Clara-Cecilie Günther

Statistical analysis of biological data – diagnostic tests, gene ontology and gene expression

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2009

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-471-1596-1 (printed ver.)
ISBN 978-82-471-1597-8 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2009:107

Printed by NTNU-trykk

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *philosophiae doctor* (PhD) at the Norwegian University of Science and Technology (NTNU). The research was funded by the Department of Mathematical Sciences, NTNU.

First and foremost I would like to thank my supervisor Mette Langaas for her invaluable support and guidance throughout these years. She has been such an enthusiastic and inspiring supervisor, always available to answer my questions. I would like to thank Øyvind Bakke for the countless hours he spent discussing with me during the final stressful months, my work benefited greatly from his suggestions and comments. I also thank Håvard Rue for his advice concerning Bayesian methodology and for letting me use his computing resources, and Stian Lydersen for sharing his knowledge of contingency tables and exact tests.

I am grateful to everyone I have collaborated with at the Department of Cancer Research and Molecular Medicine, in particular Christina Sæten Fjeldbo, Torunn Bruland, Irina P. Eide, Rigmor Austgulen and Vidar Beisvåg, for giving me the opportunity to participate in their exciting research projects.

Thanks to my colleagues at the Department of Mathematical Sciences, in particular past and present members of the statistics group, for providing a nice working environment and to the technical and administrative staff for always being helpful.

My sincere gratitude also goes to my parents and my brother for their continuous support and encouragement, and to my friends for being patient and for caring. Finally, a very special thank you to Patrick, for being there throughout all these years, you are the best!

Trondheim, April 2009

Clara-Cecilie Günther

Thesis outline

The thesis consists of the following papers, presented in chronological order.

Paper I: **GeneTools – application for functional annotation and statistical hypothesis testing.** Vidar Beisvåg, Frode K. R. Jünge, Hallgeir Bergum, Lars Jølsum, Stian Lydersen, Clara-Cecilie Günther, Heri Ramampiaro, Mette Langaas, Arne K. Sandvik and Astrid Lægreid. BMC Bioinformatics 2006, 7(1):470.

Paper II: **Fetal growth restriction is associated with reduced FasL expression by decidual cells.** Irina P. Eide, Christina V. Isaksen, Kjell Å. Salvesen, Mette Langaas, Clara-Cecilie Günther, Ann-Charlotte Iversen and Rigmor Austgulen. Journal of Reproductive Immunology 2007, 74, p. 7–14.

Paper III: **Functional studies on transfected cell microarray analysed by linear regression modelling.** Christina Sæten Fjeldbo, Kristine Misund, Clara-Cecilie Günther, Mette Langaas, Tonje Strømmen Steigedal, Liv Thommesen, Astrid Lægreid and Torunn Bruland. Nucleic Acids Research 2008, 36(15), e97.

Paper IV: **Comparison of predictive values from two diagnostic tests in large samples.** Clara-Cecilie Günther, Øyvind Bakke, Stian Lydersen and Mette Langaas. Preprint Series in Statistics no. 9, 2008, Department of Mathematical Sciences, NTNU, Trondheim, Norway.

Paper V: **Comparing positive predictive values for small samples with application to gene ontology testing.** Clara-Cecilie Günther, Øyvind Bakke and Mette Langaas. Preprint Series in Statistics no. 3, 2009, Department of Mathematical Sciences, NTNU, Trondheim, Norway.

Paper VI: **Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters.** Clara-Cecilie Günther, Øyvind Bakke, Håvard Rue and Mette Langaas. Preprint Series in Statistics no. 4, 2009, Department of Mathematical Sciences, NTNU, Trondheim, Norway.

Background

The thesis can be divided into two parts, the first one consists of paper I–III, it is biologically oriented and addresses specific biological problems, whereas the second part consists of paper IV–VI and is mainly statistically oriented with focus on hypothesis testing for categorical data.

Part I – Analysis of biological data

In molecular biology, DNA microarray is one of the present techniques used to discover genes that are differentially expressed between two or more conditions. DNA microarrays are small glass or plastic slides with single stranded DNA printed in spots in a rectangular grid across the array. Each spot represents a gene or part of a gene, and there may be thousands of spots on a single array. mRNA (messenger RNA), which carries information from DNA in the nucleus to the cytoplasm where the protein synthesis takes place, is extracted from a biological sample, e.g. a tissue sample, that represents the condition under study, and labelled with a fluorescent dye. This is called the target. The array is washed with a solution containing the target which will hybridize to the matching DNA strands on the array. Afterwards the array is scanned, and the fluorescence intensities of the spots in the resulting image is a measure of the amount of mRNA in the biological sample which is thereby a measure of the activity of the gene, Draghici (2003). The higher the intensity, the more expressed a gene is. With this technique, the gene expression of thousands of genes can be measured simultaneously. To analyze the data and to detect genes that are differentially expressed between the conditions studied, statistical methodology is needed. Before performing the experiment, an experimental design should be decided on to increase the quality of the measurements. After the experiment is finished, the microarray image is processed and the expression data normalized to remove systematic variation, Allison, Cui, Page and Sabripour (2006). To find differentially expressed genes, moderated T-tests that borrow strength across the genes to estimate the variance for each gene are commonly used, Smyth (2004). Since the hypothesis testing is carried out for many genes simultaneously, correction for multiple testing should be applied, Allison et al. (2006).

Pre-eclampsia and fetal growth restriction are examples of conditions that can be studied partly through DNA microarrays. Pre-eclampsia is a potentially deadly condition that affects about 2.5–3.0% of pregnant women, Redman and Sargent (2005). The pathology is not yet fully understood, but it has been found that pregnancies that are complicated by one or both of these conditions are characterized by superficial trophoblast invasion and insufficient spiral artery remodelling, Redman and Sargent (2005). To better understand the biological processes that causes these conditions, the gene expression of women with pre-eclampsia and/or fetal growth restriction can be compared to the gene expression of healthy pregnant women in search for genes that are differentially expressed between these groups.

The final output from a microarray experiment after processing the data is often one or more lists of significantly differentially expressed genes. These lists contain only the single genes with no information about possible connections between the genes and to interpret the results, the biologist would have to manually obtain information from many different sources and databases. Instead of treating each gene separately, it is of greater interest to study several genes simultaneously and look for biological pathways that are active in the conditions studied. Genes act together, and if several genes belonging to a gene set known to be a biological pathway are present on the lists, then this pathway is probably more active in one of the conditions compared to the other conditions. It is thus important to analyze the lists of genes resulting from the experiment in terms of the involvement of the genes in biological pathways or processes.

Gene Ontology (GO) is a controlled vocabulary that describes gene and gene product attributes, The Gene Ontology Consortium (2000), in terms of the three main categories: biological process, molecular function and cellular component. The GO is set up as a hierarchy with the three main categories on top, each one followed by a number of subcategories which is split into further subcategories. Each gene can belong to more than one GO category, it may e.g. have both a known molecular function and play a role in a biological process. There are several tools available for analysis of gene lists with respect to GO annotations that also include statistical evaluation of the GO categories present on the lists, an overview was given by Khatri and Draghici (2005). It is of particular interest to test whether a GO category is over-represented or depleted in one list of genes from a microarray experiment compared to another list of genes from the same experiment.

Even though DNA microarrays provide measurement of gene expression for a large number of genes, the analysis of functions of gene products in the living cell is often performed for one gene at a time. Transfected cell microarray is a different technique that allows analysis of protein functions in cells for many genes simultaneously, Ziauddin and Sabatini (2001). Examples of protein functions that can be analyzed include apoptosis, gene regulation and receptor binding. By using plasmids, i.e. circular DNA molecules containing the genes we want to study, or short interfering RNAs (siRNAs), i.e. double stranded RNA molecules that direct the breakdown of a specific mRNA molecule, the effects of over-expression or down-regulation of genes of interests can be studied. As an illustration of how the technique works, assume we are interested in genes related to apoptosis, i.e. programmed cell death. Plasmids containing the genes for which we want to study the effect on apoptosis are mixed with transfection reagent and other substances necessary for transfection, and printed on the array. To investigate the effect of many genes, many different plasmids are printed. Mammalian cells are then cultured on top. The cells growing on top of the printed areas are transfected and express the genes coded for in the plasmids, whereas the cells growing between spots are not transfected, Ziauddin and Sabatini (2001). After some time, e.g. 24-72 hours depending on the assay, the effect of over-expression or down-regulation is detected. In the apoptosis example, fragmented DNA can be labelled with fluorescence and the spots can be studied in a fluorescence microscope where the cells that have undergone apoptosis will be visible. Thus, genes that are over-expressed in those spots are probably apoptosis inducing genes. In other situations, the array is scanned with a laser scanner, resulting in an image of the fluorescence intensities in the spots which can be analyzed further. While a variety of statistical methods have been developed for analysis of DNA microarrays, the progress has been slower for tranfected cell microarrays. Normalization procedures and possible standard approaches for statistical analysis of the data are among the areas that need more attention.

Part II – Comparing positive predictive values

The problem of testing over-representation or depletion of GO categories between lists of differentially expressed genes is statistically the same situation as testing whether the positive predictive values of two diagnostics tests are equal. Diagnostic tests are used in medicine to e.g. detect diseases and can be evaluated by various measures. The sensitivity and specificity are common accuracy measures and tell us how likely the test results are given the true disease status. The positive predictive value is defined as the probability that a subject has the disease of interest given that the test result of the diagnostic test is positive. The negative predictive value is the probability that the subject does not have the disease given that the test result is negative. The predictive values give information about the prediction abilities of the diagnostic tests which is of important value when considering which test is better to use, in particular for the patients receiving treatment, Guggenmoos-Holzmann and van Houwelingen (2000).

Assume there are two diagnostic tests, test A and B, available for the same disease. The null hypothesis is that the positive predictive value for test A is equal to the positive predictive value for test B. This problem has been studied by Leisenring, Alonzo and Pepe (2000) who proposed a score test based on generalized estimating equations by fitting a generalized linear model. Wang, Davis and Soong (2006) also fitted a linear model using weighted least squares and Moskowitz and Pepe (2006) derived confidence intervals from normal asymptotic theory which lead to formulas for sample size estimation. Both Wang et al. (2006) and Moskowitz and Pepe (2006) assumed a multinomial distribution for the possible joint outcomes of the three variables disease status, test result for test A and test result for test B.

The available methods are all developed for large samples, using approximate asymptotic distributions. When the sample size is small, as it may be in smaller studies of diagnostic tests and in the GO setting, other solutions are needed.

Summary

Paper I presents the software GeneTools, which is a collection of three web-based gene related tools: NMC Annotation Tool, GO Annotator Tool and eGOn. These tools are implemented on top of a database containing annotations from several other databases like UniGene, Entrez Gene, Swiss Prot and GeneOntology. The database in GeneTools is updated weekly. The users submit gene reporter lists which may originate from e.g. a microarray experiment. GeneTools store these lists which makes it possible for the user to share the lists with other users and the lists can quickly be updated with the most recent annotation information. The NMC Annotation Tool provides access to the database through a single or batch-search mode after submitting the lists. The GO Annotator Tool allows the users to define their own GO annotations which are stored in the database and are available for further analysis. The eGOn (explore Gene Ontology) tool visualizes gene annotations in the GO hierarchy and provides statistical hypothesis tests applied to the submitted lists. Hypothesis testing is available for three different situations, when one of the lists is a master list, when the two lists are mutually exclusive and when they are partially overlapping. All the tests pinpoint GO categories where the probability of belonging to the GO category is different for gene reporters on one list list compared to reporters on a second list, i.e. testing if certain GO categories are over-represented or depleted in one list compared to the other.

Paper II studies FasL expression and apoptosis in decidual tissue from pregnancies with impaired placental development. The interaction between the Fas ligand (FasL) and its receptor Fas is known to be inducing apoptosis in the Fas-expressing cell. The Fas-FasL system has previously been ascribed a role in implantation and placental development by regulating trophoblast invasion and spiral artery remodelling and the expression of FasL might therefore differ in these pregnancies compared to normal pregnancies. The study groups were cases with preeclampsia and/or fetal growth restriction and controls with normal pregnancies. Expression of genes involved in the main apoptotic pathways was studied using Affymetrix GeneChip arrays and the statistical analysis was based on summary expression measures for each probe set. However, the expression of these genes did not differ significantly between the two groups. The proportion of FasL-expressing decidual cells was studied through cell counting, which showed that it was high in controls and significantly reduced in the cases.

Paper III provides a suggestion on how to analyze data from transfected cell microarrays using measurements from all spots and all experimental replicates in a study. The analysis includes measurement of fluorescence intensities, normalization and linear regression modelling. The signals were first log-transformed before normalized using negative control spots and internal control plasmids. A linear regression model that models the effect of the conditions, i.e. different nucleic

acids printed on the array, different treatments, i.e. external stimuli added to the cells for the induction of gene expression, and experimental replicates was fitted to the data. P-values were calculated for these effects and for comparisons of interest. Three studies with increasing complexity of known biological effects were performed and for each these studies simulation experiments were carried out to evaluate the number of technical and experimental replicates necessary to detect the known biological effects.

Paper IV is the first of three that address the problem of comparing positive or negative predictive values for two diagnostic tests. The statistical problem is to test whether two conditional probabilities expressed by the parameters of the multinomial distribution are equal. The existing tests for large samples are compared to the proposed likelihood ratio test through simulation experiments and it is also suggested that using maximum likelihood estimates under the null hypothesis can improve upon the existing tests. Finding the maximum likelihood estimates under the null hypothesis requires optimization of a non-linear function satisfying specific constraints and can be challenging. Two approaches are presented. The simulation experiments showed that the improved existing test performs well in terms of test size.

In **paper V**, the statistical problem is the same as in paper IV, but the application is within gene ontology testing. The sample size at each Gene Ontology category is often small and since the large sample tests having asymptotic distributions do not preserve their test size when the sample size is small, parametric bootstrapping is proposed as a method to find the distribution of the test statistics in the small sample case. The performance of four small sample parametric bootstrap tests are evaluated through simulation experiments and compared to their large sample alternatives. Both the parametric bootstrap small sample and asymptotic tests are applied to a biological example using the eGOn software from paper I.

Paper VI presents further work on the statistical problem from paper IV and V, using enumeration as an exact approach for small samples. The probability of a multinomial outcome is not completely specified by the null hypothesis because of the presence of nuisance parameters. To calculate p-values, however, it is necessary to calculate these probabilities, but in order to do so, the nuisance parameters must be dealt with. Different approaches, including maximization, estimation, integration and combinations thereof are discussed and compared in terms of test size and power for a variety of test statistics including the test statistics from paper IV. The methods in this paper are general and can be applied to other finite discrete distributions and null hypotheses as well.

References

Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus, *Nature Reviews Genetics* **7**: 55–65.

Draghici, S. (2003). *Data analysis tools for DNA microarrays*, Chapmann & Hall/CRC, chapter 2.

Guggenmoos-Holzmann, I. and van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity, *Statistics in Medicine* **19**: 1783–1792.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics* **21**(18): 3587–3595.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* **56**: 345–351.

Moskowitz, C. S. and Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs, *Clinical Trials* **3**: 272–279.

Redman, C. and Sargent, I. L. (2005). Latest advances in understanding preeclampsia, *Science* **308**: 1592–1594.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology* **3**. Iss. 1, Article 3.

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology, *Nature Genetics* **25**: 25–29.

Wang, W., Davis, C. S. and Soong, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares, *Statistics in Medicine* **25**: 2215–2229.

Ziauddin, J. and Sabatini, D. M. (2001). Microarrays of cells expressing cDNAs, *Nature* **411**: 107–110.

Paper I

BMC Bioinformatics

BioMed Central

Software

Open Access

GeneTools – application for functional annotation and statistical hypothesis testing

Vidar Beisvag*[1], Frode KR Jünge[1], Hallgeir Bergum[1], Lars Jølsum[1], Stian Lydersen[1], Clara-Cecilie Günther[2], Heri Ramampiaro[3], Mette Langaas[2], Arne K Sandvik[1,4] and Astrid Lægreid[1]

Address: [1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway, [2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway, [3]Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway and [4]Department of Medicine, St. Olav's University Hospital, Trondheim, Norway

Email: Vidar Beisvag* - vidar.beisvag@ntnu.no; Frode KR Jünge - frode.junge@ntnu.no; Hallgeir Bergum - hallgeir.bergum@ntnu.no; Lars Jølsum - lars@jolsum.net; Stian Lydersen - stian.lydersen@ntnu.no; Clara-Cecilie Günther - claracec@math.ntnu.no; Heri Ramampiaro - heri@ntnu.no; Mette Langaas - mettela@math.ntnu.no; Arne K Sandvik - arne.sandvik@ntnu.no; Astrid Lægreid - astrid.lagreid@ntnu.no

* Corresponding author

Published: 24 October 2006

BMC Bioinformatics 2006, **7**:470 doi:10.1186/1471-2105-7-470

Received: 19 July 2006
Accepted: 24 October 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/470

Abstract

Background: Modern biology has shifted from "one gene" approaches to methods for genomic-scale analysis like microarray technology, which allow simultaneous measurement of thousands of genes. This has created a need for tools facilitating interpretation of biological data in "batch" mode. However, such tools often leave the investigator with large volumes of apparently unorganized information. To meet this interpretation challenge, gene-set, or cluster testing has become a popular analytical tool. Many gene-set testing methods and software packages are now available, most of which use a variety of statistical tests to assess the genes in a set for biological information. However, the field is still evolving, and there is a great need for "integrated" solutions.

Results: *GeneTools* is a web-service providing access to a database that brings together information from a broad range of resources. The annotation data are updated weekly, guaranteeing that users get data most recently available. Data submitted by the user are stored in the database, where it can easily be updated, shared between users and exported in various formats. *GeneTools* provides three different tools: i) *NMC Annotation Tool*, which offers annotations from several databases like UniGene, Entrez Gene, SwissProt and GeneOntology, in both single- and batch search mode. ii) *GO Annotator Tool*, where users can add new gene ontology (GO) annotations to genes of interest. These user defined GO annotations can be used in further analysis or exported for public distribution. iii) eGOn, a tool for visualization and statistical hypothesis testing of GO category representation. As the first GO tool, eGOn supports hypothesis testing for three different situations (master-target situation, mutually exclusive target-target situation and intersecting target-target situation). An important additional function is an evidence-code filter that allows users, to select the GO annotations for the analysis.

Conclusion: *GeneTools* is the first "all in one" annotation tool, providing users with a rapid extraction of highly relevant gene annotation data for e.g. thousands of genes or clones at once. It allows a user to define and archive new GO annotations and it supports hypothesis testing related to GO category representations. *GeneTools* is freely available through www.genetools.no

Background

Microarray technology allows researchers to monitor transcript levels of thousands of genes in a single experiment [1]. Typically it confronts the researcher with vast amounts of numerical data as a starting point from which to begin to investigate how molecular mechanisms are involved in a specific biological setting. Typically, scientists have to manually query several resources/databases for information. Although these can be highly informative individually, the collection of available content would be more useful if provided in an integrated manner. High-throughput, automated annotation summaries can expedite this step and today several resources like Source [2], GeneCards [3] and NetAffx [4] already offer this.

In order to understand how cells function within a tissue, e.g. in a given state one can use data-driven methods, such as hierarchical clustering and self-organizing maps [5,6], which identify groups of genes with similar expression patterns. However, a complementary approach is to view data at the level of biological background knowledge such as a gene's involvement in a biological processes or pathway. The leading controlled vocabulary for such functional information is Gene Ontology (GO) [7]. Annotation of genes with GO terms creates a biological knowledge profile, in three layers dependent on the top-level GO branch used (biological process, molecular function or cellular component).

Several tools are suited for analysis of the GO hierarchy and for statistical evaluation of GO category representations between gene lists [8]. Comparisons of gene lists are important in order to answer questions such as "are genes involved in process P overrepresented among the total of differentially expressed genes in an experiment" or "does treatment A induce more genes involved in process P than treatment B?".

A potential problem using such tools, is that the existing annotation databases are incomplete and for most organisms only a subset of the known genes are functionally annotated [8]. Moreover, a major part of the available annotations e.g. those inferred from electronic annotations may be imprecise or incorrect.

The present paper describes *GeneTools*, a package of web-based tools for gene annotation. *GeneTools* is built on top of an underlying database that is updated on a weekly basis to provide information as recent as possible. The annotation data is accessible through two user interfaces, the *NMC Annotation Tool* which offers general functional annotation information in both single- and batch search mode, and the *e*GOn tool which can annotate, display and perform statistical hypothesis testing to assess the degree of similarity of GO category representation between different gene lists. An important function in *e*GOn is the possibility to filter on evidence codes. Also, additional user defined GO annotations can be added to the database through the *GO Annotator Tool* for use in further analysis. Another unique feature in *GeneTools* is that user submitted data is stored in the database and can be shared with other users.

Finally, a significant part of this paper deals with how the hypothesis testing for GO category representations is performed, which we think has been inadequately described for many other resources.

Implementation

GeneTools is a web service. It runs on most web browsers, including IE 5.0 or higher, Netscape 7 or higher and Mozilla Firefox 1.0 or higher, and is platform-independent. *GeneTools* is implemented in the PHP programming language. We have chosen to implement this tool as a web service to make it as user-friendly as possible, as most of the users are not bioinformaticians able to perform programming. However, more advanced use of the service is possible as described later in this chapter.

GeneTools is the front-end of a MySQL database containing annotation data from the following publicly available resources: UniGene [9], EntrezGene [10] (including GOA [11], Proteome, MGD [12] and RDG [13] annotations), Gene Ontology [14], SwissProt [15], and HomoloGene [16]. Information from 64 organisms available through UniGene is included, but the most comprehensive information is available for human, rat and mouse genes. All these databases are stored as local copies, enabling quick access to the data in response to the user query. Since many of the resources on which *GeneTools* draws continuously change their information content, the *GeneTools* database is updated on a weekly basis to ensure that it contains the most up-to-date information, continuously updating the stored gene reporter lists. An automated process checks for updates of the outside databases, downloads these files, and populates database tables accordingly. This ensures that the connections between external databases made within *GeneTools* are as accurate as possible. Thus, both the mapping of clones to genes and the functional attributes associated with those genes are dynamic and current. All data and graphics from searches and analysis can be exported in various formats (txt, XML or as Excel files).

Due to the heterogeneous nature of annotation information, bioinformaticians and systems biology researchers may want to perform more high-level analysis than offered through our web service. We therefore offer an API solution, based on web services description language (WSDL), for external resources wishing to use data from our database. Typically new and important tools like Taverna [17] can easily utilize this system using SOPE/RPC. Currently our API solution is utilized by the Norwegian Microarray Consortium (NMC) which updates their local BASE (BioArray Software Environment) [18] servers with information from this database. Moreover, SciCraft [19], a general data analysis tool, uses data from the *GeneTools* database in its microarray data analysis tool box. We will also offer R code for the statistical testing in *e*GOn upon request. The structure of our *GeneTools* database is built so that it can be used in the future as part of local or external data warehouses.

Results and discussion

Inputs

Figure 1 gives an overview of *GeneTools* with its single search and batch search (gene reporter lists) inputs, its underlying database structure and associated tools for analysis. The ability to simultaneously collect data from numerous sources for e.g. thousands of genes from microarray experiments in batch is especially important and made very user friendly through *GeneTools*.

Single search

The database enables searching by gene symbols/names, GenBank accession numbers, UniGene cluster IDs, Swiss-Prot entry names and several unique clone IDs (IMAGE clone IDs, University of Iowa clone IDs, Operon oligo IDs, TAIR IDs and a subset of selected Affymetrix and Agilent IDs).

The names and symbols of genes/proteins may be highly ambiguous [20]. We therefore recommend using primary gene IDs, like GeneBank accession numbers or specific probe IDs when querying the database. However, if gene names or symbols are used, caution is advised because only official names/symbols associated with UniProt knowledgebase will be recognized.

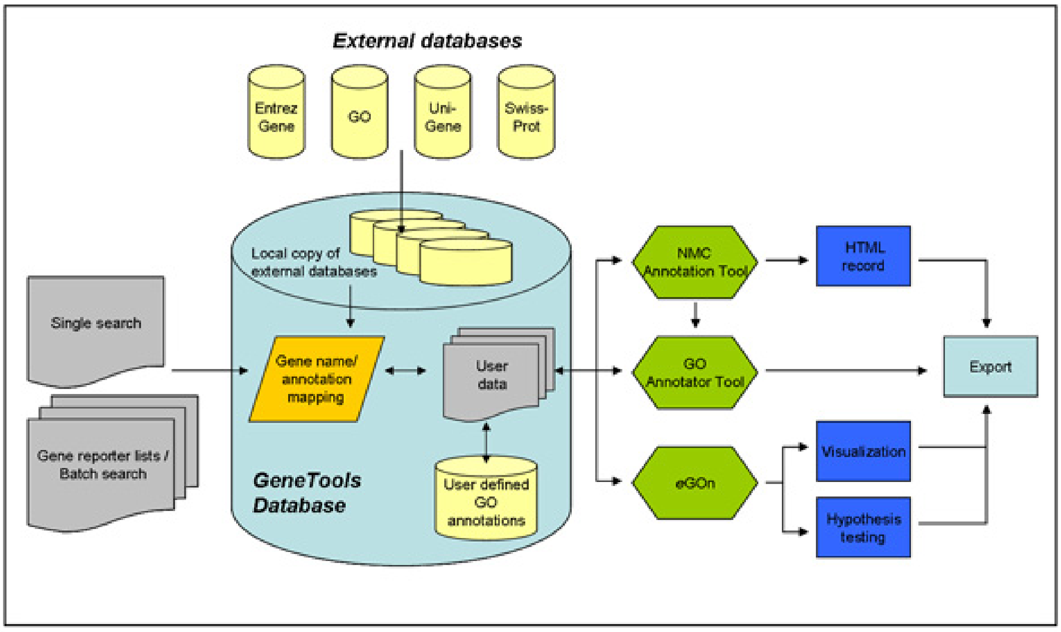


Figure 1
Flowchart of the *GeneTools* program and the underlying database. The underlying database is updated on a weekly basis with annotation information from several external databases including UniGene, Swiss-Prot, Entrez Gene and GO. User data are submitted to the database as text files of gene reporters and analysis of the annotation data can be performed through three user interfaces: the *NMC Annotation Tool*, the *GO Annotator Tool* and eGOn. Analysis results and annotation data can be exported in various formats.

