Appendix 3. Sample size justification and detailed power statement for the main phase

As indicated in the main text, the global aim translates into several more specific sub-hypotheses. As a consequence, the BeLOVE cohort does not have one single hypothesis that can be taken as a single basis for a confirmatory analysis strategy and a related sample size calculation. The aim of this section is, therefore, to justify that the recruited number of patients is sufficient to generate results with good precision under various different data scenarios.

Even though many different types of analyses and models will be applied, the time-to-event analyses are in the major focus of BeLOVE. The Cox proportional hazards regression model is the basic approach for modeling time to event data. The other models discussed above for recurrent or competing events are all extensions of this well-known Cox-model. Therefore, it seems reasonable to base the power considerations on the Cox model approach.

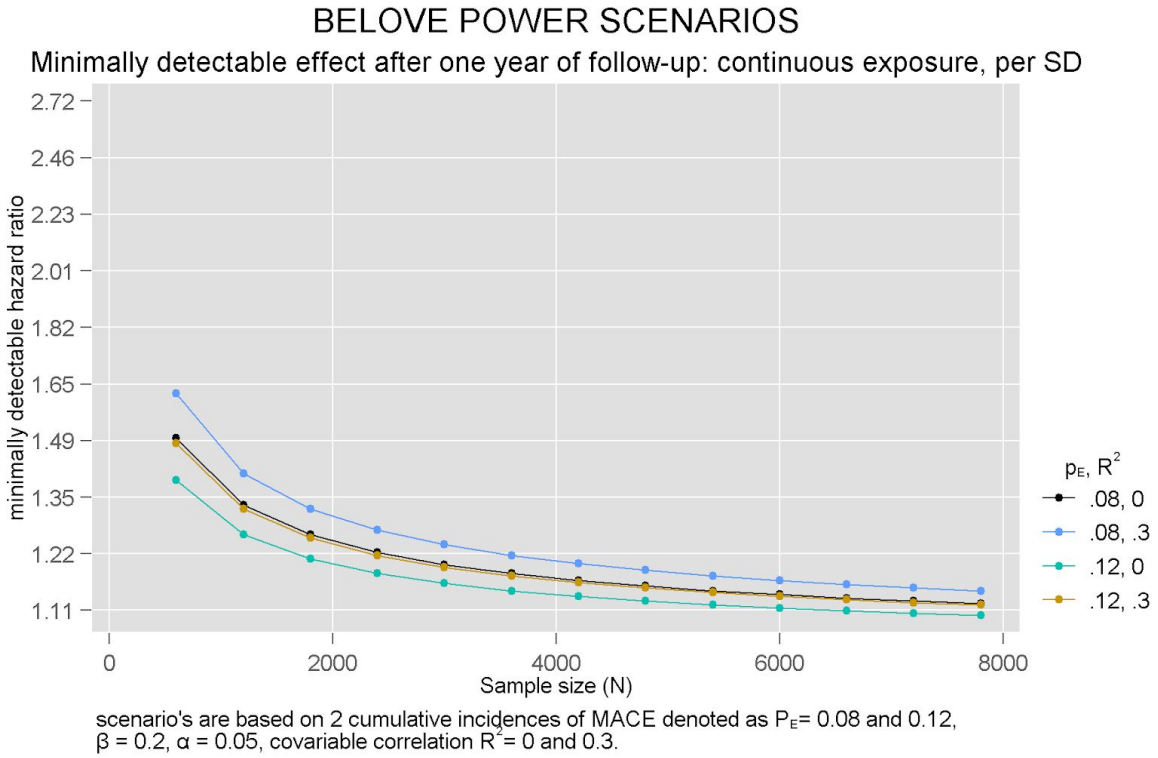
The power of a Cox model is influenced by several factors:

