White paper

Strategies for design of protein biomarker studies

Introduction

Protein biomarker-based studies have great potential to drive the development of precision medicine, but in order to maximize their impact, aspects such as study design and statistical analysis need to be carefully considered.

The requirements for a proper study design include a well- defined study objective, adequate sample size, control of confounding factors and biases and appropriate statistical analysis. Moreover, application of standardized protocols for sample handling to ensure equal sample quality is of the highest importance. Advice on sample handling and processing is provided in Olink's white paper *"Pre-analytical variation in protein biomarker research"* (1).

The study design should consider all procedures involved in the study, from initial planning, through sample collection and analysis to the final report. If a study is properly designed, any factors which distort or bias the results of a test procedure can be minimized. This white paper describes important aspects of study design to consider when planning your future research.

Planning phase

The planning phase of a study is very important. To have confidence in the study results and conclusions, the study objectives for a project and other considerations necessary for a successful outcome need to be addressed before beginning the project. For example, an under-powered study could be a waste of limited resources if no firm conclusions could be drawn with confidence.

TERMINOLOGY

In a group comparison, the Power of a study is the probability that a true difference in protein levels will be detected as statistically significant in a hypothesis test. Power analysis can be used to calculate the minimum sample size required to reach a certain power. It is always recommended to have as high power as possible, but in practice, studies are often designed to have 80% power based on assumptions of parameters such as effect size.

All studies will have biases, and researchers should always strive for these to be acknowledged and avoided. The design of experiments must therefore be thoroughly considered from a statistical perspective. It is recommended to have interdisciplinary teams who work together designing the study.

Power of a study

To translate an experimental objective into a statistical or analytical plan it is important to understand how to derive accurate power calculations for the study. Different study questions require different statistical methods, and these in turn require different sample sizes to detect significant changes. The statistical method used may have a very significant effect on the power analysis. Without an understanding of the statistical tools best suited for a given experiment, it is impossible to carry out an appropriate power calculation. A power calculation for many of the most commonly used statistical tests can be solved through a straightforward equation, see Dell *et al.* (2), but for more complex analyses simulations may be necessary.

Power analysis

Ideally, power analyses should always be carried out prior to running samples and collecting data. A power analysis conducted prior to the research study is typically used to determine an appropriate sample size to achieve adequate power. Power analysis can also be conducted after a study has been completed, and such retrospective calculations may be useful when designing a scaled-up study from a smaller pilot for example.

Power calculation can also be used to determine the minimum measurable effect size, meaning the quantification of the difference between two groups, that is likely to be detectable in a study with that sample size.

Sample size calculations

The goal of performing sample size calculations during the planning phase of the study is to ensure that the outcome is informative. If the sample size is too small, the probability of producing meaningful results is low. On the other hand, using too many samples will be unnecessarily costly, time-consuming and may be considered unethical.

If an outcome is rare and therefore occurs infrequently, it is imperative to increase the sample size in order to detect potential differences. It may also be advisable to oversample the rare group since power is largely determined by the smallest group in the study. A balanced study design between groups is therefore recommended.

Increasing the sample size is like increasing the resolution of a picture. With just a few samples, the picture is so fuzzy that we would only be able to see differences between the most distinct data. However, if the sample size is large, the picture becomes sharp enough to determine even very small differences between data.

Variables

Power is directly related to spread (standard deviation), effect size, sample size and significance level. These variables are described below. In general, it is possible to calculate or estimate one of these five variables if the others are kept constant. In reality, it is common to explore multiple levels of the constant variables to see how the target variable (e.g. power) is affected.

IMPORTANT!

All power analyses performed prior to the study are built on *assumptions* of the variables and can only be used as a guide when designing a study. These assumptions can come from domain knowledge, historical data, literature searches or pilot studies.

Spread

Spread of data, commonly measured as the standard deviation, describes how similar observed data points are, and must be estimated in order to perform a power calculation. If the standard deviation is high for a variable, a larger sample size is needed to reach a certain power compared to a variable with low standard deviation.