Batch search

Input of gene reporter lists for batch search is done by uploading tab-delimited text files to the server. After submission, the gene reporters are automatically mapped to a UniGene cluster, and functional annotations/attributes (e.g. GO annotation) are associated with the specific gene/protein (Figure 1). Uploaded gene reporter lists are stored and can easily be managed in folders or shared with other users. If new annotation information becomes available for any of the stored gene reporter lists, the user will be notified.

Updates

The user may at any time choose to update a stored gene reporter list, thus incorporating the most recent annotation information from the weekly update of the *GeneTools* database in the analysis. The updating process is fast even for lists of thousands of gene reporters. The user receives a specified report detailing which gene reporters are associated with new annotation information and the changes made.

Tools, analyses and outputs

NMC annotation Tool

A major challenge when using genomic scale methods like microarrays, is to handle annotation information from the resulting comprehensive gene reporter lists. Thus, one of the most important features of *GeneTools* is the ability to simultaneously extract pre-existing annotation data from a wide variety of database resources for thousands of genes in a batch. Since the *GeneTools* database is weekly updated and the *NMC Annotation Tool* provides user friendly functionalities for associating new annotation information with the reporters in uploaded gene lists, the *NMC Annotation Tool* is particularly useful when it is important to always have access to the most recent information on the genes and clones being examined. The *NMC Annotation Tool* enables the user to query the *GeneTools* database by singe gene search or by batch search after submission of a gene reporter list for a microarray experiments. Given the massive amount of data available through *GeneTools* (Figure 1), information overload can be a potential problem. Therefore, we have provided the user with the option to select (in the "preferences" menu) the information to be shown on the screen for single search and batch view and to select which information to export. However, we will stress that this option should be used cautiously, because it may introduce selection bias and important information may be lost.

Single search outputs

The single search function captures the collection of features attributable to the given gene and its products, when a gene is defined by a unique UniGene cluster. Whenever available, each single search result view will contain all or a subset of the following categories of data:

I. Data from Unigene, including e.g. A. gene cluster, name and symbol. B. protein similarities with selected organisms (with direct link to Entrez protein). C. chromosome localization information. D. UniGene associated sequences with cluster.

II. Data from Homologene: Shows homologous genes for human, rat and mouse.

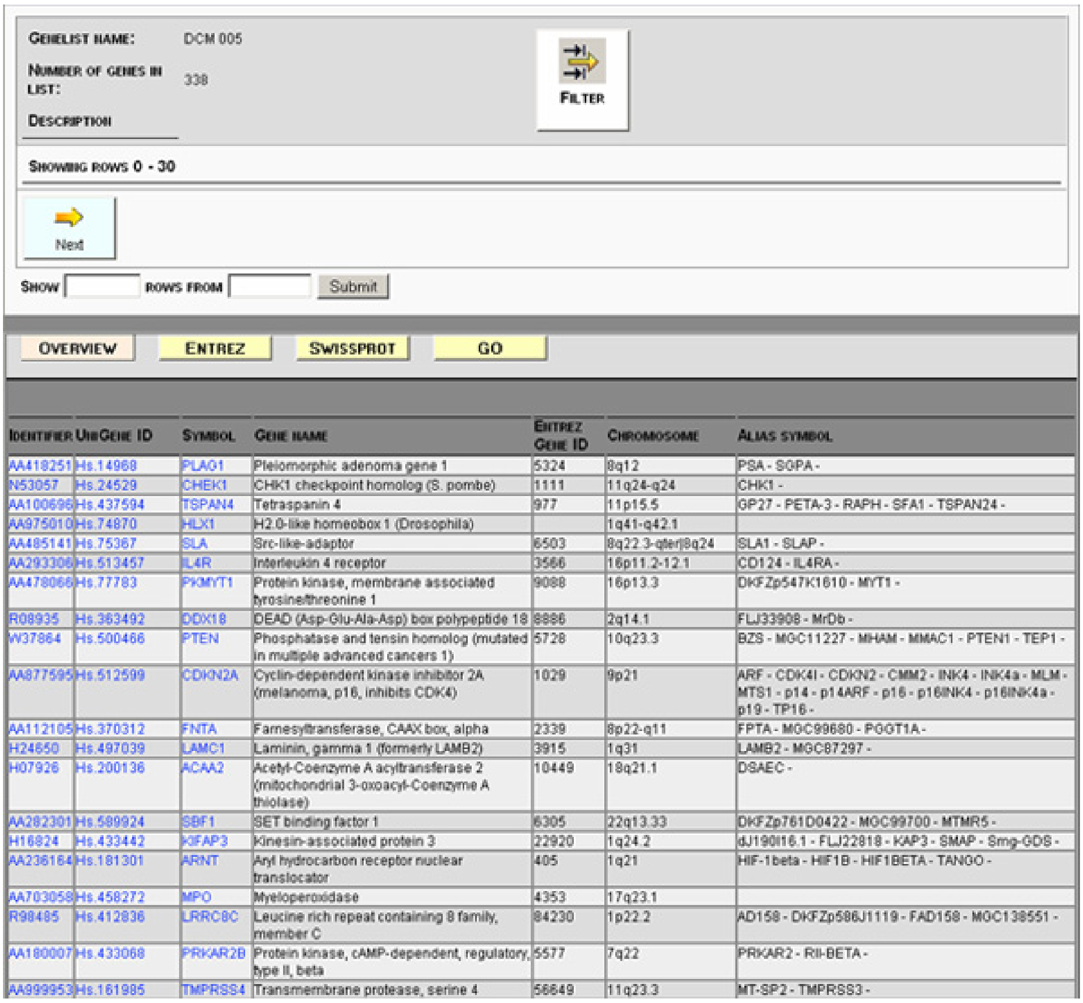
III. Data from Entrez Gene: A. gene name, symbol and aliases. B. biological roles and summary of functions curated by Entrez (Ref.seq summary). C. gene ontology (GO) annotations with direct link to references and links to alternative ontologies like KEGG. D. direct link to curated PubMed Gene RIFs (reference into function).

IV. Data from Swiss Prot: A. protein names and aliases. B. biological role and function information curated by Swiss Prot. C. protein sequence information. D. direct links to various external sources associated with current protein are offered for each gene reporter.

Batch search outputs

One of the most important and unique features of the *NMC Annotation Tool* is the batch search mode which utilizes all of our database sources for gene reporter lists from microarray experiments. For instance, the users can easily extract biological function, chromosomal localization, and get access to publications (GeneRIFs) that describe gene functions. The results for reporter gene list from a batch search can be viewed in a user-friendly tabular form (Figure 2). Moreover, the annotation data displayed on the screen are associated with hyperlinks to the underlying database or to the single search view. The annotation data can be exported in several formats for printing or storage (XML and XLS).

NMC Annotation Tool provides several features not available in other gene annotation tools. To our knowledge, few other application stores users' gene reporter lists allowing update of the reporter lists at any time with the most recent UniGene, Entrez Gene and GO information. This is important since the clusters in UniGene change rapidly and new GO annotations are being added continuously. To achieve this, the submitted gene reporter lists can easily be updated with all new information. Information about the external databases included in *GeneTools* and their last updates can be found from a link named "database status" in the menu, and provides useful documentation for publishing purposes. Information about commercial arrays supported by *GeneTools* (currently Affymetrix, Operon and Agilent) is also given. To our knowledge, a similar



Genelist name: DCM 005

Number of genes in list: 338

Description

Showing rows 0 - 30

Show ____ rows from ____

Identifier	UniGene ID	Symbol	Gene name	Entrez Gene ID	Chromosome	Alias symbol
AA418251	Hs.14968	PLAG1	Pleiomorphic adenoma gene 1	5324	8q12	PSA - SGPA -
N53057	Hs.24529	CHEK1	CHK1 checkpoint homolog (S. pombe)	1111	11q24-q24	CHK1 -
AA100696	Hs.437594	TSPAN4	Tetraspanin 4	977	11p15.5	GP27 - PETA-3 - RAPH - SFA1 - TSPAN24 -
AA975010	Hs.74870	HLX1	H2.0-like homeobox 1 (Drosophila)		1q41-q42.1	
AA485141	Hs.75367	SLA	Src-like-adaptor	6503	8q22.3-qter|8q24	SLA1 - SLAP -
AA293306	Hs.513457	IL4R	Interleukin 4 receptor	3566	16p11.2-12.1	CD124 - IL4RA -
AA478066	Hs.77783	PKMYT1	Protein kinase, membrane associated tyrosine/threonine 1	9088	16p13.3	DKFZp547K1610 - MYT1 -
R08935	Hs.363492	DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	8886	2q14.1	FLJ33908 - MrDb -
W37864	Hs.500466	PTEN	Phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	5728	10q23.3	BZS - MGC11227 - MHAM - MMAC1 - PTEN1 - TEP1 -
AA877595	Hs.512599	CDKN2A	Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	1029	9p21	ARF - CDK4I - CDKN2 - CMM2 - INK4 - INK4a - MLM - MTS1 - p14 - p14ARF - p16 - p16INK4 - p16INK4a - p19 - TP16 -
AA112105	Hs.370312	FNTA	Farnesyltransferase, CAAX box, alpha	2339	8p22-q11	FPTA - MGC99680 - PGGT1A -
H24650	Hs.497039	LAMC1	Laminin, gamma 1 (formerly LAMB2)	3915	1q31	LAMB2 - MGC87297 -
H07926	Hs.200136	ACAA2	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)	10449	18q21.1	DSAEC -
AA282301	Hs.589924	SBF1	SET binding factor 1	6305	22q13.33	DKFZp761D0422 - MGC99700 - MTMR5 -
H16824	Hs.433442	KIFAP3	Kinesin-associated protein 3	22920	1q24.2	dJ190I16.1 - FLJ22818 - KAP3 - SMAP - Smg-GDS -
AA236164	Hs.181301	ARNT	Aryl hydrocarbon receptor nuclear translocator	405	1q21	HIF-1beta - HIF1B - HIF1BETA - TANGO -
AA703058	Hs.458272	MPO	Myeloperoxidase	4353	17q23.1	
R98485	Hs.412836	LRRC8C	Leucine rich repeat containing 8 family, member C	84230	1p22.2	AD158 - DKFZp586J1119 - FAD158 - MGC138551 -
AA180007	Hs.433068	PRKAR2B	Protein kinase, cAMP-dependent, regulatory, type II, beta	5577	7q22	PRKAR2 - RII-BETA -
AA999953	Hs.161985	TMPRSS4	Transmembrane protease, serine 4	56649	11q23.3	MT-SP2 - TMPRSS3 -

Figure 2

Typical "overview" result output for a submitted gene reporter list. Input gene reporter and associated UniGene cluster, gene name, symbol and chromosome localization is shown for all the gene reporters in the submitted lists. Several of the information boxes are hyperlinked redirecting the user to the original source. More specific annotations can be found under the "tabs" named Entrez, SwissProt and GO. By clicking on the gene reporter ID, a single search window for the selected gene reporter will appear.

variety of important features is not available in gene annotation tools like Source [2], GeneCards [3], NetAffx [4], GeneCruiser [21], Onto-Tools [22], GARBAN [23] and GeneLynx [24].

GO annotator tool (user defined GO annotations)

The introduction of Gene Ontology (GO) [25] as a standardised vocabulary for describing genes, gene products and their biological functions represents an important milestone in the possibilities to handle and include biological background information in functional genomics analyses. Many databases today provide GO annotations for a variety of organisms including humans and other species. However, GO is still incomplete and significant extensions to its structure are needed before all available biological knowledge can be represented as GO annotations in public databases. Also, besides the human research filed other organisms e.g. common model organisms like rat and mouse are still lagging behind when it comes to raising the quality of curation of GO annotations. Thus, a high proportion of GO annotations offered in the rat genome database (RGD) [13] and the mouse genome database (MGD) [12] are associated with the IEA (inferred by electronic annotation) evidence code, which implies a lower degree of certainty than some users may require.

To overcome at least some of these problems, *GeneTools* allows a user to define their own GO annotations to genes of interest. The *GO Annotation Tool* (accessible through "single search" mode in the *NMC Annotation Tool*) enables the addition of new, user defined GO annotations as well as the curation of GO annotations e.g. annotations with evidence code IEA. *GO Annotation Tool* is supported by a GO term search system, simplifying the browsing for GO terms. Evidence codes and references (e.g. PMID) according to GO standards and free text can be added (Figure 3). New annotations are stored in the database and can be included in further analysis (e.g. added to the GO analysis in the *e*GOn tool). We are in the process of making an export function, where these user defined GO annotations can be exported to the GOA database [11] by an email service. GOA will curate these annotations and make them available for others through the GO annotation database [26].

Explore Gene Ontology (eGOn)

Controlled vocabularies facilitate query and retrieval of knowledge from many different sources using a common query structure. Three separate important activities are needed to enable this: the production and maintenance of the ontologies themselves; the creation of associations (or annotations) between the GO terms and gene products, and the development of tools that facilitate the creation, maintenance and use of the ontologies.

*e*GOn visualizes gene annotations in the GO hierarchy and offers a collection of statistical tests that translate the GO annotation information associated with the reporters in gene lists from functional genomics experiments to provide insight into the biological mechanisms involved.

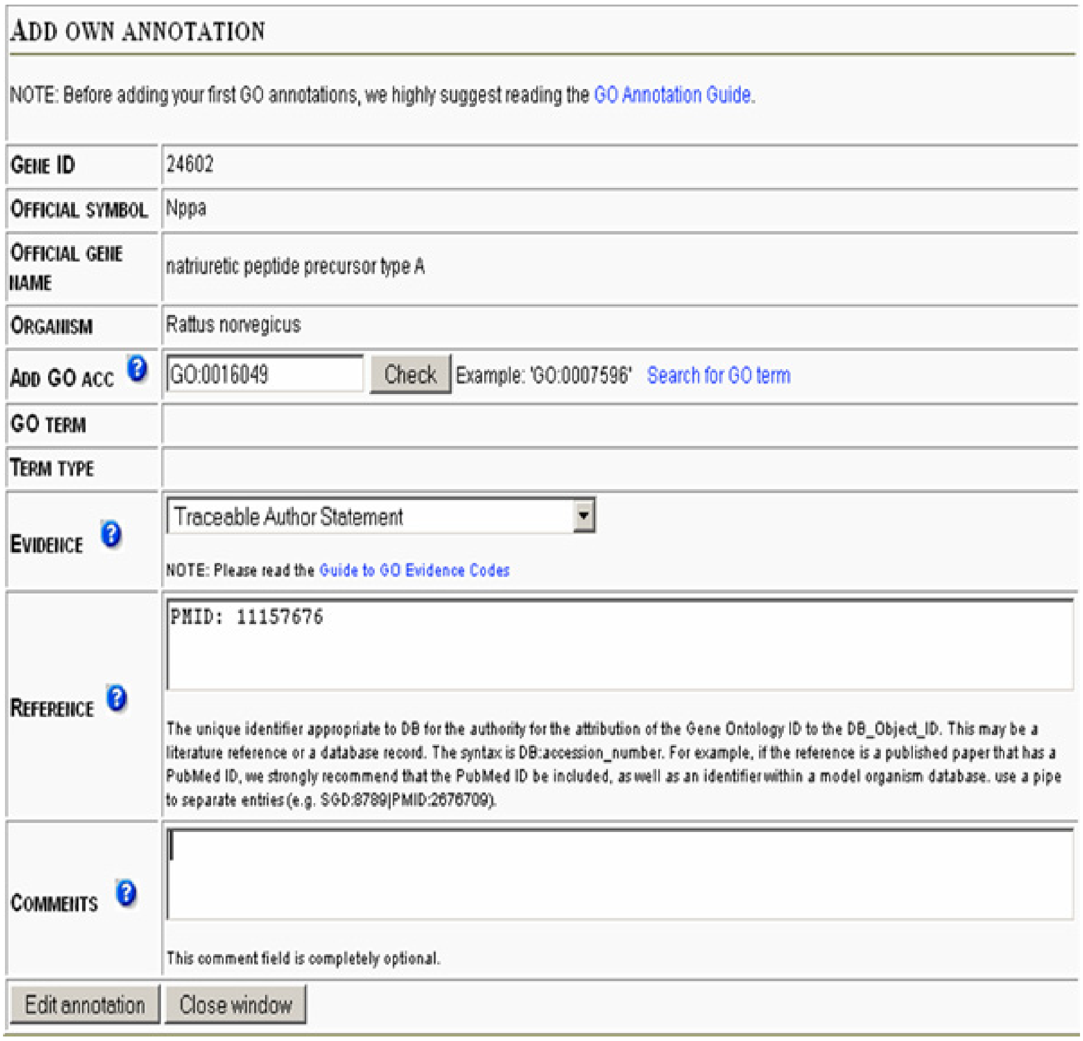
A wide range of resources are available for GO analysis [27]. In a recent review, Khatri et al. [8] question how such resources are built and used. Khatri et al. point out that existing annotation databases are incomplete, that a proportion of the annotations may be imprecise or incorrect, that name space mapping (how to connect a probe sequence to a gene/protein) is a problem, and that available statistical tests are not always validated. We think that the tool *e*GOn of the *GeneTools* suite meets many of these challenges since it enables filtering of annotations by evidence code, it allows the entry of new annotations and curation via the *GO Annotator Tool* and it provides a series of robust statistical tests that are thoroughly validated and documented.

For GO annotations, *GeneTools* uses Entrez Gene which offers curated data from the GO database that includes all registered GO annotations [26]. Some annotations available in the GO database will not be included using the Entrez curated GO annotations but the quality of annotation is most likely better. *e*GOn offers the possibility to filter the GO annotations from a gene reporter list by evidence codes. A substantial proportion of GO annotations are inferred by electronic methods (evidence code IEA), potentially being imprecise and possibly biasing further analysis. Thus, in a given analysis, it may be beneficial to exclude IEA annotations and only use more robust annotations, like e.g. annotations derived from "traceable author statement" (TAS), "inferred from direct assay" (IDA) or "inferred by curator" (IC). In other situations it may be desirable to include electronic annotations in order to obtain a sufficient amount of data to do a valid analysis, e.g. for rat and mouse genes where most of the annotations up to now are IEA. Another possibility which to our knowledge is not in use by any GO analysis tool today, might be to perform some kind of weighting by the type of evidence code for the statistical calculations.

An essential feature of *e*GOn is the possibility to compare and analyze annotated genes from two or more gene reporter lists in the GO-tree. *e*GOn both visualizes these comparisons within the GO-tree and formally calculates the degree of GO category representation similarity between the gene lists using statistical tests (Figure 4).

Testing statistical hypotheses of association between gene reporter lists

To investigate and better interpret the relevance of biological annotations of lists of gene reporters, statistical hypothesis testing can be a valuable tool. Let us for example consider a microarray experiment where the objective of the study is to compare the differentially expressed genes from heart failure tissue between cases and controls where the cases are patients with coronary artery disease (CAD) or dilated cardiomyopathy (DCM) and the controls are tissue from non-failing hearts [28].



ADD OWN ANNOTATION

NOTE: Before adding your first GO annotations, we highly suggest reading the GO Annotation Guide.

GENE ID	24602
OFFICIAL SYMBOL	Nppa
OFFICIAL GENE NAME	natriuretic peptide precursor type A
ORGANISM	Rattus norvegicus
ADD GO ACC	GO:0016049 Example: 'GO:0007596'
GO TERM	
TERM TYPE	
EVIDENCE	Traceable Author Statement NOTE: Please read the Guide to GO Evidence Codes
REFERENCE	PMID: 11157676 The unique identifier appropriate to DB for the authority for the attribution of the Gene Ontology ID to the DB_Object_ID. This may be a literature reference or a database record. The syntax is DB:accession_number. For example, if the reference is a published paper that has a PubMed ID, we strongly recommend that the PubMed ID be included, as well as an identifier within a model organism database. use a pipe to separate entries (e.g. SGD:8789|PMID:2676709).
COMMENTS	This comment field is completely optional.

Figure 3

User interface for the *GO Annotator Tool*. To add a new GO annotation, the user selects a gene, adds a GO term, chooses an appropriate evidence code and adds a reference article (PMID). The GO annotations are then stored in the database and an exported function to GOA for world wide distribution is under development. A link to the *GO Annotator Tool* can be launched from the top of the page of the result window from a single gene search, in the *NMC Annotation Tool* mode.

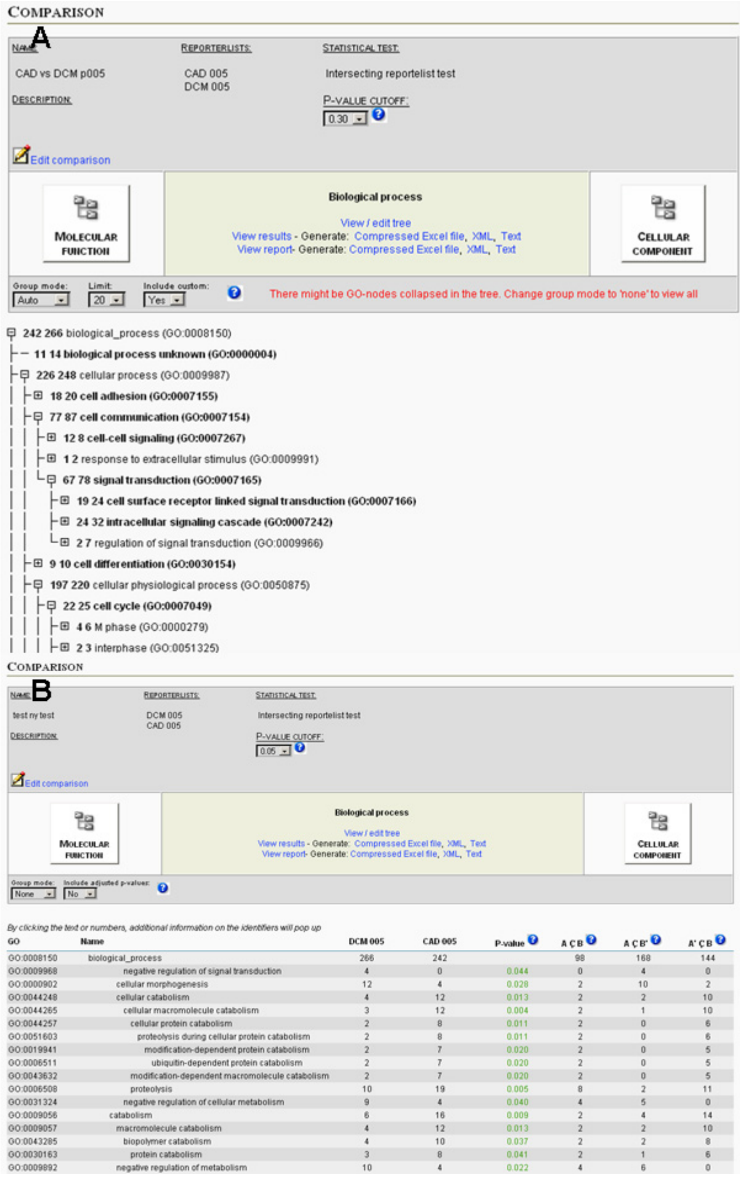
To formally state the statistical hypothesis, consider a randomly chosen gene and a given GO category denoted G. Define the following three events:

- A = the gene is in gene reporter list A

- B = the gene is in gene reporter list B

- G = the gene is a member of GO category G.

