additive model (GAM) is an expansion of the multivariable linear regression model. In GAM the parameters that have a non-linear relationship with the outcome are represented by smoothing functions.[[1]](#endnote-2),[[2]](#endnote-3) The plots of these smoothing functions - or splines - show the independent contributions of each variable to the outcome. The effective degrees of freedom quantify the flexibility of the smoothing splines, with a value of 1 indicating a linear relationship between the variable and the outcome, and higher values indicating a non-linear relationship.

To account for the expected asymmetrical, positively skewed distribution of the cost outcome, we also compared the performance of OLS with log transformation and generalized linear models (GLM) with gamma-family and log-link for the prediction of cost.[[3]](#endnote-4),[[4]](#endnote-5) In OLS with log transformation, we assumed that the log-transformed costs show an approximately linear relationship to the model input parameters. The predicted cost values from an OLS model are on the log-USD scale and require a back-transformation to the original scale of USD. We used simple exponentiation for back-transformation to the original scale of the dependent variable. GLM is a more recent and commonly recommended method for cost prediction, which does not require back-transformation.31 In GLM with gamma-family and log-link, values of the cost outcome are drawn from a gamma distribution rather than a normal distribution and the log-transformed costs have a linear relationship with the model parameters. The predicted cost values are calculated by exponentiation of the regression equation.

**Table 1 Cross-validation Results of Various Metamodeling Techniques in Comparison with the CEPAC-P Model for the *EID* Strategy**

|  |  |  |
| --- | --- | --- |
|  | **Life expectancy** | **Lifetime per-person HIV-related cost** |
| **R2 Statistic** | **OLS** | **GAM** | **OLS** | **GAM** | **OLS:log** | **GLM: gamlog** |
| Training dataset 1(2,500 simulations) | 0.99 | 0.99 | 0.98 | 0.93 | 0.88 | 0.89 |
| Validation dataset 1(2,500 simulations) | 0.99 | 0.99 | 0.98 | 0.91 | 0.89 | 0.92 |
| Training dataset 2(2,500 simulations) | 0.99 | 0.99 | 0.98 | 0.92 | 0.88 | 0.89 |
| Validation dataset 2(2,500 simulations) | 0.99 | 0.99 | 0.98 | 0.91 | 0.89 | 0.90 |

We conducted 5,000 CEPAC-P model microsimulations. We divided these 5,000 parameter sets into a training dataset (the 2,500 simulations used in metamodel development) and a validation dataset (the 2,500 simulations not used in metamodel development). **EID**, early infant diagnosis; **GAM**, generalized additive model; **GLM:gamlog**, generalized linear model with gamma-family and log-link; **OLS**, ordinary least squares; **OLS:log**, ordinary least squares with log transformation.

1. Hastie T, Tibshirani R. Generalized additive models. Stat Sci 1986;1(3):297-318. [↑](#endnote-ref-2)
2. Wood SN. Generalized Additive Models: An introduction with R. Boca Raton, FL: Chapman and Hall/CRC; 2006. [↑](#endnote-ref-3)
3. Moran JL, Solomon PJ, Peisach AR, Martin J. New models for old questions: generalized linear models for cost prediction. J Eval Clin Pract 2007;13(3):381-9. [↑](#endnote-ref-4)
4. Malehi AS, Pourmotahari F, Angali KA. Statistical models for the analysis of skewed healthcare cost data: a simulation study. Health Econ Rev 2015;5:11. [↑](#endnote-ref-5)