A study may have multiple sources of variation and the spread may therefore be hard to estimate. If no good estimate is available it may still be worth performing a power analysis with both low and high variability levels to see what difference it makes to the calculated sample size or power.

Effect size

An effect size can be standardized and it contains information about the variability in the measurements.

One common way of standardizing the effect size is by dividing the observed mean difference by the estimated standard deviation, a value called Cohen's *d*. If there is no overlap between the data in the experimental group and the control group, there is a substantial difference, and the effect size is high. But if the overlap between the groups is larger than the difference between the groups, the effect size is less significant.

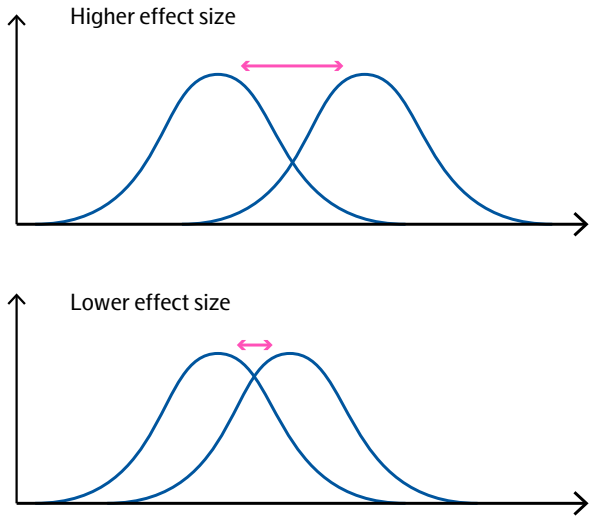

Fig 1. The higher effect size pulls the two distributions apart and differences are easier to detect.

The standardized effect size, Cohen's *d*, is the equivalent to a Z-score of a standard Normal distribution. For example, an effect size of 0.8 means that the score of the average person in the experimental group is 0.8 standard deviations above the average person in the control group. The power is higher when detecting larger effect sizes. Hence, if the effect size is small, a larger number of samples are needed to have the power to detect it.

Sample size

This is the number of samples in each group included in the study. It is often the variable we want to estimate, and it is most often the only factor the researcher can control to affect the power of the study. However, in cases where it is not possible to increase the number of samples, a power analysis can be used to determine the smallest effect size at which the proposed study will have a reasonably high power. Increasing sample size will always increase the power of a study.

Significance level

The significance level (often denoted α) is the probability of a false positive result. The significance level is commonly set to 0.05 if one statistical test is carried out. However, Olink's panels each contain 92 assays and one test is commonly applied to each assay. This means that the significance level should be adjusted accordingly for multiple testing. A common way of doing this is by dividing the significance level with the number of tests which is called a Bonferroni correction. After correction for 92 test the corrected level is around 0.000544.

Each data has an associated probability value called a p-value, which is defined as the probablility that a difference is observed by chance when no true difference exists given the sampling distribution.

If the p-value is less than the significance level after necessary correction ($p < 0.05$), the result is statistically significant. If the p-value is greater than the significance level ($p > 0.05$), the result is statistically non-significant. Increasing the significance level will increase power, but also increase the probability of detecting a difference when there is none. Therefore, increasing the number of tests (running more panels) will require running more samples to maintain the same level of power.

Of the four variables discussed (spread, effect size, sample size and significance level), effect size has by far the biggest influence on power, but all should be taken into consideration.

Not the whole story

One major shortcoming of tests of statistical significance is that they are blind to the study design. They do not tell the whole story. For example, if a treatment group consisted of only men over age 70 and the control group consisted of only women under age 50, it would be impossible to conclude whether the difference in groups was due to age, gender, or the treatment being studied. The researcher in this hypothetical study would still be able to calculate a p-value, but the result would not be useful.

Possible confounding factors such as age, gender, other underlying diseases, or differences in in sample handling, so called pre-analytical variation, need to be considered during the planning phase.

Considerations for different types of clinical studies

There are several distinct types of clinical study recognized, each of which has its own benefits and challenges. The scope can range from more basic disease vs control comparisons (case-control studies) to large cohort studies that collect a wealth of clinical data and may include a range of clinical end-points.