In this example the list A would be the list of differentially expressed genes between CAD and controls while list B would be the differentially expressed genes between DCM and controls. At the given GO category G (e.g. catabolism), we are interested in investigating whether the prob-



COMPARISON

NAME: CAD vs DCM p005

DESCRIPTION:

REPORTERLISTS: CAD 005 DCM 005

STATISTICAL TEST: Intersecting reportelist test

P-VALUE CUTOFF: 0.30

Edit comparison

Biological process

View / edit tree
View results - Generate: Compressed Excel file, XML, Text
View report- Generate: Compressed Excel file, XML, Text

Group mode: Auto Limit: 20 Include custom: Yes

There might be GO-nodes collapsed in the tree. Change group mode to 'none' to view all

```
⊟ 242 266 biological_process (GO:0008150)
├─ 11 14 biological process unknown (GO:0000004)
├⊟ 226 248 cellular process (GO:0009987)
│ ├⊞ 18 20 cell adhesion (GO:0007155)
│ ├⊟ 77 87 cell communication (GO:0007154)
│ │ ├⊞ 12 8 cell-cell signaling (GO:0007267)
│ │ ├⊞ 1 2 response to extracellular stimulus (GO:0009991)
│ │ └⊟ 67 78 signal transduction (GO:0007165)
│ │   ├⊞ 19 24 cell surface receptor linked signal transduction (GO:0007166)
│ │   ├⊞ 24 32 intracellular signaling cascade (GO:0007242)
│ │   └⊞ 2 7 regulation of signal transduction (GO:0009966)
│ ├⊞ 9 10 cell differentiation (GO:0030154)
│ ├⊟ 197 220 cellular physiological process (GO:0050875)
│ │ ├⊟ 22 25 cell cycle (GO:0007049)
│ │ │ ├⊞ 4 6 M phase (GO:0000279)
│ │ │ ├⊞ 2 3 interphase (GO:0051325)
```

COMPARISON

NAME: test ny test

DESCRIPTION:

REPORTERLISTS: DCM 005 CAD 005

STATISTICAL TEST: Intersecting reportelist test

P-VALUE CUTOFF: 0.05

Edit comparison

Biological process

View / edit tree
View results - Generate: Compressed Excel file, XML, Text
View report- Generate: Compressed Excel file, XML, Text

Group mode: None Include adjusted p-values: No

By clicking the text or numbers, additional information on the identifiers will pop up

GO	Name	DCM 005	CAD 005	P-value	A Ç B	A Ç B'	A' Ç B
GO:0008150	biological_process	266	242		98	168	144
GO:0009968	negative regulation of signal transduction	4	0	0.044	0	4	0
GO:0000902	cellular morphogenesis	12	4	0.028	2	10	2
GO:0044248	cellular catabolism	4	12	0.013	2	2	10
GO:0044265	cellular macromolecule catabolism	3	12	0.004	2	1	10
GO:0044257	cellular protein catabolism	2	8	0.011	2	0	6
GO:0051603	proteolysis during cellular protein catabolism	2	8	0.011	2	0	6
GO:0019941	modification-dependent protein catabolism	2	7	0.020	2	0	5
GO:0006511	ubiquitin-dependent protein catabolism	2	7	0.020	2	0	5
GO:0043632	modification-dependent macromolecule catabolism	2	7	0.020	2	0	5
GO:0006508	proteolysis	10	19	0.005	8	2	11
GO:0031324	negative regulation of cellular metabolism	9	4	0.040	4	5	0
GO:0009056	catabolism	6	16	0.009	2	4	14
GO:0009057	macromolecule catabolism	4	12	0.013	2	2	10
GO:0043285	biopolymer catabolism	4	10	0.037	2	2	8
GO:0030163	protein catabolism	3	8	0.041	2	1	6
GO:0009892	negative regulation of metabolism	10	4	0.022	4	6	0

Figure 4
Result report output from eGOn. Gene reporter lists submitted to eGOn can be visualized in tree-view, as result-view or as report-view. In the tree-view (A) the nodes may be collapsed or expanded producing the desired level of detail and the resulting structure can be saved as a template for future use. Several preset levels can also be selected. By clicking on a GO node the gene reporter associated with this GO node in the GO-tree can be interactively examined and links are offered to single gene view in the *NMC Annotation Tool*. In result view p-values for all GO categories are shown and for the report view (B), only the GO categories that fit the user's p-value cut-off are shown.

ability of belonging to GO category G is different for genes on gene list A and genes on gene list B. For each gene on list A, there is a conditional probability P(G|A) of belonging to GO category G, and for each gene on list B, there is a conditional probability P(G|B) of belonging to GO category G. Under the null hypothesis these two probabilities are equal. From this the following null hypothesis and alternative hypothesis can be formulated.

H_0: P(G|A) = P(G|B) vs. H_1: P(G|A) ≠ P(G|B)

By using the laws of conditional probability, we have the following additional interpretation. For a chosen GO category G, the ratio between the probability of membership of gene reporter list A and membership of gene reporter list B, is the same as the ratio between the probability of being a member of gene reporter list A to the probability of being a member of gene reporter list B in the whole GO-tree. Statistically we need to distinguish between three situations, to correctly handle the possible dependencies between gene reporter lists A and B. An illustration of these situations is given in figure 5. Different statistical hypothesis tests are suitable for the three situations. In *e*GOn we have implemented three tests for these situations: the master-target test, the mutually exclusive target-target test and the intersecting target-target test. In brief, all three tests are parametric and the tests for the master-target situation and the mutually exclusive target-target situation are based on the same implementation of Fisher's exact test, but with different inputs. The intersecting target-target test is based on a test statistic by Leisenring et al. [29]. The test of Leisenring is designed to test if the positive predictive value (PPV) of two medical diagnostic tests is equal. A further description of the different situations and the corresponding tests can be found in the next chapters. Moreover, a detailed description of the statistical tests is offered in the supplementary material (additional file 1).

Master-target situation

In the master-target situation the GO categories (e.g. biological processes) of the genes of interest (e.g. differentially expressed) from a given experiment (target list) are compared with the distribution of GO categories for all gene reporters represented as physical probes on the microarray (master list) used in the experiment. The purpose is to find whether, in any of the GO categories, the genes of interest are over- or underrepresented compared to the genes represented on the microarray. For our heart failure example, list M would be a list of all the genes investigated on the microarray and list B would be the genes that are found to be differentially expressed between the DCM hearts and the controls (Figure 5).

This type of comparison between two gene reporter lists is useful and most GO tools offer tests for this. Statistically this situation can be transformed into a problem where we for each GO category under consideration want to test if two independent binomial proportions are equal (for details, see Günter et al. [30]). Several statistical approaches can be used, e.g. Fisher's exact test, Pearson's asymptotic Chi-square-test, a conditional mid-p test, or an unconditional test. We refer to Agresti [31] for a presentation of these tests, and to Khatri and Dragici [8] for an overview of different statistical tests implemented in the various GO-tools available in the master-target situation. In *e*GOn we have chosen the Fisher's exact test for the master-target situation and we call this the master-target test. The implementation is based on a translation to PHP of a JAVA-script by Langsrud [32]. The use of this two sided test is further explained by Zeeberg et al. [33].

Mutually exclusive target-target situation

In the mutually exclusive target-target situation there are no common genes in the two lists compared, in the heart failure example list A1 could be the list of differentially expressed genes that are up-regulated for the CAD hearts compared to the controls, while list A2 contains the genes that are down-regulated for the CAD hearts compared to the controls. The purpose with this type of comparison is to find which e.g. biological processes as defined by GO categories are differentially represented in the up- and down-regulated genes in the same experiment (Figure 5).

Statistically this situation is very similar to the master-target situation and can be transformed into a problem where we for each GO category under consideration want to test if two independent binomial proportions are equal. The same statistical tests as listed for the master-target test can be used. In *e*GOn we have chosen to implement the Fisher's exact test for the mutually exclusive target-target situation, called the mutually exclusive target-target test, using the same implementation, but with different inputs, as for the master-target test.

Intersecting target-target situation

When two gene reporter lists are compared and a number of gene reporters are represented on both lists, the intersecting target-target test is used to investigate whether the GO categories represented by these genes are over- or under represented in the experiments behind the two lists. In our heart failure example, list A could be the differentially expressed genes between CAD hearts and controls while list B would be the differentially expressed genes between DCM hearts and controls (Figure 5).

In Günther et al. [30], three different statistical tests are presented in the situation where the two gene lists are intersecting. All three tests are constructed for use with

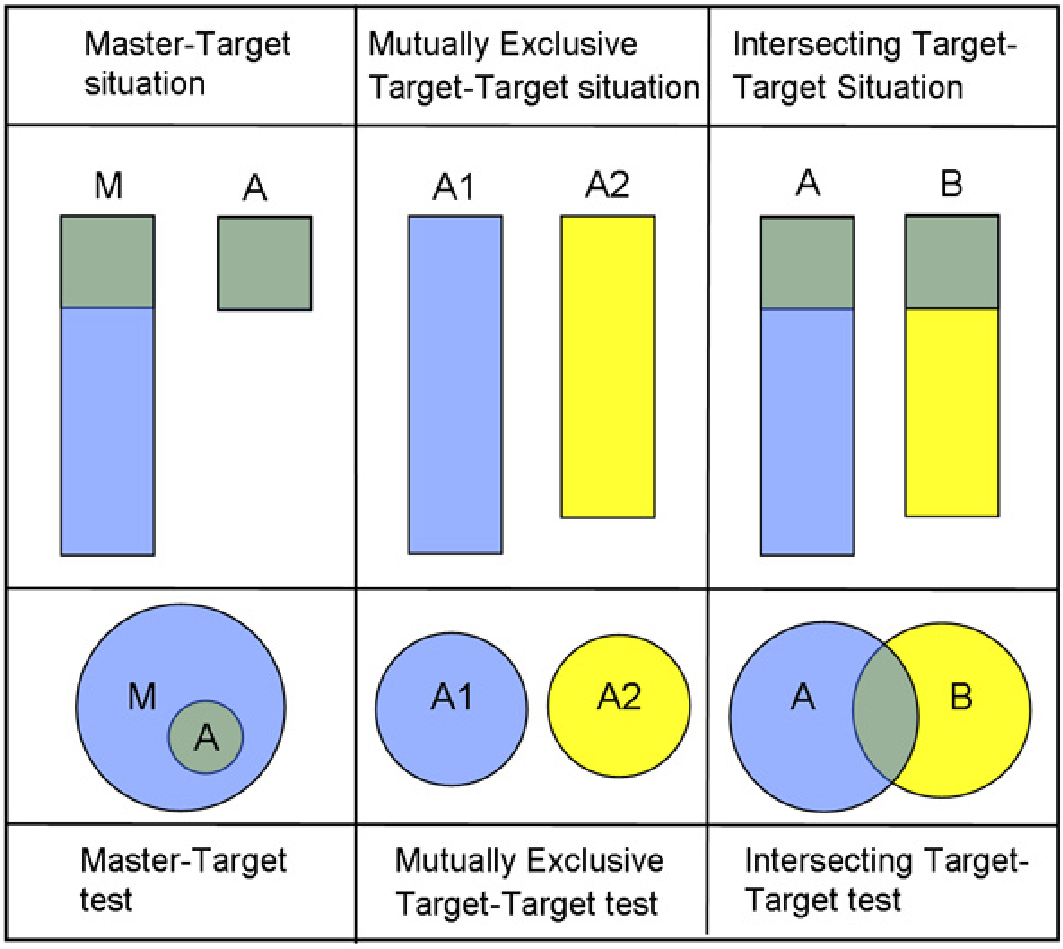

Master-Target situation
Mutually Exclusive Target-Target situation
Intersecting Target-Target Situation
M
A
A1
A2
A
B
M
A
A1
A2
A
B
Master-Target test
Mutually Exclusive Target-Target test
Intersecting Target-Target test

Figure 5

Three different situations covered by the statistical testes in e GOn. *Master-target situation*: When one gene reporter list is a subset of the other list (the master list) the master-target test can be used in the comparison. *Mutually exclusive target-target situation*: If the gene reporters do not have any reporters in common (e.g. lists of up- vs. down regulated genes form the same experiment) the mutually exclusive target-target test can be used. *Intersecting target-target situation*: if the two lists compared include common gene reporters, from e.g. two experiments, then the intersecting target-target test can be used.

large samples, and are based on an asymptotic relation to the Chi-square distribution. In *e*GOn we have chosen to implement the test based on Leisenring et al. [29], originally constructed for comparing positive predictive values of two diagnostic tests, tests A and B, with respect to a disease G. This test uses a score statistic based on generalized estimating equations to fit a generalized linear model. We have translated this test into the setting of comparing two gene lists at a given GO category. Further details can be found in Günther et al. [30] or in the supplementary material (additional file 1).

Methodical considerations

The statistical tests for association between two gene reporter lists under consideration are based only on the gene lists submitted to *e*GOn, and the raw data underlying the statistical analyses producing the gene reporter lists are not submitted to *e*GOn. This means that *e*GOn does not offer permutation based methods for addressing the dependence structure between the genes. The statistical tests in *e*GOn are thus based on the assumption that under the null hypothesis the genes on the lists (or subsets of the lists in the intersecting target-target situation) act independently, as is also commonly assumed in other GO-tools. This should be taken into consideration when analysis is performed, and duplicate genes/reporters, close family members or pathways partners may be removed. This can easily be done by the filtering tool in *GeneTools*.

The p-values produced by the statistical test can be displayed for all GO categories or only those satisfying a certain p-value cut-off. Adjusted p-values can be calculated for a selected set of GO categories and is dependent on how the GO hierarchy is collapsed/expanded, using the step-up procedure of Benjamini and Hochberg [34] for controlling the False Discovery Rate (FDR). Setting a cut-off at 0.05 for the adjusted p-value will control the (FDR) at level 0.05. The Benjamini-Hochberg step-up procedure controls the FDR under certain dependence structures (for example positive regression dependency, see Benjamini and Yekutieli [35] for a detailed presentation). However, the dependency structure among the selected GO-categories in the GO-tree is not known, and questions remain about controlling the FDR in hierarchical structures.

One important "consensus point" within statistical inference discussed by Allison et al. [36] is that gene set testing is desirable, and has become a popular and widely accepted analytical tool. However, one problem with gene class testing, according to Allison et al. [36], is that the null hypotheses of these tests are not, or poorly defined. By formally stating the null and alternative hypotheses, we think our paper has addressed these concerns in a thorough manner. An important consideration when searching for statistically significant GO categories within a gene reporter list (our master-target test) is the choice of the reference (master) list of gene reporters from which the p-values for each GO category in the results are calculated. Some tools use the total set of genes in a genome as a reference (the master list). We do not think this is the best solution since the observed number of gene reporters for a specific GO category should be compared with the number of gene reporters that could appear if a random selection was taken from the list of all genes that was under study in the experiment.

In *e*GOn p-values can be shown for the whole GO tree and unlike most other tools several preset levels can be chosen and users can modify the tree as they like. In addition a result report view is accessible, showing only the GO nodes which satisfy a specific pre set p-value cutoff. Unique in the *e*GOn tool, we offer statistical tests for comparisons between gene reporter lists. The master-target test and mutually exclusive target-target test are both used in different variations in several programs today, but no other GO-tool, to our knowledge, offers tests for the intersecting target-target situation. However, the statistical test of FatiGO [37] is valid for the mutually exclusive target-target situation, and was in a simulation study found to preserve the test size when the gene reporter lists are of equal length [30]. Our intersecting target-target test is valid when the two gene reporter lists are intersecting, potentially constituting a useful test, since it offers the opportunity to compare gene reporter list for different experiments (as previously described by the heart diseases example). In this way both our target-target tests may answer questions not necessary answered by the standard master-target tests applied to most tools.

Future plans

GeneTools was released in September 2005 and has steadily gained popularity since then. In October 2006 over 1 700 users from 60 countries were registered and over 4 000 gene reporter lists were submitted to the database. We plan to continue adding new features to *GeneTools*, including more information from external databases like e.g. Ensembl and OMIM. Furthermore, we hope to provide developers of other tools an extended version of our API and extend the export function to support SBML (systems biology markup language) [38] which will make more high-level analysis possible. We think the need for central and publicly available resources which curate biological data will only continue to grow and that *GeneTools* and similar tools will be essential for biologists and bioinformaticians to efficiently analyze genome-scale datasets. Today their main utility is for gene expression analysis, but in the future proteomic and SNP data need to be analyzed by similar tools. In addition, an important future use of annotation tools will be in systems biology approaches that are now evolving rapidly.

Conclusion

GeneTools is a flexible and user friendly "all in one" annotation tool, where the users can rapidly extract gene annotation data for e.g. thousands of genes or clones at once. The user can add "user defined" GO annotation to gene products and all annotation information is stored in a database which can easily be shared with other users and exported in different formats. *e*GOn is the first tool that can perform hypothesis testing for three different situations, looking for over- or under-representation of GO categories between gene reporter lists.

Availability and requirements

Project name: *GeneTools*

Project Homepage: http://www.genetools.no

Operating System: Platform independent

Programming Language: PHP

Underlying Database: mySQL

Authors' contributions

VB initiated and coordinated the project and wrote the manuscript. AL co-initiated the project and together with AKS they supervised the project and were involved in drafting and reviewing the manuscript. ML, CCG and SL devised the statistical algorithms. FKJ, HB and HR designed and built the underlying database. LJ, HB and FKJ contributed equally in writing the program code and maintain the underlying database. FKJ, HB, and VB designed the *GeneTools* web interface. All authors read and contributed to revising the manuscript for intellectual content and approved the final manuscript.

Additional material

Additional file 1

eGOnv2_statistics.pdf. As supplementary materials a detailed description of the background for the statistical tests in eGOn is offered.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-7-470-S1.pdf]

Acknowledgements

We wish to thank Martin Kuiper for reviewing the manuscript and øyvind Møll for his great help with debugging and implementing the final features of the software. In addition, we wish to thank Vladimir Yankovski and Jan Komorowski for their contributions to early stages of *GeneTools*.

VB is the recipient of a research fellowship from the National Council on Cardiovascular Diseases. CCG is funded by a NTNU grant.

GeneTools was developed with support from the National Microarray Technology platform (Norwegian Microarray Consortium, NMC), which is supported by the Functional genomics program (FUGE) of the Norwegian Research Council.

References

1. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93:**10614-10619.
2. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
3. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.** *Bioinformatics* 1998, **14:**656-664.
4. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA: **NetAffx: Affymetrix probesets and annotations.** *Nucleic Acids Res* 2003, **31:**82-86.
5. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
6. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.
8. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
9. **UniGene website** [http://www.ncbi.nlm.nih.gov/./UniGene/]
10. **Entrez website** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene]
11. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32:**D262-D266.
12. Blake JA, Richardson JE, Davisson MT, Eppig JT: **The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. The Mouse Genome Informatics Group.** *Nucleic Acids Res* 1997, **25:**85-91.
13. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ: **Rat Genome Database (RGD): mapping disease onto the genome.** *Nucleic Acids Res* 2002, **30:**125-128.
14. **GeneOntology website** [http://www.geneontology.org/]
15. **SwissProt website** [http://ca.expasy.org/sprot/]
16. **HomoloGene website** [http://www.ncbi.nlm.nih.gov/HomoloGene/]
17. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services.** *Nucleic Acids Res* 2006, **34:**W729-W732.
18. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3:**SOFTWARE0003.
19. Alsberg B, Kirkhus L, Tangstad T, Andressen E: **Data analysis of microarrays using SciCraft.** *Knowledge exploration in life science informatics, proceedings lecture notes in artificial intelligence* 2004, **3303:**58-68. Ref Type: Journal (Full)
20. Liu H, Hu ZZ, Torii M, Wu C, Friedman C: **Quantitative assessment of dictionary-based protein named entity tagging.** *J Am Med Inform Assoc* 2006, **13:**497-507.
21. Liefeld T, Reich M, Gould J, Zhang P, Tamayo P, Mesirov JP: **GeneCruiser: a web service for the annotation of microarray data.** *Bioinformatics* 2005, **21:**3681-3682.
22. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA: **Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate.** *Nucleic Acids Res* 2003, **31:**3775-3781.
23. Martinez-Cruz LA, Rubio A, Martinez-Chantar ML, Labarga A, Barrio I, Podhorski A, Segura V, Sevilla Campo JL, Avila MA, Mato JM: **GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data.** *Bioinformatics* 2003, **19:**2158-2160.
24. Lenhard B, Wahlestedt C, Wasserman WW: **GeneLynx mouse: integrated portal to the mouse genome.** *Genome Res* 2003, **13:**1501-1504.
25. **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34:**D322-D326.
26. **Gene Onology: Current annotations website** [http://www.geneontology.org/GO.current.annotations.shtml]
27. **Gene Ontology: Tools website** [http://www.geneontology.org/GO.tools.shtml]
28. Beisvag V, Lehre PK, Midelfart H, Aass H, Geiran O, Sandvik AK, Laegreid A, Komorowski J, Ellingsen O: **Aetiology-specific patterns in end-stage heart failure patients identified by functional**

annotation and classification of microarray data. *Eur J Heart Fail* 2006, **8:**381-389.

29. Leisenring W, Alonzo T, Sullivan S: **Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs.** *Biometrics* 2000, **56:**345-351.
30. Günther CC, Langaas M, Lydersen S: **Statistical Hyhpothesis Testing of Association Between Two Lists of Genes for a Given Gene Class. 4-1-2006. Preprint Statistics 1/2006.** [http://www.math.ntnu.no/preprint/statistics/2006/]. Department of Mathematical Sciences, The Norwegian University of Science and Technology Ref Type: Report
31. Agresti A: *Categorical Data Analysis* 2nd edition. John Wiley & Sons, New York; 2002.
32. Langsrud Ø: **Fisher's exact test.** 2004 [http://www.matforsk.no/ola/fisher.htm]. Ref Type: Computer Program
33. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4:**R28.
34. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **57:**289-300.
35. Yekutieli D, Benjamini Y: **The Control of the FDR multiple testing under dependency.** *The Annuals of Statistics* 2001, **29:**1165-1188.
36. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7:**55-65.
37. Al-Shahrour F, az-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
38. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, *et al.*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19:**524-531.

Paper II

ELSEVIER

Journal of Reproductive Immunology 74 (2007) 7–14

JOURNAL OF
REPRODUCTIVE
IMMUNOLOGY

www.elsevier.com/locate/jreprimm

Fetal growth restriction is associated with reduced FasL expression by decidual cells

Irina P. Eide [a,d,*], Christina V. Isaksen [b,e], Kjell Å. Salvesen [b,d], Mette Langaas [c], Clara-Cecilie Günther [c], Ann-Charlotte Iversen [a], Rigmor Austgulen [a]

[a] *Department of Cancer Research and Molecular Medicine, DMF, Norwegian University of Science and Technology (NTNU), Medisinsk teknisk forskningssenter, Olav Kyrres gt. 3, N-7489 Trondheim, Norway*
[b] *Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway*
[c] *Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway*
[d] *Department of Obstetrics and Gynecology, St. Olavs Hospital, Trondheim, Norway*
[e] *Department of Pathology, St. Olavs Hospital, Trondheim, Norway*

Received 1 May 2006; received in revised form 27 November 2006; accepted 28 November 2006

Abstract

The Fas-Fas ligand (FasL) system contributes to immune tolerance at the feto-maternal site and has been ascribed a role in implantation and placental development by regulating trophoblast invasion and spiral artery remodelling. In the present study, we have examined FasL expression in decidual tissue from pregnancies with impaired placental development. Women with pre-eclampsia (PE) and/or fetal growth restriction (FGR) were enrolled as cases ($n = 33$), and women with normal pregnancies were used as controls ($n = 27$). Decidua basalis tissue was obtained by vacuum suction of the placental bed after delivery. FasL expression by extravillous trophoblasts (EVTs) and decidual cells (DeCs), together with EVT apoptosis, were assessed by immunohistochemistry. Levels of soluble FasL in maternal serum and apoptosis-related gene expression in decidual tissue were determined.

The proportion of FasL-expressing DeCs was high in controls (72.0 ± 10.2%), with a significant reduction among cases (58.1 ± 19.7%; $p = 0.002$), especially in those with FGR (54.3 ± 19.9%; $p < 0.001$). EVTs had a lower proportion of FasL expression than DeCs, with a less pronounced reduction in cases compared to controls (10.9 ± 3.9 and 8.3 ± 4.0%, respectively; $p = 0.02$). Decidual FasL expression correlated with placental growth. The EVT apoptosis rate did not differ between cases and controls (1.1 ± 1.9 and 1.1 ± 1.3%, respectively).

These findings indicate a reduction of immune privilege in decidua of PE/FGR pregnancies by reduced FasL expression and that DeCs may have a central role in the Fas-FasL-based feto-maternal immune balance.

Keywords: FasL; Decidual cells; Extravillous trophoblasts; Fetal growth restriction; Pre-eclampsia

1. Introduction

Fas ligand (FasL) is a member of the tumour necrosis (TNF) superfamily (Bohana-Kashtan and Civin, 2004). Its expression is confined to activated T lymphocytes, natural killer (NK) cells and cells in immune privileged sites (Medvedev et al., 1997; Bohana-Kashtan

* Corresponding author at: Department of Cancer Research and Molecular Medicine, DMF, Norwegian University of Science and Technology (NTNU), Medisinsk teknisk forskningssenter, Olav Kyrres gt. 3, N-7489 Trondheim, Norway. Tel.: +47 73550290; fax: +47 73598801.
E-mail address: irina.p.eide@ntnu.no (I.P. Eide).

0165-0378/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.jri.2006.11.002

and Civin, 2004). In contrast, its receptor Fas (belonging to the TNF receptor superfamily) is expressed by most cell types (Bohana-Kashtan and Civin, 2004). The interaction between Fas-FasL induces apoptosis of the Fas-expressing cell (Bohana-Kashtan and Civin, 2004). The Fas-FasL system has been ascribed a role in immune tolerance at the feto-maternal interface (Straszewski-Chavez et al., 2005), and has been considered a potential pathway to prevent excessive infiltration of leukocytes into the decidua (Hunt et al., 1997; Mor et al., 1998; Kauma et al., 1999; Qiu et al., 2005). Both extravillous trophoblasts (EVTs) (Runic et al., 1996; Uckan et al., 1997; Hammer and Dohr, 2000) and decidual cells (DeCs) (Hunt et al., 1997; Makrigiannakis et al., 2003; Kayisli et al., 2003) express FasL. Apoptotic leukocytes have been detected in close proximity to FasL-expressing EVTs in in vivo materials (Mor et al., 1998), and trophoblast-induced FasL-mediated apoptosis of activated lymphocytes has been confirmed in vitro (Coumans et al., 1999). In a recent study, an inverse relationship was reported between FasL-expressing DeCs and maternal leukocytes (Qiu et al., 2005). Based on these data, Qiu et al. (2005) hypothesized that FasL-expressing DeCs may regulate infiltration of maternal Fas-positive leukocytes at the feto-maternal interface.

Superficial trophoblast invasion and insufficient spiral artery remodelling characterize pregnancies complicated by pre-eclampsia (PE) and/or fetal growth restriction (FGR) (Redman and Sargent, 2005). The Fas-FasL system is involved in these processes (Murakoshi et al., 2003; Ashton et al., 2005; Harris et al., 2005) and, in accordance with this, abnormal Fas-FasL expression has been reported in PE/FGR. FasL exists also in a soluble (s) form. The physiological functions of sFasL are not completely understood, but it is assumed to play a role in the regulating functions of the immune system. In women with PE/FGR, reduced leukocyte expression of FasL has been reported (Kuntz et al., 2001), whereas observations regarding sFasL are conflicting; some report elevated levels in PE/FGR (Kuntz et al., 2001; Hu et al., 2005) and others find sFasL concentrations corresponding to those in normal pregnant women (Laskowska et al., 2006). Similarly, observations at the local site are diverging—reduced (Allaire et al., 2000; Yue et al., 2005), unchanged (Hu et al., 2005) and increased (Koenig and Chegini, 2000) FasL expression have been reported in PE/FGR pregnancies.

The closest interaction between fetal and maternal cells occurs in decidua and thus this tissue will probably provide the most sensitive detection of changes in FasL expression, disturbing immune privilege and physiological implantation processes. Thus, in this study, we have used decidual tissue to examine FasL expression and apoptosis in pregnancies with PE and/or FGR. Concomitantly, maternal sFasL levels were assessed to learn whether disease is associated with changes.