- The **applied two-sided significance level**, which controls the type I error which is set to 5% as proposed by all study relevant guidelines.
- The **anticipated power** value (usually set as 80% or 90%).
- The **number or percentage of observed events** at the investigated follow-up time point. For BeLOVE, it is assumed that at the 1-year follow-up 8-12% of the patients experience a cardiovascular event of any type (Amarenco et al., 2016; Jernberg et al., 2015; Maggioni, 2015; Wells et al., 2008). For the individual disease entities, the event rates might differ. Note that the BeLOVE study is planned with a much longer follow-up of up to 10 years. With an increasing follow-up. The number of expected events will increase and as a consequence, the power will increase as well. Therefore, the considerations made in here for the one-year follow-up define a conservative scenario and even better power values can be expected for longer follow-up times.
- The **type, the distribution and the number of predictors** included in the model. For the sake of simplicity, we will assume two types: continuous and binary. For continuous we will model the exposure per standard deviation increase (standardized effect) and for binary predictor x we assume equal groups (50% prevalence of the exposure of interest). The accuracy of a prediction model generally increases with more predictors, so this simple assumption defines a conservative scenario.
- The **anticipated treatment effect** between the two groups which are defined by the binary predictor. A Cox model quantifies a treatment effect in terms of the hazard ratio (HR) or the logarithm of the HR (Alternative coefficient b_1=ln(HR)), where an HR > 1 indicates a favorable outcome for the group x=0.
- The **anticipated degree of correlation among predictor of interest and all other predictors in the model**, which is given as pseudo-R^2, which lays within [0;1]. Values close to 0 indicate that the predictor of interest is independent of all other covariates. As there are always multiple factors influencing the final outcome, a correlation among predictors of 0.3 seems often more reasonable.

As there are a number of factors influencing the sample size, we will evaluate a number of parameter constellations as specified in the table below.

Significance level α	Power 1-b	Percentage of observed events/Year	predictor	Treatment effect HR ($\ln(HR)=b_1$)	Degree of correlation among predictor R^2
5%	80%	8% 12%	continuous, SD^{-1} Binary, 50%	1.11 (0.1) 1.82 (0.7)	0.0 0.3

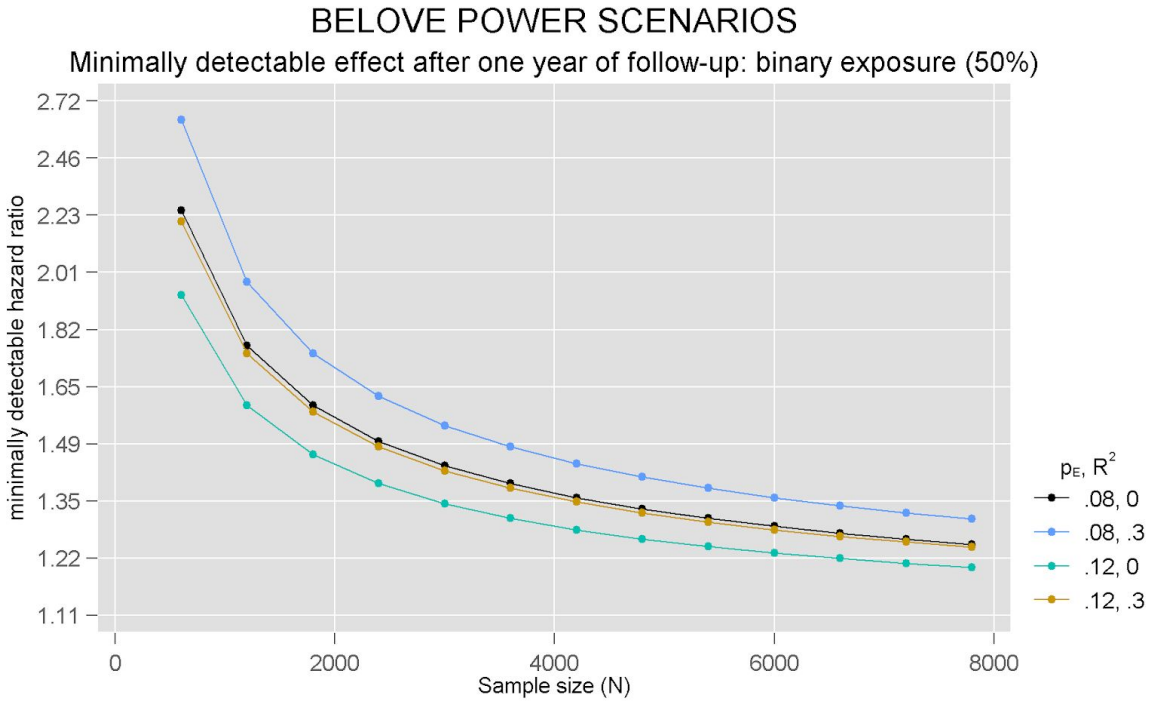
Based on these parameters, we can construct a figure that provides us with a good overview of the precision and minimally detectable effects for continuous exposures modeled per standard deviation increase that can be expected from a clinical cohort study like BeLOVE. Specifically, the figure investigates the effect of a sample size from 1200 (number of patients per disease entity, second marker reading from left to right) to 6000 (minimal total number of patients in the smallest of the two BeLOVE subcohort) and beyond.