Studies can be prospective (where all details such as parameters to be measured or clinical interventions to be made are planned prior to the start of the study) or retrospective (where material from a completed study is examined or re-interrogated after the fact).

If samples are taken from multiple time points from the same individuals over the course of a study, it is described as longitudinal. We will now consider some aspects of study design that relate to the various types of clinical study.

Discovery and validation in cohort studies

Olink's panels are ideal for exploratory screenings of a large number of proteins, where the researcher is looking for patterns or relative differences between the analytes.

One example of this broad screening approach is the study performed by Bryan *et al*. where they ran five different Olink panels. They found novel proteins that were significantly associated with urothelial bladder cancer and possible prognostic staging marker candidates (3).

In clinical settings, cases are typically symptomatic and have undergone a variety of procedures leading to the diagnosis. Controls can consist of, or include, patients with other diseases, and there may be differences in sample collection and processing procedures between cases and controls. Depending on the sample collection demographics, there may also be a genetic bias generated within a specific cohort. All of these factors can potentially affect the findings.

The confidence level in the conclusions from a protein biomarker-based cohort study can be greatly enhanced by validating the findings in a second, independent cohort of samples. In one of many such published examples of this multi-cohort approach, Tromp *et al*. identified a number of protein markers that distinguish between the two major forms of heart failure. The proteins were first identified in a ~1500 sample discovery cohort before the findings were confirmed in an independent 850 sample validation cohort (4).

Batch-to-batch and run-to-run variation in longitudinal studies

In a longitudinal study, variables relating to an individual or group of individuals are assessed over a period of time, with continuous or repeated monitoring of risk factors and health outcomes.

A key consideration is batch-to-batch reproducibility. A change of reagent batch may introduce bias between samples run at different times. Similarly, even if the same batch is used a bias may be introduced between timepoints in the longitudinal study if laboratory conditions fluctuate. Therefore, it is common and recommended to include bridging samples that are run at every timepoint of the longitudinal study. The protein levels measured for these samples can be used as a common reference between batches and used to normalize and alleviate any potential bias.

TERMINOLOGY

Bridging samples are samples run on plates from the different batches. The bridging samples should represent the data set in regard to samples and controls. These samples are used for reference sample normalization. Olink recommends that at least 8 different bridging samples are used in such cases.

Make sure to have a sufficient sample supply for bridging through the study. Discuss the full study design with your Olink representative to ensure that the appropriate numbers of kits and bridging samples are included.

In the production of new reagent kits, Olink has QC processes to limit variation between batches, and to ensure a consistent performance. This is important since a new antibody batch could cause a shift in the data generated. To address this potential risk, Olink has introduced a thorough QC procedure with strict acceptance criteria.

However, when kits are used in longitudinal studies, potential signal differences across multiple batches have to be considered and adjusted for in the statistical analysis.

Olink continuously improves products and services. For the latest information on panel compositions and updates, contact your Olink representative or support@olink.com.

Sample considerations

Before conducting any study using Olink panels there are several things to bear in mind.

Samples and controls

Cases and controls within a study should all use the same sample matrix to be able to compare data between groups. It is not possible to directly compare data between matrices, but it is possible for example to determine whether a biomarker found in plasma can also be measured in CSF. The inclusion criteria for the control group needs to be designed to fit the study questions. The control group often consists of samples from healthy individuals, but it is not always the best option. For example, in a study assessing potential biomarkers for secondary prevention of diseases, healthy controls are not a good comparison.

Control group design can be tricky, as was pointed out by Yeh *et al*. They evaluated whether a matched sibling case and control study design would yield more statistical power in uncovering significant early diagnostic biomarkers for breast cancer. They hypothesized that sisters would serve as well-matched controls as they are naturally controlled for race, ethnicity and a large proportion of genetic background. However, samples from biological sisters did not generally appear to be more similar to each other than to other individual samples and were not well separated (5).