2. Materials and methods

2.1. Study groups

Since we aimed to study FasL expression in impaired placental development, cases with suspected placental insufficiency (with PE and/or FGR) were recruited. PE was defined as persistent hypertension (blood pressure of ≥140/90 mmHg) plus proteinuria (≥0.3 g/24 h or ≥2+ according to a dipstick test), developing after 20 weeks of pregnancy (Gifford et al., 2000). FGR implied birth weight ≤2 standard deviations (S.D.) below the expected birth weight as related to gestational age (GA) and sex (Marsal et al., 1996). Due to tissue sampling procedures, only cases delivered by caesarean section (CS) could be included. Healthy women with normal pregnancies undergoing CS for various reasons considered irrelevant to the aim of this study (e.g. breech presentation and previous CS) served as controls.

Materials were collected at St. Olavs Hospital (the University Hospital in Trondheim). The study was approved by the Regional Committee for Medical Research Ethics. Informed consent was obtained from all participants.

2.2. Sample collection and preparation

Decidual tissue was obtained by vacuum suction of the placental bed after the placenta was delivered (Staff et al., 1999; Harsem et al., 2004). Some tissue was snap-frozen in liquid nitrogen and stored at −80 °C, some was fixed in 10% neutral buffered formalin and paraffin-embedded. Tissue for gene expression was immediately submerged in a RNA stabilization solution (RNA*later*, Ambion, Huntington, UK), incubated at 4 °C overnight and stored at −80 °C. Peripheral maternal blood was collected, centrifuged and serum was stored at −80 °C.

2.3. Placental growth and histology

Placental weight and placental weight ratio (PWR: observed placental weight/expected placental weight) were registered. Expected placental weight according to GA was obtained from published standards (Benirschke and Kaufmann, 2000). All placentas were examined by one pathologist in accordance with established clinical routines.

2.4. Expression of FasL by DeCs and EVTs

The expression of FasL in decidual tissue was assessed by a double immunofluorescence technique using an antibody against FasL in combination with antibodies to detect decidual cells (anti-prolactin) or trophoblasts (anti-cytokeratin 7). A polyclonal rabbit antibody (pAb) against FasL (Q-20; Santa Cruz Biotechnology, Santa Cruz, CA; diluted 1:20), with highly ranked specificity (Hammer and Dohr, 2000), was used for FasL detection. Tissue was cut in three serial sections at 5 μm on a freeze microtome. One section was stained with haematoxylin and eosin (HE) and used for quality control. Only specimens containing both DeCs and EVTs, with a low contamination of blood and/or placental tissue, were included for further studies. The two remaining cryosections were fixed in acetone (10 min at 4 °C). Bovine serum albumin (2%) was added for 10 min to inhibit non-specific binding and, subsequently, pAb against FasL (Q-20) was added for 1 h either in combination with a mouse monoclonal antibody (mAb) against cytokeratin 7 (CK7, clone OV-TL 12/30; DakoCytomation, Glostrup, Denmark; diluted 1:300) or mouse mAb against human prolactin (clone PRL02; Neo Markers, Fremont, CA, USA; diluted 1:2). Sections were incubated with secondary antibodies (TRITC-conjugated swine anti-rabbit pAb (code R0156; DakoCytomation; diluted 1:30) and FITC-conjugated goat anti-mouse pAb (code F0479; DakoCytomation; diluted 1:10) for 30 min. All incubations were performed at room temperature in a dark moist chamber, and phosphate-buffered saline (PBS) was used for washing between all incubation periods. Sections were counterstained with 1000 ng/mL 4,6-diamidino-2-phenylindole (DAPI 1, Abbott Laboratories AS, Gentofte, Denmark) and examined with a fluorescent microscope (Nicon Eclipse E600) at ×600 magnification using CytoVision 3.6 software (Applied Imaging, Newcastle upon Tyne, UK). Placental tissue (with FasL-expressing villous trophoblasts, Hofbauer cells and fetal blood vessels) (Hammer and Dohr, 2000) were used as positive controls, whereas sections incubated with serum from non-immunized rabbits (Code X0936; DakoCytomation; diluted 1:20) served as negative controls in the FasL experiments. Glands in decidua were used as internal positive control for mAb against CK7, and sections of pituitary glands were included as a positive control for anti-prolactin (not shown).

The ratio of FasL-expressing DeCs and EVTs was calculated as the percentage of FasL-positive DeCs and EVTs in each section (number of FasL-positive prolactin or CK7-positive cells among 100 of the corresponding cell type).

2.5. Soluble FasL in maternal blood

Levels of sFasL in maternal serum were assessed by a sandwich enzyme-linked immunoassay (ELISA, R&D Systems, Abingdon, UK), according to the manufacturer's standard procedure. Samples were tested in duplicates (50 μL each). The detection level of the assay was 2.7 pg/mL.

2.6. Apoptosis-related gene expression in decidua

The expression of apoptosis-relevant genes in decidual tissue was studied by Affymetrix GeneChip analysis. Total RNA was purified from decidual tissue using a RNeasy Midi Kit according to the manufacturer's instructions (Qiagen, Crawley, UK), and RNA concentration and purity were measured using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Rockland, DE, USA). Double stranded cDNA was synthesized using 5 μg total RNA. Synthesis of cDNA (cDNA synthesis kit; Invitrogen, Carlsbad, CA, USA) and biotin-labeled cRNA (BioArray^{TM} HighYield^{TM} kit; Enzo Life Sciences, NY, USA), cRNA fragmentation, target hybridization, washing, staining and scanning were performed using standard Affymetrix protocols (Affymetrix, Santa Clara, USA). Fragmented biotinylated cRNA targets were hybridized to Human Genome Focus Arrays (Affymetrix) using 10 μg fragmented biotinylated cRNA. The arrays were scanned using a Agilent GeneArray Scanner controlled by MAS 5.0. Gene expression data were analyzed with a Gene Map Annotator and Pathway Profiler (GenMAPP, Gladstone Institute, CA, www.GenMAPP.org).

2.7. Trophoblast apoptosis

Apoptosis of EVTs was assessed in two serial sections of formalin-fixed decidual tissues by counting mAb M30 (clone M30; Roche, Mannheim, Germany; diluted 1:50) (Kadyrov et al., 2001) positive cells in one section and relating numbers to mAb CK7 (clone OV-TL 12/30; Dako Cytomation; diluted 1:1000) positive cells in section two. The staining was performed in an automated slide stainer (DakoCytomation) after deparaffination, rehydration and heat-induced antigen retrieval. Endogenous peroxidase activity was blocked, and primary antibodies added. Incubation time was 30 and 40 min for anti-CK7 and M30, respectively. Secondly, a peroxidase-conjugated polymer with antibodies against rabbit/mouse (ChemMate Dako Envision Detection Kit Peroxidase/DAB) (DakoCytomation) was used according to the manufacturer's standard procedure.

Antibody-diluent without antibody was used as a control for non-specific staining, and sections from colon adenocarcinoma were used as positive controls for M30. The sections were counterstained in haematoxylin for 1 min and analyzed at ×200 magnification in a light microscope.

The trophoblast apoptosis ratio in decidual tissues was calculated as M30-positive cells among CK7-positive cells. Cell counting was performed in areas that: (1) contained at least 100 trophoblasts, and (2) were easily reproduced from one section to the other. Areas that fulfilled these criteria were demarcated with Indian ink in paired serial sections. One case with 35 EVTs in the whole section was excluded from further apoptosis analyses due to statistical considerations.

Cell counting was performed by two independent individuals, blinded for the case/control status at ×200 magnification, using a 13 mm × 13 mm reticule. The mean count was used for statistical analysis.

2.8. Statistical analysis

The clinical data and the results of cell counting were expressed as the mean ± the corresponding S.D. The Mann–Whitney test was used for comparison between groups. Spearman's ranked correlation test was used to examine correlation between FasL expression, EVT apoptosis, placental weight and clinical data. $p < 0.05$ was considered significant. SPSS Version 12 was used for all statistical analyses.

Statistical analysis of the Affymetrix GeneChip arrays data was based on summary expression measures for each gene probe set. These summary measures were computed from quantile-normalized and log-transformed perfect match values for each probe pair using the robust multiarray average method (Irizarry et al., 2003). Tests for significant differential expression for cases versus controls were performed using moderated *T*-tests (Smyth, 2004). To account for multiple testing, we calculated adjusted *p*-values controlling the false discovery rate and inserted the estimated value of the proportion of non-differentially expressed genes (Langaas et al., 2005). All analyses were performed using the R statistical package and the Affy & Limma sofware packages from the Bioconductor project (Gentleman et al., 2004).

3. Results

Sixty women were included in the study (33 cases and 27 controls). Cases and controls differed significantly with respect to GA, birth weight and placental weight (Table 1).

Table 1
Clinical information

	Cases ($n = 33$)	Controls ($n = 27$)
Maternal age (years)	30 ± 5	31 ± 5
Gestational age (weeks) at delivery	32 ± 4*	39 ± 1
Systolic blood pressure (mmHg)	148 ± 20*	121 ± 14
Diastolic blood pressure (mmHg)	93 ± 12*	71 ± 13
Birth weight (g)	1525 ± 668*	3662 ± 501
BWR[a]	0.71 ± 0.16*	1.10 ± 0.15
Placental weight (g)	323 ± 116*	644 ± 144
PWR[b]	0.96 ± 0.25*	1.45 ± 0.33

Values are expressed as mean ± S.D., unless stated otherwise. All case groups were compared to controls.

[a] BWR: birth weight ratio (observed birth weight/expected birth weight).

[b] PWR: placental weight ratio (observed placental weight/expected placental weight).

* $p < 0.001$.

3.1. Placental weight and histology

The PWR was significantly lower among cases (0.96 ± 0.25) than the control group (1.45 ± 0.33) ($p < 0.001$). Reduced PWR was observed in pregnancies with FGR ($p < 0.001$), whereas pregnancies with isolated PE had a PWR comparable with controls. Abnormal morphological findings in the placenta were more frequent among cases than controls; 13 (39.4%) cases demonstrated advanced villous maturation compared with 3 (11.1%) controls, and 4 (12.1%) cases had multiple infarcts compared with none in the control group.

3.2. Expression of FasL by DeCs and EVTs

FasL staining of DeCs demonstrated a distinct punctuate staining pattern in the cytoplasm (Fig. 1A), with a less intense staining of EVTs (Fig. 1C). The proportion of FasL-expressing DeCs was reduced among cases (58.1 ± 19.7%) compared with controls (72.0 ± 10.2%) ($p = 0.002$) (Fig. 1B). The reduction was most pronounced in pregnancies complicated with FGR (54.3 ± 19.9%) ($p = 0.001$), whereas isolated PE did not differ significantly from controls. A positive correlation was observed between the proportion of FasL expression by DeCs and PWR ($p < 0.001$).

EVTs demonstrated a much lower proportion of FasL-expressing cells than DeCs, both in cases and controls (Fig. 1D). As with DeCs, FasL expression by EVTs was reduced among cases (8.3 ± 4.0%) compared with

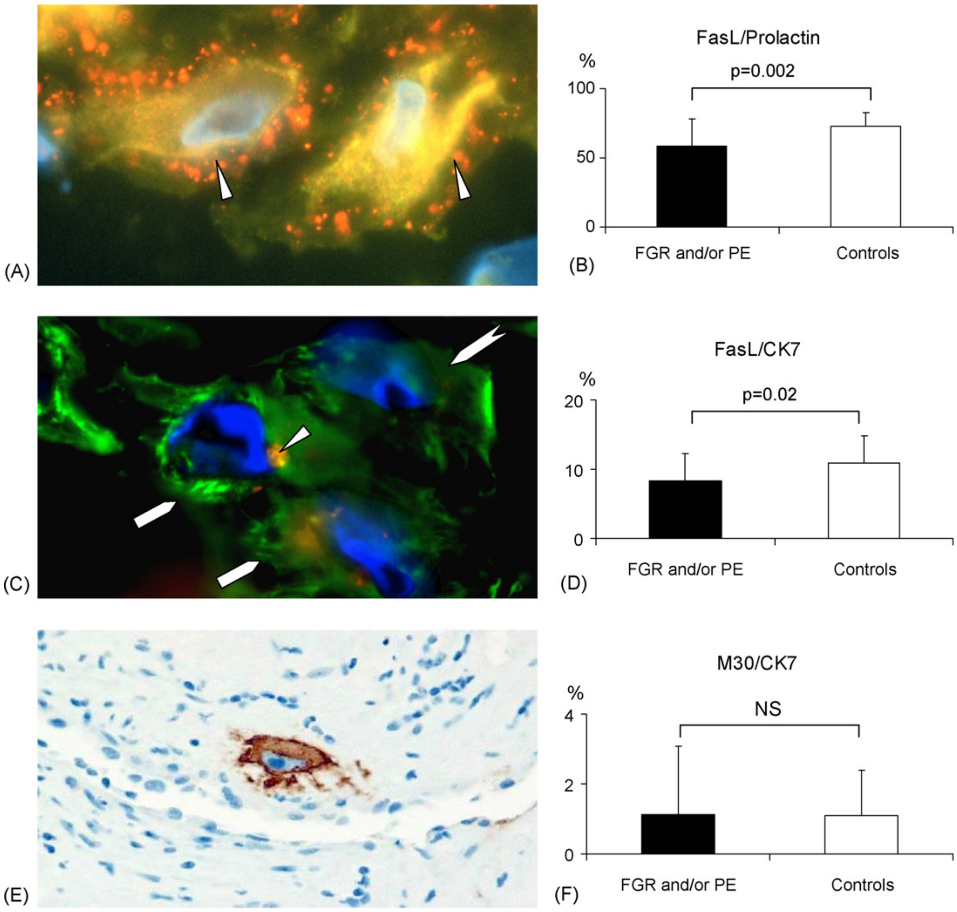


Fig. 1. Double immunofluorescence and single immunohistochemical staining on cryosections (A and C) and formalin-fixed, paraffin-embedded section (E) of decidua basalis obtained from women at delivery. The immunofluorescence staining was performed on two serial sections: (A) demonstrates FasL staining (red signal) of decidual cells (DeCs) with anti-prolactin (green signal) as DeC marker, and (C) illustrates FasL expression (red signal) in extravillous trophoblasts (EVTs) with anti-CK7 (green signal) as EVT marker. Two FasL-positive DeCs are shown in (A), whereas both FasL-positive (⇨) and FasL negative (➢) trophoblasts are presented in (B). DeCs demonstrated a distinct and punctuate FasL staining in cytoplasm (▻) (A), whereas the FasL staining of EVTs was less intense (▻) (B). Trophoblast apoptosis was detected by M30 monoclonal antibody (brown staining) (E). The proportions of FasL-positive DeCs (B), FasL-positive EVTs (D) and apoptotic EVTs (F) were calculated. Results are displayed as mean values with the corresponding S.D.

controls (10.9 ± 3.9%) ($p = 0.02$) (Fig. 1C). A positive correlation was observed between the proportion of FasL-positive EVTs and PWR ($p = 0.045$).

3.3. Soluble FasL levels in maternal serum

Levels of sFasL in maternal serum did not differ between cases ($n = 14$) and controls ($n = 12$); 77.6 ± 16.4 and 81.3 ± 13.5 pg/mL, respectively. No correlation was observed between decidual FasL expression and sFasL concentrations.

3.4. Apoptosis-related gene expression in decidua

Gene expression analyses in decidual tissue were performed on 37 of the specimens collected (18 cases and 19 controls). Due to operating costs, the number of samples subjected to gene expression analyses was restricted. The pregnancies included were randomly selected, and the gene expression analysis of cases and controls did not differ from the total case and control group with respect to GA, birth weight and blood pressure.

In total, 82 genes involved in main apoptotic pathways (both activation and inhibition, including genes involved in the Fas-FasL pathway) were studied. Expression of apoptotic pathway genes did not differ between groups. Different expression of genes known to change in PE was verified (Gallery et al., 1999), suggesting that the experimental setting had the statistical power required to detect differences between study groups.

3.5. Trophoblast apoptosis

The apoptosis rate was low both in cases and in controls, and the proportion of apoptotic EVTs (M30+/CK7+) did not differ between groups (1.1 ± 1.9 and 1.1 ± 1.3%, respectively) (Fig. 1E and F). No correlation was observed between the proportion of apoptotic EVTs and FasL expression by DeCs and by EVTs.

4. Discussion

In the present investigation we have found that FasL was expressed by both DeCs and EVTs. This observation is in accordance with previous reports (Hammer and Dohr, 1999; Kayisli et al., 2003). The physiological proportion of FasL-expressing DeCs was much higher (72%) than that of EVTs (11%). Both were reduced in association with placental disease (i.e. PE and/or FGR), but the reduction in FasL expressed by DeCs was most pronounced.

Our findings are in agreement with earlier observations indicating that decidual cells are involved in immune interactions at the feto-maternal site, with possible consequences for implantation and placental development (Olivares et al., 1997; Ruiz et al., 1997; Kitaya et al., 2000; Dimitriadis et al., 2002). Correlations between PWR and the proportion of FasL-positive DeCs ($p < 0.001$) and EVTs ($p = 0.045$) found in the present study suggest a possible association between decidual FasL expression and placental development and growth.

Trophoblast expression of FasL has been suggested as a protective mechanism against maternal leukocyte-induced apoptosis (Hammer and Dohr, 2000). Reduced, enhanced and unchanged FasL expression have been reported in PE/FGR (Koenig and Chegini, 2000; Allaire et al., 2000; Hu et al., 2005; Yue et al., 2005). All these studies were, however, performed on placental tissue/villous trophoblasts. Only one previous study has focused on changes in decidual FasL expression in association with disease (Koenig and Chegini, 2000). Koenig and Chegini (2000) found raised expression of Fas-L in cases, but their diagnosis is pregnancy-induced hypertension and the clinical information given is sparse. Accordingly, we are not sure whether the cases included in the study of Koenig and Chegini (2000) may be compared to the cases with quite severe pregnancy complications included in the present study. Since PE/FGR pregnancies are commonly delivered preterm, sampling of GA-matched healthy controls is difficult and the mean GA among cases was 6 weeks shorter than controls (Table 1). However, since FasL expression on villous trophoblasts has been reported to decrease towards term (Balkundi et al., 2000), it is likely that the observed differences in FasL expression are underestimated due to the fact that GA-matching between groups cannot be carried out.

Some investigators have reported increased EVT apoptosis in PE (Genbacev et al., 1999), whereas others found reduced apoptosis of interstitial EVTs and increased apoptosis of endovascular EVTs in PE and FGR (Kadyrov et al., 2003, 2006). No difference and generally low proportion of apoptotic EVTs (both interstitial and endovascular) has been shown in the present study (1.1% in both cases and controls). Contrasting reports on EVT apoptosis can be partly explained by use of different methods and reflect difficulties in apoptosis assessment. As decidual samples were taken from women with active disease, the presence of EVT apoptosis may have preceded the time of analyses; thus, the possibility of altered EVT apoptosis in placental insufficiency cannot be excluded.

Some investigators have reported raised maternal sFasL concentrations in PE (Kuntz et al., 2001; Hu et al., 2005), whereas we and others (Laskowska et al., 2006) found similar sFasL levels in cases and controls. Discrepancies between studies may be due to differences in diagnostic criteria and small study groups. The fact that decidual tissue from cases with placental disease demonstrated altered decidual FasL expression, whereas normal sFasL serum levels remained unchanged, makes it tempting to assume that the local immunological environment is more important for successful pregnancy than systemic regulatory events (Saito and Makino, 2005). The specific role of DeCs, compared to that of EVTs, in FasL-based immune privilege (Hunt et al., 1997; Kauma et al., 1999) remains to be elucidated, but some murine experiments are indicated. Hunt et al. (1997) found decidual necrosis and poor pregnancy outcome in mice lacking functional FasL on both fetal trophoblasts and maternal decidual cells. Thus, a normal pregnancy outcome is obtained if non-functional fetal FasL is combined with functional FasL expressed by maternal decidual cells (Rogers et al., 1998). This is supported by observations made in the present study that DeCs seem to be a fundamental contributor to appropriate FasL-based balance between mother and fetus in decidual tissues.

Acknowledgements

Toril Rolfseng and Borgny Ytterhus, Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway, performed the double immunofluorescence staining. Affymetrix GeneChip Analysis was supported by Norwegian Microarray Consortium. The study was supported by the Norwegian University of Science and Technology and St. Olavs Hospital HF.

References

Allaire, A.D., Ballenger, K.A., Wells, S.R., McMahon, M.J., Lessey, B.A., 2000. Placental apoptosis in preeclampsia. Obstet. Gynecol. 96, 271–276.

Ashton, S.V., Whitley, G.S., Dash, P.R., Wareing, M., Crocker, I.P., Baker, P.N., Cartwright, J.E., 2005. Uterine spiral artery remodeling involves endothelial apoptosis induced by extravillous trophoblasts through Fas/FasL interactions. Arterioscler. Thromb. Vasc. Biol. 25, 102–108.

Balkundi, D.R., Hanna, N., Hileb, M., Dougherty, J., Sharma, S., 2000. Labor-associated changes in Fas ligand expression and function in human placenta. Pediatr. Res. 47, 301–308.

Benirschke, K., Kaufmann, P., 2000. Pathology of the Human Placenta. Springer, New York, p. 921.

Bohana-Kashtan, O., Civin, C.I., 2004. Fas ligand as a tool for immunosuppression and generation of immune tolerance. Stem Cells 22, 908–924.

Coumans, B., Thellin, O., Zorzi, W., Melot, F., Bougoussa, M., Melen, L., Zorzi, D., Hennen, G., Igout, A., Heinen, E., 1999. Lymphoid cell apoptosis induced by trophoblastic cells: a model of active foeto-placental tolerance. J. Immunol. Methods 224, 185–196.

Dimitriadis, E., Robb, L., Salamonsen, L.A., 2002. Interleukin-11 advances progesterone-induced decidualization of human endometrial stromal cells. Mol. Hum. Reprod. 8, 636–643.

Gallery, E.D., Campbell, S., Arkell, J., Nguyen, M., Jackson, C.J., 1999. Preeclamptic decidual microvascular endothelial cells express lower levels of matrix metalloproteinase-1 than normals. Microvasc. Res. 57, 340–346.

Genbacev, O., Difederico, E., Mcmaster, M., Fisher, S.J., 1999. Invasive cytotrophoblast apoptosis in pre-eclampsia. Hum. Reprod. 14 (Suppl. 2), 59–66.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Gifford, R.W., August, P.A., Cunningham, G., Green, L.A., Lindheimer, M.D., Mcnellis, D., Roberts, J.M., Sibai, B.M., Taler, S.J., 2000. Report of the National High Blood Pressure Education Program Working Group on high blood pressure in pregnancy. Am. J. Obstet. Gynecol. 183, S1–S22.

Hammer, A., Dohr, G., 1999. Apoptotic nuclei within the uterine decidua of first trimester pregnancy arise from CD45 positive leukocytes. Am. J. Reprod. Immunol. 42, 88–94.

Hammer, A., Dohr, G., 2000. Expression of Fas-ligand in first trimester and term human placental villi. J. Reprod. Immunol. 46, 83–90.

Harris, L.K., Baker, P.N., Keogh, R.J., Cartwright, J.E., Whitley, G.S., Aplin, J.D., 2005. Trophoblast-induced smooth muscle cell apoptosis during spiral artery remodelling involves Fas/Fas ligand interactions. Placenta 26, A77.

Harsem, N.K., Staff, A.C., He, L., Roald, B., 2004. The decidual suction method: a new way of collecting decidual tissue for functional and morphological studies. Acta Obstet. Gynecol. Scand. 83, 724–730.

Hu, W.S., Wang, Z.P., Dong, M.Y., Wang, H.Z., 2005. Expression of Fas and FasL in serum and placenta of preeclamptic pregnancy and its significance. Zhejiang Da Xue Xue Bao Yi Xue Ban 34, 499–502.

Hunt, J.S., Vassmer, D., Ferguson, T.A., Miller, L., 1997. Fas ligand is positioned in mouse uterus and placenta to prevent trafficking of activated leukocytes between the mother and the conceptus. J. Immunol. 158, 4122–4128.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P., 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31, e15.

Kadyrov, M., Kaufmann, P., Huppertz, B., 2001. Expression of a cytokeratin 18 neo-epitope is a specific marker for trophoblast apoptosis in human placenta. Placenta 22, 44–48.

Kadyrov, M., Kingdom, J.C., Huppertz, B., 2006. Divergent trophoblast invasion and apoptosis in placental bed spiral arteries from pregnancies complicated by maternal anemia and early-onset preeclampsia/intrauterine growth restriction. Am. J. Obstet. Gynecol. 194, 557–563.

Kadyrov, M., Schmitz, C., Black, S., Kaufmann, P., Huppertz, B., 2003. Pre-eclampsia and maternal anaemia display reduced apoptosis and opposite invasive phenotypes of extravillous trophoblast. Placenta 24, 540–548.

Kauma, S.W., Huff, T.F., Hayes, N., Nilkaeo, A., 1999. Placental Fas ligand expression is a mechanism for maternal immune tolerance to the fetus. J. Clin. Endocrinol. Metab. 84, 2188–2194.

Kayisli, U.A., Selam, B., Guzeloglu-Kayisli, O., Demir, R., Arici, A., 2003. Human chorionic gonadotropin contributes to maternal immunotolerance and endometrial apoptosis by regulating Fas-Fas ligand system. J. Immunol. 171, 2305–2313.

Kitaya, K., Yasuda, J., Yagi, I., Tada, Y., Fushiki, S., Honjo, H., 2000. IL-15 expression at human endometrium and decidua. Biol. Reprod. 63, 683–687.

Koenig, J.M., Chegini, N., 2000. Enhanced expression of Fas-associated proteins in decidual and trophoblastic tissues in pregnancy-induced hypertension. Am. J. Reprod. Immunol. 44, 347–349.

