UvA-DARE (Digital Academic Repository)

Statistical data processing in clinical proteomics
Berkenbos-Smit, S.

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Outlook

We have discussed some aspects of the analysis of clinical proteomics data. By tailoring the data analysis method (Chapters 6 and 7) it is possible to find effects in the data that would otherwise remain hidden. The combination of cross validation and permutation testing forms a thorough statistical validation which creates a solid foundation to continue developing differences between patient groups into clinically valuable biomarkers. Nevertheless, there remain many open issues regarding the analysis of proteomics data and we discuss some of those here. These issues are the subject of ongoing and future research. We briefly touched upon the issues of power calculations and increasingly complex data sets at the end of Chapter 2. In this chapter we elaborate on these issues.

Power calculations
Power calculations provide the relationship between sample size, effect size and the power of a statistical test. When the effect size is known or estimated, the sample size can be calculated given the power desired. An appropriate sample size, not too many or too few, gives rise to effective experimental designs at controlled costs. For clinical proteomics and other omics disciplines power calculations are not standard procedure. The reason for this is twofold. First, in clinical proteomics studies the effect size is usually unknown. The search for differentially expressed proteins is performed in a shotgun approach. Whether differentially expressed proteins will be measured and, if so, how large an effect can be expected is not known beforehand. Estimates for these could probably be obtained from pilot studies with 5-10 observations per class.\textsuperscript{159} The second problem stems from the high-dimensionality of the data. While power calculations are well developed in univariate analysis, results for multivariate data are very limited. Recently some results have been obtained for multiple testing problems\textsuperscript{102,159,160} using the (local) false discovery rate\textsuperscript{22} However, the issue is still open for high-dimensional data. Computer simulations using biological knowledge might be a good approach.
Increasing complexity of data sets
The improvement in mass spectrometry technology and the development of hyphenated techniques, for example liquid chromatography coupled to mass spectrometry (LC-MS, see for example Chapters 6 and 7) leads to ever more complex data sets. Different platforms and different measuring parameters, e.g. different columns, allow for measuring different parts of the proteome. Integration of the resulting data can be achieved at several levels. The data sets may be combined to form one larger set in which they are analyzed together. Alternatively, each set is analyzed separately and the results are combined to give an expanded view. Another form of increasingly complex data sets results from the integration of different types of ‘omics’ data, for example gene expression and proteomics data. The findings in one data set can be used to confirm findings in the other, or together they can bring to light new discoveries. The best method for fusing data sets remains a topic for future research.

Towards clinical use
The goal in clinical proteomics research is to find protein markers that are of clinical use, for example in population screening programs. Finding a protein that is differentially expressed in one experiment does not necessarily translate to a clinical application. Pepe identifies several phases of development for markers intended for population screening. The work presented in this thesis could be considered first phase studies where many leads are discovered and prioritized. Between this phase and actual use as a screening tool lie the phases of clinical assay development and evaluation. A challenge in these phases is setting acceptable thresholds for type I and type II errors (false positives and false negatives). A type I error means unnecessary psychological burden for the person tested falsely positive. In population screening, a test with a high type I error results in many costly follow-up procedures that would not have been performed without the screening programme. On the other hand, a high type II error leads to many people being falsely reassured. A good screening tool strikes an acceptable balance between the two.