It is important to remember that proper documentation around sample handling can provide valuable input when interpreting the data. Read more in Olink's white paper *"Pre-analytical variation in protein biomarker research"* (1).

Quality control

Selection of appropriate quality controls must also be addressed. For many other multiplex immunoassays, validation steps are needed after the runs. To avoid this, Olink has developed a built-in QC system, using internal controls, for its multiplex biomarker panels. This system allows full control over the technical performance of assays and samples.

In addition to the controls provided by Olink, a pooled plasma sample or another customized bio matrix pool should be included on all plates to enable further QC. An example of additional QC is to assess potential variation between runs and plates, for example to calculate inter-assay and intra-assay CV. Read more about the QC system in the Olink white paper *"Data normalization and standardization"* (6).

Consider optimal dilutions

Each Olink panel is optimized for the expected physiological and pathological ranges found in plasma and serum for the 92 biomarker assays included. A few of our panels are designed to be run diluted (from 1:10 up to 1:2025) so as to fit the physiological ranges within the measurable ranges of our assays, these dilutions are a part of the assay and no pre-dilution is necessary.

Dilution factors are set with plasma and serum in mind, so when running other sample matrices, these dilutions may need adjustment depending on the expected protein concentrations in the specific matrix. Guidelines for sample preparation are available for a selected number of matrices. Please contact support@olink.com for further information.

When samples are to be sent to Olink Analysis Service, the requested dilution step of the samples will be performed in-house before the analysis.

For tissue or cell lysate, we recommend a range from 0.5–1.0 µg/µL of total protein concentration determined by, for example, BCA or Bradford. Cells can also be sent with a concentration indicated as cells/µL. To ensure that the proteins are measured in the optimal range for each assay, it is recommended to run samples in at least two different starting concentrations, (i.e. dilutions). It is important that all tissue or cell lysate samples that are to be analyzed together have the same starting concentration of proteins or cells.

Range of detection

The analytical measuring range needs to be determined to characterize the performance of the test. To understand the capabilities and limitations of Olink panels, the lower limit of quantification (LLOQ) and the upper limit of quantification (ULOQ) is listed for each assay in the validation data document for each panel. Use this information to ensure that the panel is fit for the specific purpose.

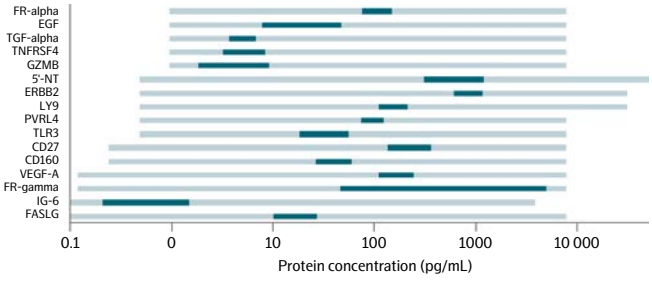

Fig 2. Example diagram of dynamic range with highlighted normal plasma levels.

Data sharing and comparing

Share data

One way to drive further improvement in the proteomics field's use of statistics is continued growth in open data sharing. Many research questions are addressed by multiple teams, and it may be misleading to emphasize the statistically significant findings of one single team.

There are several collaborative frameworks set up to share Olink data. For example, SCALLOP (www.olink.com/scallop) is a consortium for discovery and follow-up of genetic associations with proteins (pQTLs) on the Olink Proteomics platform (7).

To be a member of the SCALLOP consortium you have to be the principle investigator of a study collection with Olink and genome-wide genotyping data.

In a major study by Folkersen *et al*, it was recommended that a large sample size is needed to detect these genetic protein associations (8), and this shows why collaborative working environments are of increasing importance.

Compare NPX values and combine studies

To compare NPX values between studies it is important to remember that the studies need to be normalized to remove any bias between them. If the studies can be considered randomized or if the factors that differ between the studies are not important, median centering intensity normalization can be utilized. Otherwise, bridging samples should be used for normalization analogous to when combining data from timepoints in a longitudinal study described above. If possible, these bridging reference samples should be representative for the study, for example include all study groups, and should be matrix matched with other samples.