Kuntz, T.B., Christensen, R.D., Stegner, J., Duff, P., Koenig, J.M., 2001. Fas and Fas ligand expression in maternal blood and in umbilical cord blood in preeclampsia. Pediatr. Res. 50, 743–749.

Langaas, M., Ferkingstad, E., Lindqvist, B.H., 2005. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Stat. Soc., Ser. B 67, 555–572.

Laskowska, M., Laskowska, K., Leszczynska-Gorzelak, B., Oleszczuk, J., 2006. Evaluation of the maternal and umbilical vein serum sFas/sFasL system in pregnancies complicated by preeclampsia with intrauterine growth retardation. Eur. J. Obstet. Gynecol. Reprod. Biol. 126, 155–159.

Makrigiannakis, A., Zoumakis, E., Kalantaridou, S., Mitsiades, N., Margioris, A., Chrousos, G.P., Gravanis, A., 2003. Corticotropin-releasing hormone (CRH) and immunotolerance of the fetus. Biochem. Pharmacol. 65, 917–921.

Marsal, K., Persson, P.H., Larsen, T., Lilja, H., Selbing, A., Sultan, B., 1996. Intrauterine growth curves based on ultrasonically estimated foetal weights. Acta Paediatr. 85, 843–848.

Medvedev, A.E., Johnsen, A.C., Haux, J., Steinkjer, B., Egeberg, K., Lynch, D.H., Sundan, A., Espevik, T., 1997. Regulation of Fas and Fas-ligand expression in NK cells by cytokines and the involvement of Fas-ligand in NK/LAK cell-mediated cytotoxicity. Cytokine 9, 394–404.

Mor, G., Gutierrez, L.S., Eliza, M., Kahyaoglu, F., Arici, A., 1998. Fas-fas ligand system-induced apoptosis in human placenta and gestational trophoblastic disease. Am. J. Reprod. Immunol. 40, 89–94.

Murakoshi, H., Matsuo, H., Laoag-Fernandez, J.B., Samoto, T., Maruo, T., 2003. Expression of Fas/Fas-ligand, Bcl-2 protein and apoptosis in extravillous trophoblast along invasion to the decidua in human term placenta. Endocr. J. 50, 199–207.

Olivares, E.G., Montes, M.J., Oliver, C., Galindo, J.A., Ruiz, C., 1997. Cultured human decidual stromal cells express B7-1 (CD80) and B7-2 (CD86) and stimulate allogeneic T cells. Biol. Reprod. 57, 609–615.

Qiu, Q., Yang, M., Tsang, B.K., Gruslin, A., 2005. Fas ligand expression by maternal decidual cells is negatively correlated with the abundance of leukocytes present at the maternal–fetal interface. J. Reprod. Immunol. 65, 121–132.

Redman, C.W., Sargent, I.L., 2005. Latest advances in understanding preeclampsia. Science 308, 1592–1594.

Rogers, A.M., Boime, I., Connolly, J., Cook, J.R., Russell, J.H., 1998. Maternal-fetal tolerance is maintained despite transgene-driven trophoblast expression of MHC class I, and defects in Fas and its ligand. Eur. J. Immunol. 28, 3479–3487.

Ruiz, C., Montes, M.J., Abadia-Molina, A.C., Olivares, E.G., 1997. Phagocytosis by fresh and cultured human decidual stromal cells: opposite effects of interleukin-1 alpha and progesterone. J. Reprod. Immunol. 33, 15–26.

Runic, R., Lockwood, C.J., Ma, Y., Dipasquale, B., Guller, S., 1996. Expression of Fas ligand by human cytotrophoblasts: implications in placentation and fetal survival. J. Clin. Endocrinol. Metab. 81, 3119–3122.

Saito, S., Makino, T., 2005. IX International Congress on Reproductive Immunology, 11–15 October 2004, Hakone, Japan. J. Reprod. Immunol. 68, 121–126.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article 3 http://www.bepress.com/sagmb/vol3/iss1/art3.

Staff, A.C., Ranheim, T., Khoury, J., Henriksen, T., 1999. Increased contents of phospholipids, cholesterol, and lipid peroxides in decidua basalis in women with preeclampsia. Am. J. Obstet. Gynecol. 180, 587–592.

Straszewski-Chavez, S.L., Abrahams, V.M., Mor, G., 2005. The role of apoptosis in the regulation of trophoblast survival and differentiation during pregnancy. Endocr. Rev. 26, 877–897.

Uckan, D., Steele, A., Cherry, Wang, B.Y., Chamizo, W., Koutsonikolis, A., Gilbert-Barness, E., Good, R.A., 1997. Trophoblasts express Fas ligand: a proposed mechanism for immune privilege in placenta and maternal invasion. Mol. Hum. Reprod. 3, 655–662.

Yue, X.Y., Zhang, X., Cui, S.H., Wang, X.Q., 2005. Expression of Fas antigen and ligand, placental growth factor in placenta of pregnant women with pre-eclampsia. Zhonghua Fu Chan Ke Za Zhi 40, 320–322.

Paper III

Paper IV

COMPARISON OF PREDICTIVE VALUES FROM TWO DIAGNOSTIC TESTS IN LARGE SAMPLES

CLARA-CECILIE GÜNTHER*, ØYVIND BAKKE*, STIAN LYDERSEN# AND METTE LANGAAS*
* Department of Mathematical Sciences.
# Department of Cancer Research and Molecular Medicine.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

DECEMBER 2008

SUMMARY

Within the field of diagnostic tests, the positive predictive value is the probability of being diseased given that the diagnostic test is positive. Two diagnostic tests are applied to each subject in a study and in this report we look at statistical hypothesis tests for large samples to compare the positive predictive values of the two diagnostic tests. We propose a likelihood ratio test and a restricted difference test, and we perform simulation experiments to compare these tests with existing tests. For comparing the negative predictive values of the diagnostic tests, i.e. the probability of not being diseased given that the test is negative, we propose negative predictive versions of the same tests. The simulation experiments show that the restricted difference test performs well in terms of test size.

1 INTRODUCTION

Diagnostic tests are used in medicine to e.g. detect diseases and give prognoses. Diagnostic tests can for example be based on blood samples, X-ray scans, mammography, ultrasound or computed tomography (CT). Mammography is used for detecting breast cancer, a blood sample may show if an individual has an infection, fractures may be detected from X-ray images, gallstones in the gallbladder can be found using ultrasound, and CT scans are useful for identifying tumours in the liver. A diagnostic test can have several outcomes or the outcome may be continuous, but it can often be dichotomized in terms of presence or absence of a disease and we will only consider diagnostic tests for which the disease status is binary.

When evaluating the performance of diagnostic tests, the sensitivity, specificity and positive and negative predictive values are the common accuracy measures. The sensitivity and specificity are probabilities of the test outcomes given the disease status. The sensitivity is the probability of a positive test outcome given that the disease is present and the specificity is the probability of a negative outcome given no disease. These measures tell us the degree to which the test reflects the true disease status.

The predictive values are probabilities of disease given the test result. The positive predictive value (PPV) is the probability that a subject who has a positive test outcome has the disease and the negative predictive value (NPV) is the probability that a subject who has a negative test outcome does not have the disease. The predictive values give information about the prediction capabilities of the test. For a perfect test both the PPV and NPV are 1, the test result will then give the true disease status for each subject.

When there are several diagnostic tests available for the same disease, we are interested in knowing which test is the best to use, but depending on what we mean by best, there are different methods available. If we want to find the best test regarding the ability to give a correct test outcome given the disease status then e.g. McNemar's test, see Alan Agresti (2002), can be used for comparing the sensitivity or specificity of two tests evaluated on the same subjects.

A test that has a high sensitivity and specificity will be most likely to give the patient the correct test result. However, for the patient it is utterly important to be correctly diagnosed and thereby getting the right treatment. We need to take into account the prevalence of the disease. If the prevalence is low, the probability that the patient does have the disease when the test result is positive, will be small even if the sensitivity of the applied test is high. Therefore, comparing the positive or negative predictive values is often more relevant in clinical practise as discussed by Guggenmoos-Holzmann and van Houwelingen (2000).

In the remainder of this work, we wish to test if the positive or negative predictive values of two diagnostic tests are equal. In this report we apply existing tests by Leisenring, Alonzo and Pepe (2000) and Wang, Davis and Soong (2006), we propose a likelihood ratio test, and suggest improvements for some of the already existing tests in the large sample case.

In Section 2 we describe the model and the structure of the data and define the predictive values. The null hypothesis, along with our proposed methods and already existing methods are presented in Section 3. A simulation study is conducted to compare the methods in Section 4. In Section 5 the methods are applied to data from the literature. We also present an alternative model and test statistic for the likelihood ratio test in Section 6. The results are summarised in the conclusions in Section 7.

2 Model and data

Next we define the random variables and the model used to describe the situation when comparing the predictive values.

2.1 Definitions

Two tests, test A and test B, are evaluated on each subject in a study. Each test can have a positive or negative outcome, i.e. indicating whether the subject has the disease under study or not. The true disease status for each subject is assumed to be known. For each subject, we define three events:

- D: The subject has the disease.
- A: Test A is positive.
- B: Test B is positive.

Let D^*, A^* and B^* denote the complementary events. The situation can then be illustrated by a Venn diagram as in Figure 1. There are eight mutually exclusive events and we define the random variable

N_i, $i = 1, ..., 8$, to be the number of times event i occurs. In total there are $N = N_1 + \ldots + N_8$ subjects in the study. Table 1 gives an overview of the notation for the eight random variables in terms of the events A, B, D and their complements.

Notation	Alternative notation	Explanation
N_1	$N_{A\cap B\cap D^*}$	number of non-diseased subjects with positive tests A and B
N_2	$N_{A\cap B^*\cap D^*}$	number of non-diseased subjects with positive test A and negative test B
N_3	$N_{A^*\cap B\cap D^*}$	number of non-diseased subjects with negative test A and positive test B
N_4	$N_{A^*\cap B^*\cap D^*}$	number of non-diseased subjects with negative tests A and B
N_5	$N_{A\cap B\cap D}$	number of diseased subjects with positive tests A and B
N_6	$N_{A\cap B^*\cap D}$	number of diseased subjects with positive test A and negative test B
N_7	$N_{A^*\cap B\cap D}$	number of diseased subjects with negative test A and positive test B
N_8	$N_{A^*\cap B^*\cap D}$	number of diseased subjects with negative tests A and B

TABLE 1: Notation for the random variables defined by the events A, B and D and their complements.

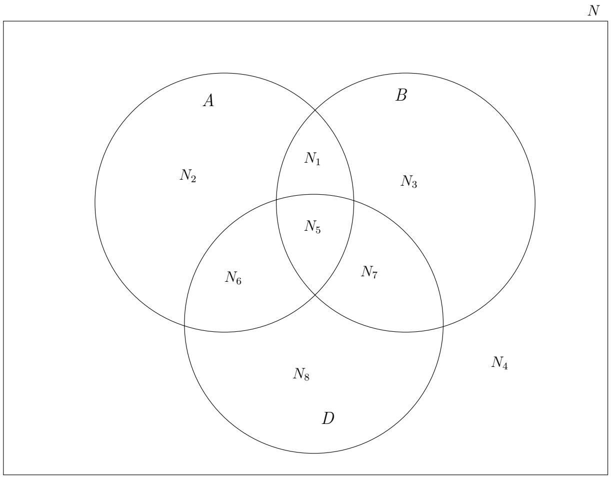


FIGURE 1: Venn diagram for the events D, A and B showing which events the random variables $N_1, ..., N_8$ correspond to.

To each of the eight mutually exclusive events there corresponds an unknown probability p_i, $i = 1, \ldots, 8$, where $\sum_{i=1}^{8} p_i = 1$, which is the probability that event i occurs for a randomly chosen subject. The positive predictive values of test A and test B can be expressed in terms of these probabilities and are given as

$$\text{PPV}_A = P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6}$$

and

$$\text{PPV}_B = P(D|B) = \frac{P(D \cap B)}{P(B)} = \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7}.$$

Similarly, the negative predictive values of test A and B are

$$\text{NPV}_A = P(D^*|A^*) = \frac{P(D^* \cap A^*)}{P(A^*)} = \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8}$$

and

$$\text{NPV}_B = P(D^*|B^*) = \frac{P(D^* \cap B^*)}{P(B^*)} = \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8}.$$

The predictive values are dependent on the prevalence of the disease, $P(D)$, which is the probability that a randomly chosen subject has the disease. For the positive predictive value,

$$\text{PPV}_A = P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{P(A|D) \cdot P(D)}{P(A|D) \cdot P(D) + (1 - P(A^*|D^*)) \cdot (1 - P(D))},$$

where $P(A|D)$ is the sensitivity and $P(A^*|D^*)$ is the specificity of test A. When $P(A) = P(B)$ testing if $\text{PPV}_A = \text{PPV}_B$ is equivalent to testing if $P(A|D) = P(B|D)$, i.e. testing whether the sensitivities of the two tests are equal. We assume that the prevalence among the subjects in the study is the same as the prevalence in the population, and this can be achieved with a cohort study in which the subjects are randomly selected.

2.2 The multinomial model

Given the total number of subjects N in the study, the random variables $N_1, N_2, ..., N_8$ can be seen to be multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and N, where $\sum_{i=1}^{8} p_i = 1$. The joint probability distribution of $N_1, N_2, ..., N_8$ is

$$P\left(\bigcap_{i=1}^{8}(N_i = n_i)\right) = N! \prod_{i=1}^{8} \frac{{p_i}^{n_i}}{n_i!}.$$

The expectation of N_i is

$$\text{E}(N_i) = \mu_i = Np_i$$

for $i = 1, ..., 8$, and the variance is

$$\text{Var}(N_i) = Np_i(1 - p_i).$$

The covariance between N_i and N_j is

$$\text{Cov}(N_i, N_j) = -Np_ip_j$$

for $i \neq j$. This leads to the covariance matrix

$$\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{N}) = N(\text{Diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p}),$$

for the multinomial distribution, Johnson, Kotz and Balakrishan (1997). The general unrestricted maximum likelihood estimator of p_i is

$$\hat{p}_i = n_i/N \tag{1}$$

for $i = 1, ..., 8$.

2.3 DATA

For a number of subjects under study, we observe for each $i = 1, ..., 8$, the number of times event i occurs among the N subjects, n_i. Table 2 shows the observed data in a 2^3 contingency table. In the following, let $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8)$ be the vector of the observed data. Using the unrestricted maximum likelihood estimators for $\boldsymbol{p}$, we can then estimate the positive and negative predictive values of test A and B as follows:

$$\widehat{\text{PPV}}_A = \frac{n_5 + n_6}{n_1 + n_2 + n_5 + n_6}, \quad \widehat{\text{PPV}}_B = \frac{n_5 + n_7}{n_1 + n_3 + n_5 + n_7}$$

$$\widehat{\text{NPV}}_A = \frac{n_3 + n_4}{n_3 + n_4 + n_7 + n_8}, \quad \widehat{\text{NPV}}_B = \frac{n_2 + n_4}{n_2 + n_4 + n_6 + n_8}.$$

		Subjects without disease		Subjects with disease	
		Test B		Test B	
		+	−	+	−
Test A	+	n_1	n_2	n_5	n_6
	−	n_3	n_4	n_7	n_8

TABLE 2: Observed data $n_1, ..., n_8$ presented in a 2^3 contingency table.

3 METHOD

Assume that we would like to test the null hypothesis that the positive predictive values are equal for test A and B, i.e. $\text{PPV}_A = \text{PPV}_B$. The null hypothesis can be written as

$$\text{H}_0^P : P(D|A) = P(D|B), \text{ i.e. } \text{H}_0^P : \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6} = \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7}. \quad (2)$$

Alternatively, if we would like to test whether the negative predictive values are equal for test A and B, i.e. if $\text{NPV}_A = \text{NPV}_B$, the null hypothesis is

$$\text{H}_0^N : P(D^*|A^*) = P(D^*|B^*), \text{ i.e. } \text{H}_0^N : \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8} = \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8}. \quad (3)$$

Our alternative hypotheses will be that the predictive values are not equal, i.e.
$H_1^P : P(D|A) \neq P(D|B)$ and $H_1^N : P(D^*|A^*) \neq P(D^*|B^*)$.

3.1 LIKELIHOOD RATIO TEST

One possibility to test the null hypothesis in (2) is to use a likelihood ratio test. We first write down the test statistic and then describe how to find the maximum likelihood estimates of parameters.

3.1.1 TEST STATISTIC

In a general setting, if we want to test $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$ where $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_0^c = \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ denotes the entire parameter space, we may use a likelihood ratio test. This approach was also suggested by Leisenring et al. (2000), who faced numerical difficulties trying to implement it. The likelihood ratio test statistic is in general defined as

$$\lambda(\boldsymbol{n}) = \frac{\sup_{\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\boldsymbol{n})}{\sup_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\boldsymbol{n})}$$

where $\boldsymbol{n}$ is the observed data, Casella and Berger (2002). The denominator of $\lambda(\boldsymbol{n})$ is the maximum likelihood of the observed sample over the entire parameter space and the numerator is the maximum likelihood of the observed sample over the parameters satisfying the null hypothesis. Let $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8)$ be the vector of the random variables. When the sample size is large,

$$-2 \cdot \log\lambda(\boldsymbol{N}) \approx \chi_k^2$$

i.e. $-2 \cdot \log\lambda(\boldsymbol{N})$ is χ^2 distributed with k degrees of freedom where k is the difference between the number of free parameters in the unrestricted case and under the null hypothesis.

Let $\boldsymbol{\theta} = \boldsymbol{p} = (p_1, \ldots, p_8)$ be the parameters in the multinomial distribution and $\boldsymbol{n} = (n_1, \ldots, n_8)$ the observed data. The log-likelihood to be maximized for the multinomial distribution is

$$l(\boldsymbol{p}) = \log L(\boldsymbol{p}|\boldsymbol{n}) = c + \sum_{i=1}^{8} n_i \cdot \log(p_i) \tag{4}$$

where c is a constant.

The sum of p_1, p_2, ..., p_8 must equal 1,

$$\sum_{i=1}^{8} p_i = 1. \tag{5}$$

Under the null hypothesis that the positive predictive values for the two tests are equal, their difference δ_P is zero, i.e.

$$\delta_P = \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6} - \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7} = 0. \tag{6}$$

In the unrestricted case (i.e. $H_0 \cup H_1$), the maximum likelihood estimates for $p_1, \ldots, p_8$ are the estimates given by (1), which satisfy (5). Under the null hypothesis, the estimates cannot be given in closed form and we will need to use an optimization routine to estimate $p_1, \ldots, p_8$ by maximizing the log-likelihood (4) under the constraints (5) and (6).

Let $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6, \hat{p}_7, \hat{p}_8)$ be the unconstrained maximum likelihood estimates and $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4, \tilde{p}_5, \tilde{p}_6, \tilde{p}_7, \tilde{p}_8)$ the maximum likelihood estimates under the null hypothesis. Then, in our model, asymptotically as N is large,

$$-2 \cdot \log(\lambda(\boldsymbol{n})) = -2 \left(\sum_{i=1}^{8} n_i \cdot (\log(\tilde{p}_i) - \log(\hat{p}_i)) \right) \approx \chi_1^2. \tag{7}$$

We have one less free parameter in the restricted case because of the constraint (6).

For testing whether the negative predictive values for the two tests are equal, the constraint δ_P (6) is replaced by δ_N, where

$$\delta_N = \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8} - \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8} = 0. \tag{8}$$

3.1.2 Finding maximum likelihood estimates under the null hypotheses

To find the maximum likelihood estimates under the null hypothesis, we can either maximize the likelihood function under the given constraints using a numerical optimization routine or find the estimates analytically by solving a system of equations. In both approaches we use Lagrange multipliers and in either case we have two constraints.

Numerical maximization of the log-likelihood If we want to find the maximum likelihood estimates using an optimization routine, the goal is to find the values $\tilde{\boldsymbol{p}}$ under the null hypothesis such that $\log L(\tilde{\boldsymbol{p}}) \geq \log L(\boldsymbol{p})$ for all $\boldsymbol{p}$ that satisfies the two constraints (5) and (6).

To maximize the log-likelihood (4) under the null hypotheses, we use the R interface version of TANGO (Trustable Algorithms for Nonlinear General Optimization), see Andreani, Birgin E. G., Martinez and Schuverdt (2007) and Andreani, Birgin, Martinez and Schuverdt (2008), which is a set of Fortran routines for optimization. In order to run the program, one must specify the objective function and the constraint and their corresponding first order derivatives. We reparametrize the problem by setting

$$\begin{aligned}
p_1 &= \frac{1}{1 + e^{y_1} + ... + e^{y_7}}, \\
p_2 &= \frac{e^{y_1}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}, \\
p_3 &= \frac{e^{y_2}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}, \\
&\vdots \\
p_8 &= \frac{e^{y_7}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}
\end{aligned}$$

where $-\infty < y_i < \infty$, $i = 1, \dots, 7$. This reparametrization ensures that the constraint (5) is satisfied, in addition to restricting the estimated probabilities to be $0 \leq p_i \leq 1$, $i = 1, \dots, 8$. Let $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5, y_6, y_7)$. The constraint under the null hypothesis (2) is then

$$\delta_{P,\boldsymbol{y}} = \frac{e^{y_4} + e^{y_5}}{1 + e^{y_1} + e^{y_4} + e^{y_5}} - \frac{e^{y_4} + e^{y_6}}{1 + e^{y_2} + e^{y_4} + e^{y_6}} = 0 \tag{9}$$

and the constraint under the null hypothesis (3) is

$$\delta_{N,\boldsymbol{y}} = \frac{e^{y_2} + e^{y_3}}{e^{y_2} + e^{y_3} + e^{y_6} + e^{y_7}} - \frac{e^{y_1} + e^{y_3}}{e^{y_1} + e^{y_3} + e^{y_5} + e^{y_7}} = 0. \tag{10}$$

These constraints are both non-linear equality constraints. The TANGO program uses an augmented Lagrangian algorithm to find the minimum of the negative log-likelihood while ensuring that the H_0 constraints (9) and (10) are satisfied when testing the null hypotheses (2) and (3) respectively. The

Lagrangian multiplier is updated successively starting by an initial value that must be set. We also set the initial value of $\boldsymbol{y}$ and its lower and upper bounds. The value of $\boldsymbol{y}$ at the optimum is returned. Some computational remarks are given in Appendix D.

ANALYTICAL MAXIMIZATION OF THE LOG-LIKELIHOOD Another approach is to find the estimates analytically by solving a system of equations arising from the method of Lagrange multipliers, for an introduction see Edwards and Penney (1998). The constraint under the null hypothesis can be rewritten as

$$k(\boldsymbol{p}) = p_1 p_7 + p_2 p_7 + p_2 p_5 - p_1 p_6 - p_3 p_5 - p_3 p_6 = 0. \tag{11}$$

In addition, let $h(\boldsymbol{p})$ be the constraint that $p_1, ..., p_8$ must sum to one,

$$h(\boldsymbol{p}) = \sum_{i=1}^{8} p_i = 1, \tag{12}$$

and let $l(\boldsymbol{p})$ be the log-likelihood function given in (4).

The system of equations to be solved then consists of

$$\nabla l = \gamma \nabla h + \kappa \nabla k \tag{13}$$

where γ and κ are Lagrangian multipliers, together with the above constraints.

The partial derivatives of the log-likelihood l and the constraints h and k with respect to p_1, p_2, p_3, p_4, p_5, p_6, p_7 and p_8 are given by

$$\nabla l = \left(\frac{n_1}{p_1}, \frac{n_2}{p_2}, \frac{n_3}{p_3}, \frac{n_4}{p_4}, \frac{n_5}{p_5}, \frac{n_6}{p_6}, \frac{n_7}{p_7}, \frac{n_8}{p_8} \right), \tag{14}$$

$$\nabla k = (p_7 - p_6, p_5 + p_7, -p_5 - p_6, 0, p_2 - p_3, -p_1 - p_3, p_1 + p_2, 0), \tag{15}$$

and

$$\nabla h = (1, 1, 1, 1, 1, 1, 1, 1). \tag{16}$$

From Equations (11) – (16) we obtain the following system of equations, which consists of ten equations and ten unknown variables

$$\begin{aligned}
n_1 &= p_1(\gamma + \kappa(p_7 - p_6)) \\
n_2 &= p_2(\gamma + \kappa(p_5 + p_7)) \\
n_3 &= p_3(\gamma + \kappa(-p_5 - p_6)) \\
n_4 &= p_4\gamma \\
n_5 &= p_5(\gamma + \kappa(p_2 - p_3)) \\
n_6 &= p_6(\gamma + \kappa(-p_1 - p_3)) \\
n_7 &= p_7(\gamma + \kappa(p_1 + p_2)) \\
n_8 &= p_8\gamma \\
\sum_{i=1}^{8} p_i &= 1 \\
p_1p_7 + p_2p_7 + p_2p_5 - p_1p_6 - p_3p_5 - p_3p_6 &= 0.
\end{aligned} \tag{17}$$

The denominators of (14) have been multiplied over to the right hand side in order to allow for $p_i = 0$ as a possible solution for $n_i = 0$. Obviously, l cannot have a maximum value $p_i = 0$ if $n_i \neq 0$, as $l(\boldsymbol{p})$ would be $-\infty$ in this case. The solutions of this system of equations involve roots of third degree polynomials, and we have used the Maple 12 command solve to find solutions. Among its solutions, the one that maximizes $l(\boldsymbol{p})$ and where all $p_i \geq 0$ yields the likelihood estimates $\tilde{\boldsymbol{p}}_i$ under the null hypothesis. We can show that when $n_i = 0$, the corresponding likelihood estimate under the null hypothesis $\tilde{p}_i$ is 0 for $i = 1, 4, 5, 8$, but that this is not necessarily true for $i = 2, 3, 6, 7$. For $\tilde{p}_4$ and $\tilde{p}_8$ we have the more general result that $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$, see Appendix A for the proofs.