As an exemplary interpretation, for a significance level of 5%, a power of 90%, and 0 correlation between the predictors and an annual event proportion of 8%, the minimal effect we can detect in 6000 participants will be HR 1.17 (black line), which is lowered to HR 1.11 when a 12% annual incidence of outcome is assumed (green line). This shows that with all patients in a combined analysis, BeLOVE has enough statistical power to pick up small. Our calculations also include more

conservative scenarios, e.g. if the population sample size at 1 year is reduced due to loss-to-follow-up, or analyses based on patients with one specific disease. Moreover, there will be the need to adjust for other covariates as for the predictor of interest and these set of predictors will usually be correlated. The figure shows these different scenarios, by varying the sample size as well as plotting separate lines for a single predictor (R^2=0) as well as for several correlated predictors, for example, adjusted for age, sex, and other traditional cardiovascular risk factors (R^2 =0.3). It can be seen that if the predictor of interest is correlated to other predictors (blue and yellow lines), the required sample sizes are larger than for uncorrelated predictors (black and green lines). The flattening of the lines with increasing sample sizes indicate that the added precision obtained from increasing sample size reduces dramatically with sample sizes above N=2000.

A similar picture, but with higher minimally detectable hazard ratios, is obtained when looking at the minimally detectable differences for binary exposures with a prevalence of 50%. Given that all other aspects are similar, this increase is only due to the difference in exposure modeling.



Please note that all the scenarios above and in the graphs are quite general and based on incidence rates often associated with one-year follow-up. Longer follow-up will result in a higher number of observed events and thus increased power.

For some research questions, the high dimensionality, sparsity and unknown underlying distribution of some of the data obtained in our deep phenotyping approach, the usefulness of this traditional approach might be limited. Therefore, other, potentially non-parametric, techniques from the field of machine learning (both supervised and unsupervised) might be required [1]. Given that our main is focused on improving risk prediction these other approaches could include classifying algorithms such as random forests, K-Nearest Neighbors, Support Vector Machines Prediction Analysis for Microarrays (PAM) etc [2]. to help identify relevant clinical subgroups within the BeLOVE study population. These analyses techniques are relatively novel for the field of clinical medicine and

harbor a great potential in general, including direct clinical relevance through clinical decision aids based on these algorithms[3]. However, these techniques are not tried and tested over decades such as the more traditional regression-based techniques and need to be further adapted or even developed from scratch for the individual analyses scenarios that we will encounter within BeLOVE.

The graphs above are based numbers obtained by the “power” package from STATA 14.0 with the following details:

- power cox, sd(1.0) n(600 1200:8000) r2(0 0.3) failprob(0.08 .12) effect(hratio) power(0.8) direction(upper)
- power cox, sd(0.5) n(600 1200:8000) r2(0 0.3) failprob(0.08 .12) effect(hratio) power(0.8) direction(upper)

References

1. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci. Institute of Mathematical Statistics; 2001;16: 199–231.
2. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics. 2010;11: 447.
3. Bae J-M. The clinical decision analysis using decision tree. Epidemiol Health. 2014;36: e2014025.