For more information see our white paper, *"Data normalization and standardization"* (6).

Maximize information output

In general there will normally be a requirement that key findings with clinical implications should be independently verified with another technique, irrespective of the original method used. If you need to do that with your important Olink findings, it is a good option to run a singleplex assay.

There are numerous examples showing how assays run in our 92-plex panels correlate very well with standard single ELISAs, and show equal or better performance (9).

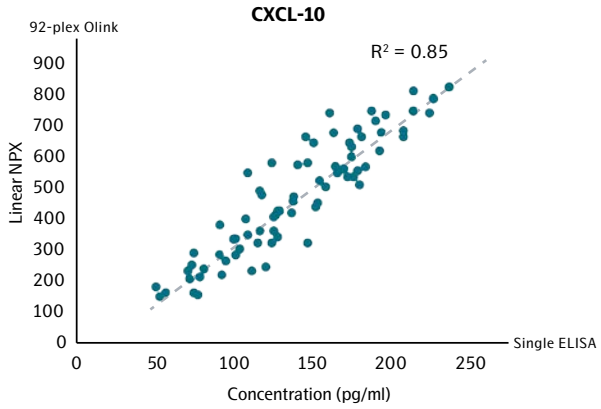

Fig 3. Correlation between conventional ELISA and Olink for CXCL-10.

Good correleation in plasma has also been shown when comparing an Olink assay with the equivalent SIMOA assay from Quanterix.

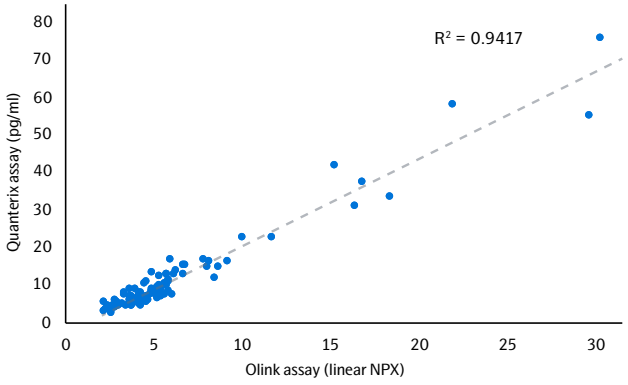

Fig 4. Correlation between Olink and SIMOA for NFL. Samples were supplied by courtecy of Prof. Tomas Olsson (KI, Sweden). The assay for Neurofilament light polypeptide uses the NF-light® antibodies from UmanDiagnostics, Umeå Sweden.

How Olink can help

As a service, Olink's data science team offers to discuss normalization approaches and study design before analysis starts, to help you get the most out of your experiment.They can also assist with customized statistical analysis and maximize the value and information output from your studies run using Olink panels.

References

1. Olink Proteomics, Pre-analytical variation in protein biomarker research, www.olink.com/white-papers
2. Dell *et al*., Sample Size Determination. Ilar Journal (2012)
3. Bryan *et al*., Multiplex screening of 422 candidate serum biomarkers in bladder cancer patients identifies syndecan-1 and macrophage colonystimulating factor 1 as prognostic indicators. Transl Cancer Res (2017)
4. Tromp *et al*., Multimorbidity in patients with heart failure from 11 Asian regions: A prospective cohort study using the ASIAN-HF registry. PLoS Med (2018)
5. Yeh *et al*., Assessing biological and technological variability in protein levels measured in pre-diagnostic plasma samples of women with breast cancer. Biomarker research (2017)
6. Olink Proteomics, Data normalization and standardization, www.olink.com/white-papers
7. Scallop consortium www.olink.com/scallop/
8. Folkersen *et al*., Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLOS Genetics (2017)
9. Siegbahn *et al*., A comparison of the proximity extension assay with established immunoassays in Advancing precision medicine: Current and future proteogenomic strategies for biomarker discovery and development. Science/AAAS (2017)

www.olink.com

Olink Proteomics, Dag Hammarskjölds väg 52B , SE-752 37 Uppsala, Sweden
1098, v2.0, 2021-06-21