A gradient based optimization routine searches for the global minimum across the negative log-likelihood surface and it can get stuck in a local minimum. In our experience this especially happens when some of the cell counts in the contingency table are small. The analytical maximization might yield more accurate estimates in these cases, see Appendix D.

3.2 Difference based tests

Other possible test statistics start out by looking at the difference of the PPVs, and then these test statistics can be standardized by using Taylor series expansion. We also suggest some improvement to these tests.

3.2.1 Test statistics

Based on the difference δ_P given in Equation (6), which equals zero under the null hypothesis, we may suggest a variety of possible test statistics.

Wang et al. (2006) suggested the test statistics

$$g_1(\boldsymbol{N}) = \frac{N_5 + N_6}{N_1 + N_2 + N_5 + N_6} - \frac{N_5 + N_7}{N_1 + N_3 + N_5 + N_7} \tag{18}$$

and

$$g_2(\boldsymbol{N}) = \log \frac{(N_5 + N_6)(N_1 + N_3 + N_5 + N_7)}{(N_5 + N_7)(N_1 + N_2 + N_5 + N_6)}. \tag{19}$$

For a more detailed description of their work, see Appendix B.1. Moskowitz and Pepe (2006) also suggest a similar test statistic to $g_2(\boldsymbol{N})$, see Appendix B.2.

Since the null hypothesis can be written

$$H_0^P : \frac{p_1 + p_3}{p_5 + p_7} = \frac{p_1 + p_2}{p_5 + p_6}$$

another test statistic to be used may be

$$g_3(\boldsymbol{N}) = \frac{N_1 + N_3}{N_5 + N_7} - \frac{N_1 + N_2}{N_5 + N_6}.$$

Another possibility is to use the log ratio of the terms, instead of their difference,

$$g_4(\boldsymbol{N}) = \log \frac{(N_1 + N_3)(N_5 + N_6)}{(N_1 + N_2)(N_5 + N_7)}$$

or we may rewrite the null hypothesis in order to obtain

$$g_5(\boldsymbol{N}) = \frac{N_5 + N_6}{N_1 + N_2} - \frac{N_5 + N_7}{N_1 + N_3}.$$

3.2.2 Standardization by Taylor series expansion

For a general test statistic, $g(\boldsymbol{N})$, we may construct a standardized test statistic by subtracting the expectation of the test statistic, $\mathrm{E}(g(\boldsymbol{N}))$, and dividing by its standard deviation, $\sqrt{\mathrm{Var}(g(\boldsymbol{N}))}$. In the large sample case the square of the standardized test statistics may then be assumed to be approximately χ_1^2-distributed,

$$T(\boldsymbol{N}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left\{ \frac{g(\boldsymbol{N}) - \mathrm{E}(g(\boldsymbol{N}))}{\sqrt{\mathrm{Var}(g(\boldsymbol{N}))}} \right\}^2 \approx \chi_1^2. \tag{20}$$

The expectation and variance of the test statistic can be approximated with the aid of Taylor series expansion as suggested by Wang et al. (2006). Let $\mathrm{E}(\boldsymbol{N}) = \boldsymbol{\mu}$ be the point around which the expansion is centered. As before, $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{N})$. A second order Taylor expansion in matrix notation is given as

$$g(\boldsymbol{N}) \approx g(\boldsymbol{\mu}) + \boldsymbol{G}^T(\boldsymbol{\mu})(\boldsymbol{N} - \boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{N} - \boldsymbol{\mu})^T \boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N} - \boldsymbol{\mu}) \tag{21}$$

where $\boldsymbol{G}$ is a vector containing the first order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$ and $\boldsymbol{G}^T$ is the transpose of $\boldsymbol{G}$. Further $\boldsymbol{H}$ is a matrix with second order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$, i.e. the Hessian matrix.

The expectation of $g(\boldsymbol{N})$ can then be approximated as

$$\mathrm{E}(g(\boldsymbol{N})) \approx g(\boldsymbol{\mu})$$

for the first order Taylor expansion and as

$$\mathrm{E}(g(\boldsymbol{N})) \approx g(\boldsymbol{\mu}) + \frac{1}{2}\mathrm{tr}(\boldsymbol{H}(\boldsymbol{\mu})\boldsymbol{\Sigma}) \tag{22}$$

for the second order Taylor expansion, since

$$\begin{aligned} & \mathrm{E}\left(\tfrac{1}{2}(\boldsymbol{N}-\boldsymbol{\mu})^T\boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})\right) \\ = \ & \mathrm{E}\left(\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})^T(\boldsymbol{N}-\boldsymbol{\mu})\right)\right) \\ = \ & \tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{H}(\boldsymbol{\mu})\mathrm{E}((\boldsymbol{N}-\boldsymbol{\mu})^T(\boldsymbol{N}-\boldsymbol{\mu}))\right) \end{aligned}$$

where we have used the result $\boldsymbol{x}^TA\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^TA\boldsymbol{x}) = \mathrm{tr}(A\boldsymbol{x}\boldsymbol{x}^T)$ where $\boldsymbol{x} = \boldsymbol{N} - \boldsymbol{\mu}$ and A is the Hessian matrix $\boldsymbol{H}$. $\mathrm{E}((\boldsymbol{N}-\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})^T)$ is the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{N}$.

The variance of $g(\boldsymbol{N})$ can be approximated as

$$\mathrm{Var}(g(\boldsymbol{N})) \approx \boldsymbol{G}^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\,\boldsymbol{G}(\boldsymbol{\mu})$$

for the first order Taylor expansion. Using a second order Taylor expansion for the variance requires finding the third and fourth order moments of $\boldsymbol{N}$.

Using the first order Taylor approximation of $\mathrm{E}(g(\boldsymbol{N}))$ and $\mathrm{Var}(g(\boldsymbol{N}))$ in the standardized test statistic of (20) yields

$$T(\boldsymbol{N}) = \frac{(g(\boldsymbol{N}) - g(\boldsymbol{\mu}))^2}{\boldsymbol{G}^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\boldsymbol{G}(\boldsymbol{\mu})} \approx \chi_1^2. \tag{23}$$

Under the null hypothesis, $g(\boldsymbol{\mu}) = 0$. $\boldsymbol{G}(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ are functions of the unknown parameters $\boldsymbol{p}$ and needs to be estimated. We can either insert the general maximum likelihood estimates $\hat{p}_i = n_i/N$ or the maximum likelihood estimates $\tilde{p}_i$ under H_0^P, as found in Section 3.1.2. When we use the standardized test statistic (23) with $g_1(\boldsymbol{N})$ and estimate the variance using the maximum likelihood estimates under H_0 we refer to it as the *restricted difference test*. If we instead use the unrestricted maximum likelihood estimates to estimate the variance, we refer to it as the *unrestricted difference test*.

We have investigated two possible improvements of the standardized test statistics. In addition to using the restricted maximum likelihood estimates to estimate the variance of (23), we have looked at the effect of using a second order Taylor series approximation to $\mathrm{E}(g(\boldsymbol{N}))$ as an attempt to arrive at a more accurate approximation to a χ_1^2 distributed test statistic. The expectation and variance in the standardized test statistic given in (20) is found using a first order Taylor series expansion and the difference between using the first order and the second order Taylor series approximation to $\mathrm{E}(g(\boldsymbol{N}))$ is the term $1/2 \cdot \mathrm{tr}(\boldsymbol{H}(\boldsymbol{\mu})\boldsymbol{\Sigma})$. For the simulation experiment in Section 4 this turned out to be very small as compared to the denominator of (23).

3.3 Test by Leisenring, Alonzo and Pepe (LAP)

Leisenring et al. (2000) present a test for the null hypothesis given in (2). We will denote this the LAP test. They define three binary random variables; D_{ij} that denotes disease status, Z_{ij} that indicates which test was used and X_{ij} that describes the outcome of the diagnostic test for test j, $j = 1, 2$, for subject i, $i = 1, \ldots, N$.

$$D_{ij} = \begin{cases} 0, & \text{non-diseased} \\ 1, & \text{diseased} \end{cases}$$

$$Z_{ij} = \begin{cases} 0, & \text{test A} \\ 1, & \text{test B} \end{cases}$$

$$X_{ij} = \begin{cases} 0, & \text{negative} \\ 1, & \text{positive} \end{cases}$$

The PPV of test A can be written as PPV$_A = P(D_{ij} = 1 \mid Z_{ij} = 0, X_{ij} = 1)$ and the PPV of test B as PPV$_B = P(D_{ij} = 1 \mid Z_{ij} = 1, X_{ij} = 1)$. Based on generalized estimation equations Leisenring et al. (2000) fit the generalized linear model

$$\text{logit}(P(D_{ij} = 1 \mid Z_{ij}, X_{ij} = 1)) = \alpha_P + \beta_P Z_{ij}.$$

Testing the null hypothesis H_0: PPV$_A$ = PPV$_B$ is equivalent to testing the null hypothesis $H_0: \ \beta_P = 0$. To derive the generalized score statistic, an independent working correlation structure is assumed for the score function and the corresponding variance function is $v_{ij} = \mu_{ij}(1 - \mu_{ij})$ where $\mu_{ij} = E(D_{ij})$. The score function is then $S_P = \sum_{i=1}^{N_p} \sum_{j=1}^{m_i} Z_{ij}(D_{ij} - \bar{D})$ which also can be written as $S_P = \sum_{i=1}^{N_p} \sum_{j=1}^{m_i} D_{ij}(Z_{ij} - \bar{Z})$. Here N_p is the number of subjects with at least one positive test outcome and m_i is the number of positive test results for subject i.

$$\bar{Z} = \frac{\sum_{i=1}^{N_p} m_i Z_i D_i}{\sum_{i=1}^{N_p} m_i}$$

is the proportion of positive B tests for the diseased subjects among all the positive tests and

$$\bar{D} = \frac{\sum_{i=1}^{N_P} m_i D_i}{\sum_{i=1}^{N_P} m_i}$$

is the proportion of positive tests for the diseased subjects among all the positive tests.

The resulting test statistic for testing the null hypothesis $H_0 : \beta_P = 0$ is obtained by taking the square of the score function divided by its variance:

$$T_{PPV} = \frac{\left\{\sum_{i=1}^{N_p} \sum_{j=1}^{m_i} D_{ij}(Z_{ij} - \bar{Z})\right\}^2}{\sum_{i=1}^{N_p} \left\{\sum_{j=1}^{m_i} (D_{ij} - \bar{D})(Z_{ij} - \bar{Z})\right\}^2}. \tag{24}$$

Under the null hypothesis, this test statistic is asymptotically χ_1^2-distributed. It is worth noting that only the subjects with at least one positive test outcome contribute to the test statistic (24).

The test statistic in (24) is general and can be used even if the disease status is not constant within a subject. Usually the disease status will be constant within the subject and the test statistic can be then simplified. By defining $T_i = \sum_{j=1}^{m_i} Z_{ij}$, the number of positive B tests subject i contributes to the analysis, the statistic can be written

$$T_{\text{PPV}} = \frac{\left\{\sum_{i=1}^{N_p} D_i (T_i - m_i \bar{Z})\right\}^2}{\sum_{i=1}^{N_p} (D_i - \bar{D})^2 (T_i - m_i \bar{Z})^2}.$$

We derived the test statistic by using our notation of the eight mutually exclusive events in Figure 1. The numerator can be separated into six terms, in three of which the disease status $D = 0$ and three where $D = 1$, by noting that $T_i = 0$ and $m_i = 1$ when only test A is positive, $T_i = 1$ and $m_i = 1$ when only test B is positive and $T_i = 1$ and $m_i = 2$ when both tests are positive. Then

$$T_{\text{PPV}} = \frac{((N_1 + N_2 + N_5 + N_6)(N_5 + N_7) - (N_1 + N_3 + N_5 + N_7)(N_5 + N_6))^2}{f(N_1, N_2, N_3, N_5, N_6, N_7)} \tag{25}$$

where

$$\begin{aligned} &f(N_1, N_2, N_3, N_5, N_6, N_7) \\ &\qquad = N_1(N_2 - N_3 + N_6 - N_7)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 \\ &\qquad\quad + N_2(N_1 + N_3 + N_5 + N_7)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 \\ &\qquad\quad + N_3(N_1 + N_2 + N_5 + N_6)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 \\ &\qquad\quad + N_5(N_2 - N_3 + N_6 - N_7)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 \\ &\qquad\quad + N_6(N_1 + N_3 + N_5 + N_7)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 \\ &\qquad\quad + N_7(N_1 + N_2 + N_5 + N_6)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7} \right)^2 . \end{aligned}$$

To compare the NPVs for test A and test B, Leisenring et al. (2000) fit the generalized linear model

$$\text{logit}(P(D_{ij} = 1 | Z_{ij}, X_{ij} = 0)) = \alpha_N + \beta_N Z_{ij}.$$

by using the generalized estimating equations method. The null hypothesis in this case is $H_0 : \beta_N = 0$. Under the assumption that disease status is constant within a subject, this leads to the test statistic

$$T_{\text{NPV}} = \frac{\left\{ \sum_{i=1}^{N_n} D_i(T_i - k_i \bar{Z}) \right\}^2}{\sum_{i=1}^{N_n} (D_i - \bar{D})^2 (T_i - k_i \bar{Z})^2}$$

where N_n is the number of subjects with at least one negative test outcome and k_i is the number of negative test results for subject i. Only the subjects with at least one negative test outcome contribute to this test statistic.

Leisenring et al. (2000) also propose a Wald test based on the estimates of the regression coefficients, but their simulation studies show that the score test performs better.

4 SIMULATION STUDY

In order to compare the test size under the null hypothesis for the tests presented in Section 3 and to assess the power of the tests under the alternative hypothesis, we perform a simulation experiment.

All the tests are asymptotic tests, but it is not clear how large the sample size has to be for the tests to preserve their test size. Therefore we will consider different sample sizes. Two different simulation strategies for generating datasets will be presented. The maximum likelihood estimates under the null hypotheses needed for the likelihood ratio test and the restricted difference test are found using TANGO as described in Section 3.1.2. All analyses are performed using the R language, R Development Core Team (2008).

4.1 Simulation experiment from Leisenring, Alonzo and Pepe

The first simulation experiment is based on the simulation experiment of Leisenring et al. (2000) and we use their algorithm to generate the data. Therefore we denote this simulation experiment the LAP simulation experiment.

4.1.1 Algorithm

We generate datasets by using the algorithm presented in Appendix B in Leisenring et al. (2000). Let I_D denote the disease status,

$$I_D = \begin{cases} 1, & \text{diseased} \\ 0, & \text{non-diseased} \end{cases}$$

and I_A and I_B the test results of test A and B,

$$I_A = \begin{cases} 1, & \text{test A positive} \\ 0, & \text{test A negative} \end{cases}$$

$$I_B = \begin{cases} 1, & \text{test B positive} \\ 0, & \text{test B negative} \end{cases}$$

In order to generate the datasets, the number of subjects tested, N, the positive and negative predictive values for both tests, the prevalence of the disease $P(D)$ and the variance σ^2 for the random effect for each subject must be set. The random effect introduces correlation between the test outcomes for each subject. Our interpretation of the simulation algorithm is as follows:

1. Set N, $P(D)$, PPV$_A$, NPV$_A$, PPV$_B$, NPV$_B$ and σ.
2. Calculate the true positive rate TP and the false positive rate FP for test A and test B defined by the equations

   $$\text{TP} = \frac{(1 - P(D) - \text{NPV}) \cdot \text{PPV}}{(1 - \text{PPV} - \text{NPV}) \cdot P(D)}$$

   and

   $$\text{FP} = \frac{1 - P(D) - \text{NPV}}{(1 - P(D))(1 - \text{NPV} - \frac{\text{PPV} \cdot \text{NPV}}{1 - \text{PPV}})}.$$

3. Given TP and FP, the parameters α_i and β_i, $i = 1, 2$, for each test are calculated from the following equations,

   $$\alpha_i = \Phi^{-1}(\text{FP})\sqrt{1 + \sigma^2}$$
   $$\beta_i = \Phi^{-1}(\text{TP})\sqrt{1 + \sigma^2},$$

   where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Case no.	N	$P(D)$	σ	PPV$_A$	PPV$_B$	NPV$_A$	NPV$_B$
1	100	0.25	0.1	0.75	0.75	0.85	0.85
2	500	0.25	0.1	0.75	0.75	0.85	0.85
3	100	0.50	0.1	0.75	0.75	0.85	0.85
4	500	0.50	0.1	0.75	0.75	0.85	0.85
5	100	0.25	1.0	0.75	0.75	0.85	0.85
6	500	0.25	1.0	0.75	0.75	0.85	0.85
7	100	0.50	1.0	0.75	0.75	0.85	0.85
8	500	0.50	1.0	0.75	0.75	0.85	0.85

TABLE 3: Specifications of the cases under the null hypotheses in the LAP simulation experiment.

4. For each subject the disease status I_D is drawn independently with probability $P(D)$.

5. A random effect $r \sim N(0, \sigma^2)$ is generated for each subject.

6. Given the disease status and the random effect r, the probability of a positive test outcome for each subject is given by

$$P(I_A = 1|I_D, r) = \Phi(\alpha_1(1 - I_D) + \beta_1 I_D + r)$$

for test A and by

$$P(I_B = 1|I_D, r) = \Phi(\alpha_2(1 - I_D) + \beta_2 I_D + r)$$

for test B. The test outcomes are drawn with these probabilities for each subject.

7. Find n_1, …, n_8 by counting the number of subjects that belongs to each of the eight events described in Section 2, e.g. n_1 is the number of subjects for which $I_D = 0$, $I_A = 1$ and I_B=1, the number of subjects that are not diseased and have positive tests A and B.

The algorithm is repeated M times, providing M datasets of $n_1, \ldots, n_8$.

4.1.2 CASES UNDER STUDY

In the simulation experiment, we suggest eight cases by varying the input parameters N, $P(D)$ and σ in the LAP simulation algorithm. The setup of the experiment is a 2^3 factorial experiment, i.e. we have three factors, N, $P(D)$ and σ, and each factor has two levels. The low level for N is 100 and the high level is 500, while the low level for $P(D)$ is 0.25 and the high level is 0.50. For σ the low level is 0.1 and the high level is 1.0. The response in this experiment is the estimated test size for each test. The cases that are under the null hypotheses H_0^P and H_0^N in equations (2) and (3) are given in Table 3. For cases not under the null hypotheses, the parameters N, $P(D)$ and σ are the same, but the remaining parameters are changed and will be described below. For each of these eight cases we simulate $M = 5000$ datasets.

We generate data under the null hypotheses (2) and (3), by setting PPV$_A$ $=$ PPV$_B$ $= 0.75$ and NPV$_1$ $=$ NPV$_2$ $= 0.85$. These datasets are used to estimate the test size under H_0 for both the PPV and NPV tests. To estimate the power of the tests, we need datasets under H_1, and for PPV we generate datasets where $\text{PPV}_A = 0.85$ and $\text{PPV}_B = 0.75$ and $\text{NPV}_1 = \text{NPV}_2 = 0.85$. To compare the power for the NPV tests, we generate datasets where $\text{NPV}_1 = 0.85$ and $\text{NPV}_2 = 0.80$ and $\text{PPV}_A = \text{PPV}_B = 0.75$.

To compare the positive predictive values of test A and B, we calculate the test statistics for the LAP test, the likelihood ratio test and the unrestricted and restricted difference tests. To compare the negative predictive values for test A and B we use the negative predictive value versions of these test statistics. We calculate p-values based on the χ^2_1 distribution. We also assess the performance of the four other difference based tests as proposed in Section 3.2.1.

4.1.3 Results

A summary of the results of the simulation experiment will follow. For each case and selected value of the nominal significance level α, let W be a random variable counting the number of p-values smaller than or equal to α. Then W is binomially distributed with size M, the number of p-values generated, and probability α. An estimate of the true significance level of the test, $\hat{\alpha}$ is then

$$\hat{\alpha} = \frac{W}{M}. \tag{26}$$

Let

$$\begin{aligned} \widetilde{W} &= W + 2 \\ \widetilde{M} &= M + 4 \\ \tilde{\alpha} &= \frac{\widetilde{W}}{\widetilde{M}}. \end{aligned}$$

A $100 \cdot (1 - \gamma)\%$ confidence interval for $\hat{\alpha}$ with limits $\hat{\alpha}_L$ and $\hat{\alpha}_U$, according to Agresti and Coull (1998) is given as

$$\hat{\alpha}_L = \tilde{\alpha} - z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1 - \tilde{\alpha})}{\widetilde{M}}} \tag{27}$$

and

$$\hat{\alpha}_U = \tilde{\alpha} + z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1 - \tilde{\alpha})}{\widetilde{M}}} \tag{28}$$

where $z_{\gamma/2}$ is the $\gamma/2$-quantile in the standard normal distribution. When the samples are drawn under H_0, $\hat{\alpha}$ will be an estimate of the test size, i.e. the probability of making a type I error, $P(\text{reject } H_0 | H_0)$. A p-value is valid, as defined by Lloyd and Moldovan (2008), if the actual probability of rejecting the null hypothesis never exceeds the nominal significance level. We choose the nominal significance level to be 0.05 and we say that the test preserves its test size if the lower confidence limit is less than or equal to 0.05, i.e. if $\hat{\alpha}_L \leq 0.05$. If $\hat{\alpha}_U < 0.05$, the test is said to be conservative, while if $\hat{\alpha}_L > 0.05$ it does not keep its test size and it is then optimistic. If the samples are drawn under the alternative H_1, $\hat{\alpha}$ is an estimate of the power of the test, i.e. $P(\text{reject } H_0 \mid H_1)$, the probability to correctly reject the null hypothesis when it is not true.

Table 4 shows the estimated test size with 95% confidence limits for the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 where the data is generated under the null hypothesis that $\text{PPV}_A = \text{PPV}_B$.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.058	0.052	0.065
Case 1 Likelihood ratio test	0.065	0.059	0.072
Case 1 Restricted difference test	0.051	0.046	0.058
Case 1 Unrestricted difference test	0.067	0.060	0.074
Case 2 LAP test	0.056	0.050	0.063
Case 2 Likelihood ratio test	0.056	0.050	0.063
Case 2 Restricted difference test	0.055	0.049	0.062
Case 2 Unrestricted difference test	0.058	0.051	0.064
Case 3 LAP test	0.051	0.046	0.058
Case 3 Likelihood ratio test	0.050	0.044	0.056
Case 3 Restricted difference test	0.048	0.043	0.055
Case 3 Unrestricted difference test	0.051	0.045	0.058
Case 4 LAP test	0.057	0.051	0.064
Case 4 Likelihood ratio test	0.057	0.051	0.064
Case 4 Restricted difference test	0.057	0.051	0.064
Case 4 Unrestricted difference test	0.057	0.051	0.064
Case 5 LAP test	0.058	0.052	0.065
Case 5 Likelihood ratio test	0.070	0.063	0.077
Case 5 Restricted difference test	0.053	0.047	0.059
Case 5 Unrestricted difference test	0.065	0.058	0.072
Case 6 LAP test	0.054	0.048	0.060
Case 6 Likelihood ratio test	0.053	0.048	0.060
Case 6 Restricted difference test	0.052	0.046	0.058
Case 6 Unrestricted difference test	0.055	0.049	0.061
Case 7 LAP test	0.053	0.047	0.060
Case 7 Likelihood ratio test	0.055	0.049	0.061
Case 7 Restricted difference test	0.049	0.044	0.056
Case 7 Unrestricted difference test	0.054	0.048	0.060
Case 8 LAP test	0.055	0.049	0.062
Case 8 Likelihood ratio test	0.055	0.049	0.062
Case 8 Restricted difference test	0.054	0.048	0.061
Case 8 Unrestricted difference test	0.055	0.049	0.062

TABLE 4: Estimated test size with 95% confidence limits when testing PPV$_A$ = PPV$_B$ for data generated under the null hypothesis using the LAP-simulation algorithm.

In Case 3, 6, 7 and 8 all four test preserve the test size. In Case 2 the unrestricted difference test is too optimistic, while the other tests preserve the test size.

In Case 1 and 5 the restricted difference test is the only test preserving the test size. The other tests are too optimistic. These cases have small cell counts, and it might be that the restricted difference test is more robust towards small cell counts than the other tests. Table 5 shows the mean observed cell counts in Case 1–8 for the data generated under the null hypothesis that $\text{PPV}_A = \text{PPV}_B$. We see that in Case 1 and 5, $\bar{n}_1$ is 0.2 and 1.1 respectively, and thereby $n_1 = 0$ in many of the datasets, and also some of the other cell counts are small.

In Case 4 none of the four tests preserve the test size, i.e. all the tests are slightly optimistic. As all the cell counts are high in this case it is not surprising that the estimated test size is the same for all the tests, however we see no apparent reason why the test size is not preserved, and thus this may perhaps be a purely random event.

For the likelihood ratio test we analysed the 2^3 factorial experiment using $\hat{\alpha}$ as the response and found that the interaction between the factors N and $P(D)$ is the most significant effect on the the test size with a p-value of 0.012. When N is at its high level, $N = 500$, the test size is less affected by $P(D)$ which makes sense, since the high value of N ensures that all the cells will have large expected values unless the corresponding cell probabilities are very small. There is also a significant interaction between N and σ, when $N = 100$, the estimated test size is higher for $\sigma = 1.0$ than for $\sigma = 0.01$ and when $N = 500$, the estimated test size is lower for $\sigma = 1.0$ than for $\sigma = 0.01$.

Table 12 (see Appendix C) shows the estimated test size with 95% confidence limits for the NPV versions of the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 where the data is generated under the null hypothesis that $\text{NPV}_1 = \text{NPV}_2$. In Case 1, 2, 4 and 8, all the cases preserve the test size. In Case 5 all the cases except the likelihood ratio test preserve the test size too. In Case 3 and 7 however, only the restricted difference test preserves the test size, none of the other tests do. It may be because it is more robust to the small cell counts in the eight cell, $\bar{n}_8$, which is 0.8 and 2.2 respectively in these two cases.

Table 13 and 14 (see Appendix C) show the estimated power with 95% confidence intervals for the PPV and NPV versions respectively of the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 for the data generated under the two alternative hypotheses. The power of the restricted difference test is generally lower than the power of the other tests, which is not surprising since it preserves its test size when the other tests do not. The power increases with the number of subjects N as we would expect. For the PPV tests, it also increases when the prevalence $P(D)$ increases. When the prevalence increases it is more likely that a random subject has the disease, therefore more subjects will have the disease and there will be more positive tests. $P(D) = 0.50$ in Case 4 and 8 where the tests have higher power than in Case 2 and 6 where $P(D) = 0.25$. We also note that in general the test power is higher when $\sigma = 1$ compared to when $\sigma = 0.1$. For the NPV tests, the power increases when N increases and when $P(D)$ decreases. When $P(D) = 0.25$, $P(D^*) = 1 - P(D) = 0.75$, and the higher this probability is the more likely it is that a random subject does not have the disease. The number of subjects that do not have the disease are then expected to be higher than when $P(D) = 0.50$ and $P(D^*) = 0.50$. As for the PPV tests, the power increases when σ increases.

Table 6 shows the estimated test size with 95% confidence intervals in Case 1–8 for the four other difference based test statistics from Section 3.2.1. When calculating the observed value of the standardized test statistics the unrestricted maximum likelihood estimates in the variance are inserted

Case no.	$\bar{n}_1$	$\bar{n}_2$	$\bar{n}_3$	$\bar{n}_4$	$\bar{n}_5$	$\bar{n}_6$	$\bar{n}_7$	$\bar{n}_8$
1	0.2	3.9	3.9	67.0	6.3	6.2	6.2	6.3
2	1.2	19.7	19.6	334.6	31.4	31.1	31.0	31.5
3	4.3	10.3	10.2	25.1	38.4	5.4	5.5	0.8
4	21.4	51.5	51.3	125.7	191.8	27.1	27.1	4.0
5	1.1	3.1	3.1	67.8	8.3	4.2	4.2	8.4
6	5.3	15.6	15.5	338.6	41.7	20.9	20.8	41.6
7	7.5	7.0	7.0	28.3	39.8	4.1	4.0	2.2
8	37.6	35.3	35.2	141.9	198.7	20.1	20.0	11.2

TABLE 5: Mean cell counts for the cases in the LAP simulation study under H_0.

since in the LAP simulation experiment, the test size for the restricted difference test was lower than the test size for the unrestricted difference test. If we compare these results with the results for the unrestricted difference test, we see that the estimated test size depend highly on the choice of test statistic. The test based on $g_2(\boldsymbol{N})$ preserves its test size in all the cases except Case 4. It is however conservative in Case 1 and 5. The test based on $g_3(\boldsymbol{N})$ preserves its test size in all the cases, but it is conservative in all except Case 4 and 8. In Case 1 and 5 it is very conservative with an estimated test size of just 0.008 and 0.007 respectively. For the fourth difference based test statistic, $g_4(\boldsymbol{N})$, the test size is preserved in all the cases except Case 4. It is conservative in Case 1, 3 and 5. The test based on $g_5(\boldsymbol{N})$ is conservative in all the cases, and more conservative than the other tests. In Case 1 and 5 the estimated test size is 0 and 0.001 which shows that this test statistic almost never rejects the null hypothesis. The tests based on $g_2(\boldsymbol{N})$ and $g_4(\boldsymbol{N})$ can be used as their estimated test size is reasonable, although conservative in some of the cases. We do not recommend using the tests based on $g_3(\boldsymbol{N})$ and $g_5(\boldsymbol{N})$ as these are even more conservative than the other tests.

4.2 Multinomial simulation experiment

In the LAP-simulation algorithm, $n_1, \ldots, n_8$ are not drawn from a particular probability distribution, but obtained from the disease status and test results which are drawn with the specified probabilities in Section 4.1.1. However, in our model for the likelihood ratio test we assume that $N_1, ..., N_8$ are multinomially distributed. This can be used in the sampling strategy and we simulate data by sampling $n_1, ..., n_8$ from the multinomial distribution given the total number of subjects N and the parameters $p_1, ..., p_8$. This is less challenging to implement than the LAP-simulation algorithm and when using the likelihood ratio test it is natural to sample data from the distribution assumed when deriving the test statistic.

4.2.1 Algorithm

Given $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and the total number of subjects N, we can generate datasets by drawing $n_1, n_2, ..., n_8$ from a multinomial distribution with parameters $\boldsymbol{p}$ and N. We first need to set $\boldsymbol{p}$ and if we want to sample under the null hypotheses, we need to ensure that $\boldsymbol{p}$ satisfy the constraints δ_P in Equation (6) and/or δ_N in Equation (8). In addition $p_1, ..., p_8$ must sum to 1.

Our simulation algorithm is as follows:

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 $g_2(\boldsymbol{N})$	0.042	0.037	0.048
Case 1 $g_3(\boldsymbol{N})$	0.008	0.006	0.011
Case 1 $g_4(\boldsymbol{N})$	0.021	0.017	0.025
Case 1 $g_5(\boldsymbol{N})$	0.000	0.000	0.001
Case 2 $g_2(\boldsymbol{N})$	0.056	0.050	0.063
Case 2 $g_3(\boldsymbol{N})$	0.043	0.038	0.049
Case 2 $g_4(\boldsymbol{N})$	0.051	0.045	0.058
Case 2 $g_5(\boldsymbol{N})$	0.008	0.006	0.011
Case 3 $g_2(\boldsymbol{N})$	0.045	0.040	0.051
Case 3 $g_3(\boldsymbol{N})$	0.037	0.032	0.042
Case 3 $g_4(\boldsymbol{N})$	0.043	0.038	0.049
Case 3 $g_5(\boldsymbol{N})$	0.001	0.000	0.002
Case 4 $g_2(\boldsymbol{N})$	0.057	0.051	0.063
Case 4 $g_3(\boldsymbol{N})$	0.056	0.050	0.063
Case 4 $g_4(\boldsymbol{N})$	0.057	0.051	0.064
Case 4 $g_5(\boldsymbol{N})$	0.042	0.036	0.048
Case 5 $g_2(\boldsymbol{N})$	0.039	0.034	0.044
Case 5 $g_3(\boldsymbol{N})$	0.007	0.005	0.010
Case 5 $g_4(\boldsymbol{N})$	0.027	0.023	0.032
Case 5 $g_5(\boldsymbol{N})$	0.000	0.000	0.001
Case 6 $g_2(\boldsymbol{N})$	0.049	0.043	0.055
Case 6 $g_3(\boldsymbol{N})$	0.040	0.035	0.046
Case 6 $g_4(\boldsymbol{N})$	0.047	0.042	0.054
Case 6 $g_5(\boldsymbol{N})$	0.007	0.005	0.010
Case 7 $g_2(\boldsymbol{N})$	0.047	0.042	0.053
Case 7 $g_3(\boldsymbol{N})$	0.035	0.031	0.041
Case 7 $g_4(\boldsymbol{N})$	0.047	0.042	0.054
Case 7 $g_5(\boldsymbol{N})$	0.001	0.000	0.003
Case 8 $g_2(\boldsymbol{N})$	0.054	0.048	0.061
Case 8 $g_3(\boldsymbol{N})$	0.052	0.046	0.058
Case 8 $g_4(\boldsymbol{N})$	0.054	0.048	0.061
Case 8 $g_5(\boldsymbol{N})$	0.042	0.036	0.048

TABLE 6: Estimated test size with 95% confidence limits when testing PPV$_A$ $=$ PPV$_B$ using the difference based tests.

Case	N	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
Case 3MN	100	0.05	0.10	0.10	0.25	0.39	0.05	0.05	0.01
Case 5MN	100	0.01	0.03	0.03	0.68	0.08	0.04	0.04	0.09

TABLE 7: Specification of the parameters in the multinomial simulation experiment.

1. Set $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and N.
2. Draw $n_1, n_2, ..., n_8 \sim$ multinom$(\boldsymbol{p}, N)$. Repeat M times.

4.2.2 CASES UNDER STUDY

We performed a small simulation study by drawing data from a multinomial distribution. Under the null hypothesis (2) we defined two cases called Case 3MN and Case 5MN. The parameters for these cases are given in Table 7.

The parameters $p_1, ..., p_8$ for each of the cases sum to one and the δ_P-constraint (6) and δ_N-constraint (8) are both satisfied. The parameters were set in order to represent Case 3 and 5 from the LAP-simulation experiment. In both of these cases $N = 100$, while $P(D)$ is 0.5 in Case 3MN and 0.25 in Case 5MN as in Case 3 and 5 in the LAP-simulation experiment. For both Case 3MN and 5MN the PPVs are equal and approximately 0.75, the NPVs are equal and approximately 0.85. However, since the datasets in the LAP simulation experiment were not drawn from a multinomial distribution, the mean and the variance of $\boldsymbol{n}$ will not be exactly the same in Case 3MN and 5MN as in Case 3 and 5.

The parameters $p_1, ..., p_8$ were found by setting the value of $P(D)$, the values of PPV$_1$ $=$ PPV$_2$ and NPV$_1$ $=$ NPV$_2$ and by considering the mean observed values for Case 3 and 5 in Table 5. These two cases were chosen because we would like to test the multinomial sampling strategy for one case where the likelihood ratio test did not preserve its test size (Case 5) as well as one case where the test size was preserved (Case 3) in the LAP simulation experiment when testing if the positive predictive values are equal.

For each of the cases we draw $M = 5000$ samples from the multinomial distribution with parameters as given in Table 7.

4.2.3 RESULTS

The estimated test size and 95% confidence limits for the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests for the two cases in the simulation study using the multinomial simulation algorithm are given in Table 8.

In Case 3MN all the tests preserve the test size. We note that the estimated test size is lower for the restricted difference test than for the other tests. In Case 5MN only the restricted difference test and the LAP test preserve their test size.

If we compare the results to Case 3 in the LAP simulation experiment we see that $\hat{\alpha}$ is higher in Case 3MN than in Case 3 for all the tests. In Case 5MN, $\hat{\alpha}$ is higher for the likelihood ratio test and lower for the other tests compared to Case 5 in the LAP simulation experiment. The datasets in the two simulation experiments are not identical, but since they are generated with approximately the same

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 3MN LAP	0.054	0.048	0.060
Case 3MN Likelihood ratio test	0.054	0.048	0.060
Case 3MN Restricted difference test	0.052	0.046	0.058
Case 3MN Unrestricted difference test	0.054	0.048	0.061
Case 5MN LAP	0.056	0.050	0.063
Case 5MN Likelihood ratio test	0.072	0.065	0.079
Case 5MN Restricted difference test	0.050	0.044	0.057
Case 5MN Unrestricted difference test	0.064	0.058	0.071

TABLE 8: Estimated test size with 95% confidence limits for testing PPV$_1$ $=$ PPV$_2$ under the null hypothesis using the multinomial simulation algorithm.

values for PPV$_1$, PPV$_2$, NPV$_1$, NPV$_2$ and $P(D)$ we find it surprising that the estimated test size for the likelihood ratio test is higher in the multinomial simulation experiment than in the LAP simulation experiment. We would expect the likelihood ratio test to perform better, i.e. have a lower test size, on datasets that are drawn from the model on which the test statistic is based, namely the multinomial model.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 3MN LAP	0.059	0.053	0.066
Case 3MN Likelihood ratio test	0.061	0.054	0.068
Case 3MN Restricted difference test	0.052	0.046	0.058
Case 3MN Unrestricted difference test	0.060	0.054	0.067
Case 5MN LAP	0.049	0.044	0.056
Case 5MN Likelihood ratio test	0.062	0.056	0.069
Case 5MN Restricted difference test	0.051	0.045	0.057
Case 5MN Unrestricted difference test	0.049	0.044	0.056

TABLE 9: Estimated test size with 95% confidence limits for testing NPV$_1$ $=$ NPV$_2$ under the null hypothesis using the multinomial simulation algorithm.

The estimated test size with 95% confidence limits for testing if the NPVs are equal in the same cases are shown in Table 9. In Case 3MN only the restricted difference test preserves its test size, while in Case 5MN the LAP test and the unrestricted difference test also preserve their test size. The likelihood ratio test does not preserve its test size in any of these cases.

5 DATA FROM LITERATURE

We will use the dataset from Weiner, Ryan, McCabe, Kennedy, Schloss, Tristani and Fisher (1979) which is the same dataset as used in Leisenring et al. (2000) and Wang et al. (2006). There were 871 subjects of which 608 subjects had coronary artery disease (CAD) and 263 subjects did not have CAD. For all the subjects the results of clinical history (test A) and exercise stress testing (EST) (test

B) were registered. The dataset is shown in Table 10.

		Coronary artery disease -		Coronary artery disease +	
		Result of EST		Result of EST	
		+	-	+	-
Result of clinical history	+	22	44	473	81
	-	46	151	29	25

TABLE 10: Data from the coronary artery disease study.

Table 11 shows the resulting p-values for comparing the positive and negative predictive values using the LAP-test, the likelihood ratio test and the restricted and unrestricted difference test.

Test	PPV	NPV
LAP	0.3706	<0.0001
Likelihood ratio test	0.3710	<0.0001
Restricted difference test	0.3696	<0.0001
Unrestricted difference test	0.3705	<0.0001

TABLE 11: Comparison of p-values for the tests using data from the coronary artery disease study.

We see that all the tests yield the same results. We do not reject the null hypothesis that the PPVs are equal, but we reject the null hypothesis that the NPVs are equal. The estimated NPVs are 0.78 for the clinical history and 0.65 for EST. Therefore the clinical history is more likely to reflect the true disease status for subjects without CAD than without EST. Since all the cell counts in Table 10 are large, it is to be expected that the p-values are equal for all the tests, as seen in our simulation experiments.

6 ALTERNATIVE MODEL

When deriving the test statistic for comparing the positive predictive values for two tests, Leisenring et al. (2000) only consider the subjects that have at least one positive test result. The subjects that do not have any positive tests do not contribute to the test statistic, i.e. there is no information in how many subjects have two negative test results. Our multinomial setting with eight probabilities is useful because the null hypothesis for both the PPV and NPV can easily be expressed using the same model. However, for testing the equivalence of the PPVs, it is interesting to consider only using the subjects with at least one positive test result also for our likelihood ratio test as this will reduce the number of parameters and thereby reducing the dimension of the optimization problem. Similarly, for testing the equivalence of the NPVs for test A and test B, we only need to look at the subjects with at least one negative test.

This situation is illustrated in Figure 2. We still have the three main events A, B and D, but we only consider the data contained in A and/or B. The sample space is divided into six mutually exclusive events, to each of which a random variable N_i^*, $i = 1, ..., 6$, corresponds. We define N_i^* to be the number of subjects for which event i occurs and n_i^* to be the observed value of N_i^*. There are N^* subjects in total, i.e. $\sum_{i=1}^{6} N_i^* = N^*$. Let q_i be the probability that event i, $i = 1, ..., 6$, occurs. q_1 is then the probability that a subject has a positive test result for both test A and B and has the disease. $N_1^*, N_2^*, ..., N_6^*$ are multinomially distributed with parameters N^* and $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6)$ where $\sum_{i=1}^{6} q_i = 1$.

The null hypothesis that the positive predictive value for test A is equal to the positive predictive value for test B can be written

$$H_0^{P,6} : \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5} - \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6} = 0 \tag{29}$$

The likelihood ratio test statistic in this case is then

$$-2 \cdot \log\lambda(\boldsymbol{n}^*) = -2 \left(\sum_{i=1}^{6} n_i^* \cdot (\log(\tilde{q}_i) - \log(\hat{q}_i) \right) \tag{30}$$

where $\boldsymbol{n}^* = (n_1^*, n_2^*, n_3^*, n_4^*, n_5^*, n_6^*)$, $\tilde{q}_i$ is the maximum likelihood estimate for q_i under the null hypothesis (29) and $\hat{q}_i = n_i^*/N^*$ is the general maximum likelihood estimate.

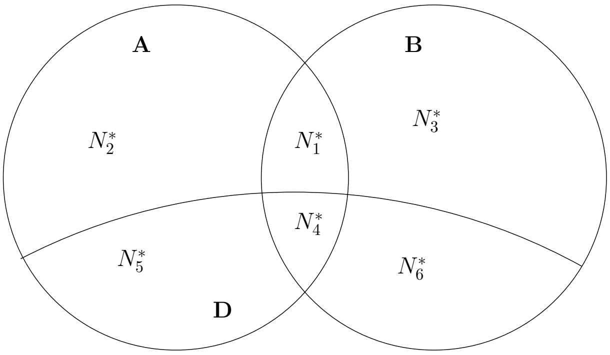


FIGURE 2: Venn diagram for the events D, A and B showing which events the random variables $N_1^*, ..., N_6^*$ correspond to.

If there is no information in the number of subjects not having at least one positive rest result, then n_4 and n_8 should not affect the value of the likelihood ratio test statistic. From the Lagrangian system of equations in Section 3.1.2, we can show that the estimates of p_4 and p_8 under H_0 are $\tilde{p}_4 = \frac{n_4}{N}$ and $\tilde{p}_8 = \frac{n_8}{N}$, see Appendix A.

Maximizing the multinomial likelihood with six parameters yields the same test statistic and thereby the same p-value as when maximizing the multinomial likelihood with eight parameters as both the restricted and unrestricted maximum likelihood estimates of q_1, q_2, q_3, q_4, q_5, q_6 are obtained from the restricted and unrestricted estimates of p_1, p_2, p_3, p_5, p_6, p_7, p_8 respectively by scaling the estimates so they sum to one (see Appendix A).

7 CONCLUSIONS

In this report we have studied large sample tests for comparing the positive and negative predictive value of two diagnostic tests in a paired design.

Based on the simulation experiments in Section 4, we have found that our restricted difference test outperforms the existing methods (Leisenring et al. (2000) and Wang et al. (2006)) as well as our likelihood ratio test with respect to test size.

A very important prerequisite of our methods is the estimation of the maximum likelihood estimates for the parameters in the multinomial distribution under the null hypothesis, and this has shown to be a challenging task as is also mentioned in Leisenring et al. (2000). We have found these estimates in two different ways, by using numerical optimization and solving a system of equations.

We have seen that when the sample size decreases, the LAP test, likelihood ratio test and unrestricted difference test do not preserve their test size. In our future work we will abandon the large sample assumption and work with small sample versions of our test statistics.

REFERENCES

Agresti, A. and Coull, B. (1998). Approximate is better than ”exact” for interval estimation of binomial proportions, *The American Statistician* 52(2): 119–126.

Alan Agresti (2002). *Categorical data analysis*, second edn, John Wiley & Sons, Inc., Hoboken, NJ, chapter 10.1.1.

Andreani, R., Birgin E. G., Martinez, J. M. and Schuverdt, M. L. (2007). On Augmented Lagrangian methods with general lower-level constraints, *SIAM Journal on Optimization* 18: 1296–1309.

Andreani, R., Birgin, E. G., Martinez, J. M. and Schuverdt, M. L. (2008). Augmented Lagrangian methods under the constant positive linear dependence constraint qualification, *Mathematical Programming* 111: 5–32.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Edwards, C. H. and Penney, D. E. (1998). *Calculus with analytic geometry*, fifth edn, Prentice-Hall International, Inc., Upper Saddle River, New Jersey, chapter 13.9.

Guggenmoos-Holzmann, I. and van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity, *Statistics in Medicine* 19: 1783–1792.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Lloyd, C. J. and Moldovan, M. V. (2008). A more powerful exact test of noninferiority from binary matched-pairs data, *Statistics in Medicine* 27(18): 3540–3549.

Moskowitz, C. S. and Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs, *Clinical Trials* 3: 272–279.

R Development Core Team (2008). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
`http://www.R-project.org`

Wang, W., Davis, C. S. and Soong, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares, *Statistics in Medicine* 25: 2215–2229.

Weiner, D. A., Ryan, T., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F.and Chaitman, B. R. and Fisher, L. (1979). Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS), *The New England Journal of Medicine* 301: 230–235.

A Proofs of properties of the maximum likelihood estimators under the null hypothesis

We show some properties of the maximum likelihood estimators under the positive predictive value constraint (6). Similar properties can be shown for maximum likelihood estimators satisfying the negative predictive value constraint (8).

We start by showing that if $n_1 = 0$, then the maximum likelihood estimate of p_1 under the null hypothesis, $\tilde{p}_1$, is 0. In the following, let $p_1, \ldots, p_8$ denote estimates, not true multinomial probabilities.

First we rewrite the constraint (11) under the null hypothesis as

$$\frac{p_1 + p_2}{p_5 + p_6} = \frac{p_1 + p_3}{p_5 + p_7} \tag{31}$$

Assume that $n_1 = 0$, $\sum_{i=1}^{8} p_i = 1$, the H_0 constraint (31) is satisfied and that $p_1 > 0$. We will prove that when $n_1 = 0$, the maximum likelihood estimate of p_1 is zero, i.e. $\tilde{p}_1 = 0$.

Let $p'_1 = 0$, $p'_2 = k(p_1 + p_2)$, $p'_3 = k(p_1 + p_3)$, $p'_4 = p_4$, $p'_5 = p_5$, $p'_6 = p_6$, $p'_7 = p_7$ and $p'_8 = p_8$ where $k = \frac{p_1+p_2+p_3}{2p_1+p_2+p_3}$.

Then $\sum_{i=1}^{8} p'_i = 1$ and $\boldsymbol{p}'$ also satisfy H_0, since

$$\frac{0 + p'_2}{p_5 + p_6} = \frac{0 + p'_3}{p_5 + p_7}.$$

We will show that $p'_2 > p_2$ and $p'_3 > p_3$, implying $\log L(p'_1, ..., p'_8) > \log L(p_1, ..., p_8)$. We start by writing down the expression for p'_2 and check if it is greater than p_2.

$$\begin{aligned} k(p_1 + p_2) &> p_2 \\ \frac{p_1 + p_2 + p_3}{2p_1 + p_2 + p_3}(p_1 + p_2) &> p_2 \\ (p_1 + p_2 + p_3) \cdot (p_1 + p_2) &> (p_1 + p_1 + p_2 + p_3)p_2 \\ (p_1 + p_2 + p_3)p_1 &> p_1 \cdot p_2 \\ p_1 + p_2 + p_3 &> p_2 \end{aligned}$$

The inequality is satisfied and therefore $p'_2 > p_2$. The same argument can be used to show that $p'_3 > p_3$.

The non-constant part of the log likelihood function is $\sum_{i=1}^{8} n_i \cdot \text{log}p_i$, and when $n_1 = 0$, the first term in the sum is 0, regardless of the value of p_1. When $p_2' > p_2$ and $p_3' > p_3$ we see that

$$\text{log}L(0, p_2', p_3', p_4, p_5, p_6, p_7, p_8) > \text{log}L(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8).$$

Therefore $\tilde{p}_1 = 0$. The same argument is valid for $\tilde{p}_5$, i.e. $\tilde{p}_5 = 0$ when $n_5 = 0$. When n_4 and/or n_8 is 0, then $\tilde{p}_4$ and/or $\tilde{p}_8$ are also 0, see below.

However, the argument does not hold for $\tilde{p}_2$, $\tilde{p}_3$, $\tilde{p}_6$ and $\tilde{p}_7$ when n_2, n_3, n_6 or n_7 is 0. Even though e.g. $\tilde{p}_2$ may sometimes be 0 when $n_2 = 0$, this is not always true. If e.g. $\tilde{p}_1 = 0$ and $\tilde{p}_3 > 0$, then $\tilde{p}_2$ cannot be equal to 0 even if $n_2 = 0$ because then the null hypothesis constraint (31) will not be satisfied. One example of this situation is the table $\boldsymbol{n} = (0, 0, 6, 0, 2, 6, 0, 0)$. The analytic solution of the Lagrangian system of equations is $\tilde{\boldsymbol{p}} = (0, 2/7, 1/7, 0, 2/7, 2/7, 0, 0)$ and we see that $\tilde{p}_2 \neq 0$ even though $n_2 = 0$.

We proceed to show that $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$. If we add the first eight Lagrangian equations in (18), we get

$$N = \gamma h(\boldsymbol{p}) + 2\kappa k(\boldsymbol{p}) = \gamma,$$

where $h(\boldsymbol{p}) = 1$ and $k(\boldsymbol{p}) = 0$ are the two constraints. Thus $\gamma = N$, and $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$ follow from (18).

So the maximum likelihood estimate $\tilde{\boldsymbol{p}}$ under the null hypothesis is among the $\boldsymbol{p} = (p_1, \ldots, p_8)$ for which $p_4 = n_4/N$ and $p_8 = n_8/N$. For such a $\boldsymbol{p}$, let $s(\boldsymbol{p}) = r \cdot (p_1, p_2, p_3, p_5, p_6, p_7)$, where $r = N/(N - n_4 - n_8)$ so that the sum of the components of $s(\boldsymbol{p})$ is 1. Let $\text{log}L$ and $\text{log}L'$ denote the log-likelihood of the original multinomial model with eight parameters and the alternative multinomial model with six parameters in Section 6. Then $\text{log}L(\boldsymbol{p}) - \text{log}L'(s(\boldsymbol{p}))$ is constant, showing that $\text{log}L(\boldsymbol{p})$ is maximal if and only if $\text{log}L'(s(\boldsymbol{p}))$ is. Furthermore, $\boldsymbol{p}$ satisfies the null hypothesis for the multinomial model with eight parameters if and only if $s(\boldsymbol{p})$ does for the multinomial model with six parameters, showing that the maximum likelihood estimates under the null hypothesis for the multinomial model with eight and six parameters are obtained from the other model by up- and downscaling, respectively.

There are also other relationships between the restricted parameter estimates that can easily be shown and used in the estimation of the parameters:

$$p_1 + p_2 + p_3 = \frac{n_1 + n_2 + n_3}{N}$$

$$p_5 + p_6 + p_7 = \frac{n_5 + n_6 + n_7}{N}$$

$$\frac{n_1}{p_1} + N = \frac{n_2}{p_2} + \frac{n_3}{p_3}$$

$$\frac{n_5}{p_5} + N = \frac{n_6}{p_6} + \frac{n_7}{p_7}$$

B Existing difference based methods

The already published difference based tests for comparing predictive values will here be described briefly.

B.1 Test by Wang

Recently Wang et al. (2006) presented two tests, one based on the difference of the PPVs and one based on the log ratio of the PPVs for testing the null hypothesis in (2). The data are assumed to be multinomially distributed.

They fit the model $\text{PPV}_A - \text{PPV}_B = \beta_1^P$ using the weighted least squares approach.[1] Testing if the positive predictive values for test A and B are equal is equivalent to testing $H_0 : \beta_1^P = 0$.

The test statistic is

$$W_1^P = \left(\sqrt{\frac{N}{\hat{\Sigma}_1^P}} (\widehat{\text{PPV}}_A - \widehat{\text{PPV}}_B) \right)^2 , \tag{32}$$

which is asymptotically χ_1^2-distributed. $\hat{\Sigma}_1^P$ is the estimated variance of $\hat{\beta}_1^P = \widehat{\text{PPV}}_A - \widehat{\text{PPV}}_B$. To compare the negative predictive values the same approach is followed by looking at the difference of the two negative predictive values. They fit the model $\text{NPV}_A - \text{NPV}_B = \beta_1^N$ and test the null hypothesis $H_0 : \beta_1^N = 0$ using the following test statistic

$$W_1^N = \left(\sqrt{\frac{N}{\hat{\Sigma}_1^N}} (\widehat{\text{NPV}}_A - \widehat{\text{NPV}}_B) \right)^2 \tag{33}$$

where $\hat{\Sigma}_1^N$ is the estimated variance of $\hat{\beta}_1^N = \widehat{\text{NPV}}_A - \widehat{\text{NPV}_B}$. W_1^N is asymptotically χ_1^2-distributed.

In the second test they consider the log ratio of the PPVs as their test statistic and fit the model $\log\frac{\text{PPV}_A}{\text{PPV}_B} = \beta_2^P$. Testing if the positive predictive values are equal is in this case equivalent to testing the null hypothesis $H_0 : \beta_2^P = 0$. The test statistic is

$$W_2^P = \left(\frac{\sqrt{N}}{\hat{\Sigma}_2^P} \log\frac{\widehat{\text{PPV}_A}}{\widehat{\text{PPV}_B}} \right)^2 \tag{34}$$

which is asymptotically χ_1^2 distributed. $\hat{\Sigma}_2^P$ is the estimated variance of $\hat{\beta}_2^P = \log\frac{\widehat{\text{PPV}_A}}{\widehat{\text{PPV}_B}}$. The same approach is followed to derive a second test for the negative predictive values by looking at the log ratio of the negative predictive values for test A and test B, the model fitted is $\log\left(\frac{\text{NPV}_A}{\text{NPV}_B}\right) = \beta_2^N$. To test if the negative predictive values are equal, the null hypothesis is $H_0 : \beta_2^N = 0$ and they use the following test statistic

$$W_2^N = \left(\frac{\sqrt{N}}{\hat{\Sigma}_2^N} \log\frac{\widehat{\text{NPV}_A}}{\widehat{\text{NPV}_B}} \right)^2 \tag{35}$$

where $\hat{\Sigma}_2^N$ is the estimated variance of $\hat{\beta}_2^N = \log\frac{\widehat{\text{NPV}_A}}{\widehat{\text{NPV}_B}}$. The test statistic in (35) is χ_1^2-distributed. They recommend using the tests based on the difference of the predictive values because it performs better than the tests based on the log ratio of the predictive values in terms of test size and power.

B.2 Test by Moskowitz and Pepe

Moskowitz and Pepe (2006) look at the relative predictive values, $\text{rPPV} = \frac{\text{PPV}_A}{\text{PPV}_B}$ and $\text{rNPV} = \frac{\text{NPV}_A}{\text{NPV}_B}$. By using the multivariate central limit theorem and the Delta method (which uses Taylor series ex-

[1] The notation in Appendix B.1 differs from the notation used in Wang et al. (2006).

pansions to derive the asymptotic variance), the following $100 \cdot (1-\alpha)\%$ confidence intervals can be estimated for log rPPV and log rNPV,

$$\log \text{rPPV} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_P^2}{N}}$$

$$\log \text{rNPV} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_N^2}{N}}$$

where $\widehat{\sigma}_P^2$ and $\widehat{\sigma}_N^2$ are the estimated variances of $\frac{1}{\sqrt{N}}\log \widehat{\text{rPPV}}$ and $\frac{1}{\sqrt{N}}\log \widehat{\text{rNPV}}$ respectively and N is the number of subjects under study. Moskowitz and Pepe (2006) do not present a hypothesis test, but based on the confidence intervals we have the asymptotically χ_1^2 distributed test statistic

$$Z_P = \left(\log\left(\frac{\sqrt{N}}{\widehat{\sigma}_P}\text{rPPV}\right)\right)^2 \tag{36}$$

for testing the null hypothesis (2) for the positive predictive values. When testing the null hypothesis (3) whether the negative predictive values are equal the test statistic

$$Z_N = \left(\log\left(\frac{\sqrt{N}}{\widehat{\sigma}_N}\text{rNPV}\right)\right)^2, \tag{37}$$

which has an asymptotic χ_1^2 distribution can be used. The test statistic in (36) only differs from the test statistic in (32) in the estimated variance. Moskowitz and Pepe (2006) use the multinomial Poisson transformation to simplify the variances.

C Results from the LAP simulation experiment

Table 12 shows the estimated test size with 95% confidence limits when comparing the negative predictive values for data generated the null hypothesis. Table 13 and 14 show the estimated test power when comparing the positive and negative predictive values respectively for data generated under the alternative hypothesis.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.052	0.046	0.059
Case 1 Likelihood ratio test	0.056	0.050	0.063
Case 1 Restricted difference test	0.048	0.043	0.055
Case 1 Unrestricted difference test	0.052	0.046	0.059
Case 2 LAP test	0.050	0.045	0.057
Case 2 Likelihood ratio test	0.050	0.044	0.057
Case 2 Restricted difference test	0.050	0.044	0.056
Case 2 Unrestricted difference test	0.050	0.045	0.057
Case 3 LAP test	0.058	0.052	0.065
Case 3 Likelihood ratio test	0.059	0.053	0.066
Case 3 Restricted difference test	0.052	0.047	0.059
Case 3 Unrestricted difference test	0.060	0.053	0.067
Case 4 LAP test	0.046	0.041	0.052
Case 4 Likelihood ratio test	0.046	0.040	0.052
Case 4 Restricted difference test 4	0.045	0.039	0.051
Case 4 Unrestricted difference test	0.047	0.041	0.053
Case 5 LAP test	0.050	0.044	0.056
Case 5 Likelihood ratio test	0.061	0.055	0.068
Case 5 Restricted difference test	0.049	0.044	0.056
Case 5 Unrestricted difference test	0.050	0.044	0.056
Case 6 LAP test	0.049	0.044	0.056
Case 6 Likelihood ratio test	0.049	0.044	0.056
Case 6 Restricted difference test	0.049	0.043	0.055
Case 6 Unrestricted difference test	0.049	0.044	0.056
Case 7 LAP test	0.060	0.053	0.067
Case 7 Likelihood ratio test	0.067	0.061	0.074
Case 7 Restricted difference test	0.056	0.050	0.063
Case 7 Unrestricted difference test	0.060	0.054	0.067
Case 8 LAP test	0.046	0.040	0.052
Case 8 Likelihood ratio test	0.047	0.041	0.053
Case 8 Restricted difference test	0.044	0.039	0.050
Case 8 Unrestricted difference test	0.046	0.040	0.052

TABLE 12: Estimated test size with 95% confidence limits when testing NPV$_A$ = NPV$_B$ for data generated under the null hypothesis using the LAP simulation algorithm.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.125	0.116	0.135
Case 1 Likelihood ratio test	0.143	0.134	0.153
Case 1 Restricted difference test	0.111	0.102	0.120
Case 1 Unrestricted difference test	0.141	0.131	0.151
Case 2 LAP test	0.396	0.383	0.410
Case 2 Likelihood ratio test	0.390	0.376	0.403
Case 2 Restricted difference test	0.380	0.367	0.394
Case 2 Unrestricted difference test	0.400	0.386	0.413
Case 3 LAP test	0.369	0.356	0.383
Case 3 Likelihood ratio test	0.361	0.348	0.375
Case 3 Restricted difference test	0.349	0.336	0.363
Case 3 Unrestricted difference test	0.369	0.356	0.383
Case 4 LAP test	0.945	0.938	0.951
Case 4 Likelihood ratio test	0.944	0.937	0.950
Case 4 Restricted difference test	0.943	0.936	0.949
Case 4 Unrestricted difference test	0.944	0.938	0.950
Case 5 LAP test	0.146	0.137	0.156
Case 5 Likelihood ratio test	0.174	0.163	0.184
Case 5 Restricted difference test	0.124	0.116	0.134
Case 5 Unrestricted difference test	0.158	0.149	0.169
Case 6 LAP test	0.463	0.450	0.477
Case 6 Likelihood ratio test	0.458	0.444	0.472
Case 6 Restricted difference test	0.449	0.435	0.463
Case 6 Unrestricted difference test	0.466	0.452	0.479
Case 7 LAP test	0.485	0.471	0.498
Case 7 Likelihood ratio test	0.484	0.470	0.498
Case 7 Restricted difference test	0.468	0.454	0.482
Case 7 Unrestricted difference test	0.485	0.471	0.498
Case 8 LAP test	0.987	0.984	0.990
Case 8 Likelihood ratio test	0.987	0.984	0.990
Case 8 Restricted difference test	0.987	0.983	0.990
Case 8 Unrestricted difference test	0.987	0.984	0.990

TABLE 13: Estimated power with 95% confidence limits when testing $\text{PPV}_A = \text{PPV}_B$ for data generated under the alternative hypothesis.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.324	0.312	0.338
Case 1 Likelihood ratio test	0.336	0.323	0.349
Case 1 Restricted difference test	0.048	0.043	0.055
Case 1 Unrestricted difference test	0.052	0.046	0.059
Case 2 LAP test	0.929	0.921	0.935
Case 2 Likelihood ratio test	0.929	0.921	0.935
Case 2 Restricted difference test	0.050	0.044	0.056
Case 2 Unrestricted difference test	0.050	0.045	0.057
Case 3 LAP test	0.113	0.105	0.122
Case 3 Likelihood ratio test	0.114	0.105	0.123
Case 3 Restricted difference test	0.052	0.047	0.059
Case 3 Unrestricted difference test	0.060	0.053	0.067
Case 4 LAP test	0.350	0.337	0.364
Case 4 Likelihood ratio test	0.349	0.336	0.362
Case 4 Restricted difference test	0.045	0.039	0.051
Case 4 Unrestricted difference test	0.047	0.041	0.053
Case 5 LAP test	0.427	0.413	0.441
Case 5 Likelihood ratio test	0.458	0.444	0.471
Case 5 Restricted difference test	0.049	0.044	0.056
Case 5 Unrestricted difference test	0.050	0.044	0.056
Case 6 LAP test	0.986	0.982	0.989
Case 6 Likelihood ratio test	0.986	0.982	0.989
Case 6 Restricted difference test	0.049	0.043	0.055
Case 6 Unrestricted difference test	0.049	0.044	0.056
Case 7 LAP test	0.136	0.127	0.146
Case 7 Likelihood ratio test	0.146	0.136	0.156
Case 7 Restricted difference test	0.056	0.050	0.063
Case 7 Unrestricted difference test	0.060	0.054	0.067
Case 8 LAP test	0.465	0.451	0.478
Case 8 Likelihood ratio test	0.463	0.450	0.477
Case 8 Restricted difference test	0.044	0.039	0.050
Case 8 Unrestricted difference test	0.046	0.040	0.052

TABLE 14: Estimated power with 95% confidence limits when testing NPV$_A$ = NPV$_B$ for data generated under the alternative hypothesis using the LAP simulation algorithm.

D Computational remarks

When using the TANGO program Andreani et al. (2007), Andreani et al. (2008), there are several parameters that can be set or modified by the user. Along with the specification of the objective function and the constraints, the initial estimates of the Lagrange multipliers, the initial values of the variables and their lower and upper bounds must be set. Other parameters have a default value, but these can be altered by the user. These parameters include tolerance limits and the maximum number of iterations.

In our simulation studies we have chosen the initial value 0.0 for all the variables with upper and lower bounds ± 200000. The initial value for the Lagrangian multiplier was set to 0.0 as advised in the program when one does not believe it should have a specific value. The feasibility and optimality tolerances are 10^{-4} by default. We found that with these tolerances, the resulting variable values depend on both the initial value of the Lagrange multiplier and the initial values of the variables. However, different initial values for the variables give more similar results than different initial values of the Lagrange multipliers. The smaller the tolerance is, the more similar the results will be, so in order to get results that do not depend on any of the initial values one should use smaller values for the tolerances and in our problems, smaller than 10^{-4}. The problem is then that it takes longer for the algorithm to converge. When performing the likelihood ratio test for one or a few datasets this is not an issue, but when performing simulation experiments with several thousand datasets this will slow down the experiment considerably.

Another problem is that of the algorithm converging to a local maximum. For example, the analytical restricted likelihood estimates for the table $\boldsymbol{n} = (0, 7, 0, 69, 5, 3, 11, 5)$ is -115.73 while it is -166.38 using the numerical estimates from TANGO. The difference is caused by the fact that $\tilde{p}_3 = 0$ using the numerical optimization routine, while it is 0.04 using the analytical optimization. For most of the large sample datasets the difference is less, with e.g. 15 of 5000 estimates in Case 1 in the LAP simulation experiment differ by more than 1.0 between the analytical and numerical estimates. We recommend using the analytical estimates.

Paper V

Comparing positive predictive values for small samples with application to gene ontology testing

Clara-Cecilie Günther, Øyvind Bakke and Mette Langaas
Department of Mathematical Sciences.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

March 2009

Summary

Motivated by the challenge of detecting Gene Ontology (GO) categories which are over-represented or depleted when comparing biological findings represented by two over-lapping lists of genes, we examine the performance of different statistical tests. One key feature with this type of data is that the sample size at each GO category often is small and thus large sample asymptotic tests are not suitable. We look at four different test statistics in combination with parametric boot-strapping, and compare the methods with their asymptotic alternatives. We find that the choice of test statistic influence which GO categories are found to be significant, and all tests under study perform increasingly conservative as the sample size decreases. We observe that this problem is statistically the same as comparing the positive predictive values of two diagnostic tests.

1 Introduction

In some biological experiments the aim is, e.g. by using DNA microarrays, to discover genes that are differentially expressed between two or more conditions. The conditions may be defined by the presence or absence of a disease or by different treatments like diets, drugs or amount of physical exercise. As an example we consider a situation where the relationship between inborn aerobic capacity and cardiac gene expression in rats was studied, Bye, Langaas, Høydahl, Kemi, Heinrich, Koch, Britton, Najjar, Ellingsen and Wisløff (2008). The rats were born with either high running capacity (HCR) or low running capacity (LCR), and half of the rats were trained, while the others remained sedentary. Thus there were four groups of rats, LCR trained, LCR sedentary, HCR trained and HCR sedentary. Several comparisons were done, and the comparison of the gene expression for the sedentary HCR rats with the gene expression for the sedentary LCR rats resulted in a list of 1540 differentially expressed genes between these two groups.

However, since such lists contains only single genes, i.e. without information about potential connections to the other genes on the list, it can be challenging to interpret the biological meaning of the results. What may be more interesting for interpretation purposes is the biological pathways that are active in the conditions under study. To do this, groups of genes instead of single genes are considered. In this paper we consider groups of genes selected from a predefined set using the Gene Ontology (GO) vocabulary, The Gene Ontology Consortium (2000). GO is a vocabulary that classifies genes into the three main categories: biological process, molecular function and cellular component and their subcategories.

Given the list of differentially expressed genes from the experiment and the list of all genes present on the microarray chip, called the master list, the biologist wants to know whether certain gene classes are over-represented or depleted in the list of differentially expressed genes compared to the master list. In the rat example, we are interested in knowing if the number of genes related to aerobic capacity among those differentially expressed between the sedentary HCR rats and the sedentary LCR rats is higher than what we would expect by chance if we compare it to the master list. The list of differentially expressed genes is contained in the master list, and the statistical hypothesis problem is to test whether two binomial proportions are equal. Common approaches are Pearson's asymptotic χ^2-test and Fisher's exact test for large and small samples, respectively.

If there are more than two conditions in the experiment, several comparisons can be done which may each result in a list of differentially expressed genes between the conditions being compared. Then we would like to see whether some specific gene classes of interest are over-represented or depleted on one of the lists compared to one of the others. The two lists may either be mutually exclusive or partly overlapping. If they are mutually exclusive the problem reduces to test whether two binomial proportions are equal as for the master list problem and the same approaches can be used. We will consider only the situation of overlapping gene lists. In the rat example, we want to compare the list of differentially expressed genes between trained HCR and LCR rats to the list of differentially expressed genes between trained HCR rats and sedentary LCR rats.

Comparing two overlapping gene lists in terms of over-represented or depleted gene classes is statistically the same situation as comparing the positive predictive values for two diagnostic tests and several hypothesis tests for this situation can be found in the literature, see Leisenring, Alonzo and Pepe (2000), Wang, Davis and Soong (2006) and Moskowitz and Pepe (2006). Günther, Bakke, Lydersen and Langaas (2008) presented a likelihood ratio test and a restricted difference test and compared them to the other existing tests. Simulation experiments showed that for smaller sample sizes these tests did not preserve their test size. When comparing gene lists, the actual sample size is the number of genes associated with each of the three main GO-categories, not the number of genes on the microarray chip, nor the total number of genes on the lists. This number is usually quite small and large sample tests are not a suitable approach. Instead small sample tests should be applied.

In this paper we evaluate small sample tests for comparing two overlapping gene lists, i.e. to test whether the probabilities that a randomly chosen gene belongs to a specific gene class are equal for the two lists. We first describe the assumed model and define the null and alternative hypotheses in Section 2, and then present the test statistics and how to calculate the p-values in Section 3. A simulation study is conducted to assess the method and described in Section 4 and in Section 5 an example in which data from the literature is given. A short discussion is given in Section 6 before we end with the conclusions in Section 7.

2 Model and data

We assume that we have two lists of genes, list A and list B. For each gene we are interested in comparing the probability that it belongs to a certain gene class D given that it is on list A, with the probability that it belongs to D given that it is on list B. Our null hypothesis is that these two probabilities are equal. By defining the three events

- D: The gene belongs to gene class D.
- A: The gene is present on gene list A.
- B: The gene is present on gene list B.

we can express the null hypothesis as

$$H_0\colon P(D \mid A) = P(D \mid B). \tag{1}$$

Statistically, this is the same problem as testing equality of the positive predictive values of two diagnostic tests for the same disease. Two diagnostic tests with binary outcomes, i.e. positive or negative, are applied to each subject in the study. The positive predictive value (PPV) is defined as the probability that the subject has the disease of study given that the test is positive. If we let event D be that the subject has the disease, A the event that the outcome of test A is positive and B the event that the outcome of test B is positive, then the positive predictive value of test A is PPV$_A = P(D \mid A)$, and the positive predictive value of test B is PPV$_B = P(D \mid B)$. Our null hypothesis is that the two positive predictive values are equal, i.e. $H_0\colon P(D \mid A) = P(D \mid B)$ as in (1).

The Venn diagram in Figure 1 shows the six mutually exclusive events defined by A, B and D. We only look at the restricted sample space, i.e. $A \cup B$, and thereby only the part of D that intersects $A \cup B$. Let A^*, B^* and D^* be the complementary events of A, B and D respectively. Günther et al. (2008) argue that when comparing positive predictive values it suffices to consider only the subjects with at least one positive test result, which equals the set $A \cup B$. In the GO setting the number of genes belonging to the GO category D that are not present on any of the lists, i.e. the event $(A^* \cup B^*) \cap D$, is unknown as is the number of genes not present on the lists that do not belong to the GO category D, therefore the part of D that intersects with $A^* \cup B^*$ is not included.

To each of the six events in the Venn diagram there corresponds the probability q_i that event i occurs, $i = 1, ..., 6$. The sum of these probabilities is one, i.e. $\sum_{i=1}^{6} q_i = 1$. Associated with each event is also a random variable N_i, $i = 1, \ldots, 6$, N_i being the number of times event i occurs. We consider one main category at a time, such that in total there are $N = \sum_{i=1}^{6} N_i$ unique genes on the two lists associated with either biological process, cellular component or molecular function. The number N will typically change between the three main categories. Given N, the random variables $N_1, N_2, \ldots, N_6$ are multinomially distributed with parameters N and
$\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6)$. The joint probability function of $N_1, N_2, \ldots, N_6$ is

$$P\left(\bigcap_{i=1}^{6}(N_i = n_i)\right) = N! \prod_{i=1}^{6} \frac{q_i^{n_i}}{n_i!}. \tag{2}$$

The expected value of $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6)$ is $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{N}) = N \cdot \boldsymbol{q}$ and the covariance matrix is $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{N}) = N(\mathrm{Diag}(\boldsymbol{q}) - \boldsymbol{q}^T\boldsymbol{q})$, Johnson, Kotz and Balakrishan (1997). We do not assume that a random gene's presence on list A is independent on its presence on list B and of whether it belongs to GO category D. This is implicitly handled by the multinomial model, since each gene yields only one observation of one of the six mutually exclusive events. We do however assume that the genes are sampled independently of each other and we will comment this further in Section 6.

Throughout this work, we assume that only N is fixed and that $\boldsymbol{N}$ are realisations of multinomial samples. Other sampling schemes are possible as well, for instance by fixing N_D, N_A and N_B, the number of genes belonging to gene class D, are present on list A and are present on list B respectively, and sampling $\boldsymbol{N}$ independently from three binomial distributions. In this report, we will not consider these approaches.

The probabilities $P(D|A)$ and $P(D|B)$ can be expressed in terms of the parameters $\boldsymbol{q}$ since

$$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5}$$

and

$$P(D|B) = \frac{P(D \cap B)}{P(B)} = \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6}.$$

Thus, the null hypothesis can be written

$$H_0 : \delta = \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5} - \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6} = 0. \tag{3}$$

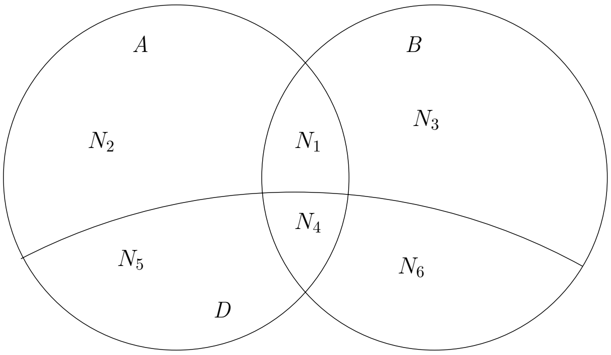


FIGURE 1: Venn diagram for the events D, A and B showing which events the random variables $N_1, \ldots, N_6$ correspond to.

There are several possible alternative hypotheses. If we are interested in whether there is an enrichment or depletion of genes belonging to gene class D on list A compared to list B, we have the two-sided alternative

$$H_1 : P(D|A) \neq P(D|B). \tag{4}$$

If we are interested in testing only whether there is an enrichment of genes belonging to gene class D on list A compared to list B, we have the one-sided alternative

$$H_1 : P(D|A) > P(D|B). \tag{5}$$

When testing whether there is a depletion of genes belonging to gene class D on list A compared to list B, the alternative hypothesis is

$$H_1 : P(D|A) < P(D|B). \tag{6}$$

In this work we will focus on the two sided alternative. We observe data $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6)$ which are realizations of $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6)$ and can be represented in a table as shown in Table 1.

Event	D^*	D
$A \cap B$	n_1	n_4
$A \cap B^*$	n_2	n_5
$A^* \cap B$	n_3	n_6

TABLE 1: The observed data classified by the events A, B and D.

3 METHOD

In this section we present the test statistics we considered to test whether the probability of a gene belonging to gene class D given that it is present on list A is equal to the probability of a gene belonging to gene class D given that it is present on list B. We also describe how to calculate the p-values.

3.1 TEST STATISTICS

To test the null hypothesis (3), we consider four test statistics: a likelihood ratio test statistic, a score test statistic and two difference test statistics. They have all been shown to be asymptotically χ^2_1 distributed when the sample size is large, Casella and Berger (2002), Leisenring et al. (2000), Wang et al. (2006), but here we will use parametric bootstrapping to approximate their distribution under the null hypothesis for small samples. We describe the test statistics briefly, more details can be found in Günther et al. (2008).

The likelihood ratio test statistic is

$$T_{\text{LR}} = -2 \cdot \log(\lambda(\boldsymbol{N}))$$

where $\lambda(\boldsymbol{N})$ is the maximum likelihood of a multinomial sample under the null hypothesis divided by the general maximum likelihood of the multinomial sample. Let $\boldsymbol{Q}$ denote the parameter space for $\boldsymbol{q}$ and $\boldsymbol{Q}_0$ the subspace of $\boldsymbol{Q}$ in which $\boldsymbol{q}$ satisfy the constraint given by the null hypothesis (3). Then,

$$\lambda(\boldsymbol{n}) = \frac{\sup_{\boldsymbol{Q}_0} L(\boldsymbol{q}|\boldsymbol{n})}{\sup_{\boldsymbol{Q}} L(\boldsymbol{q}|\boldsymbol{n})}.$$

Let $\tilde{q}_i,\ i = 1, \ldots, 6$, be the restricted maximum likelihood estimates, that is, the maximum likelihood estimates under H_0, and let $\hat{q}_i$, $i = 1, \ldots, 6$ be the unrestricted general maximum likelihood estimates for the multinomial distribution, i.e. $\hat{q}_i = n_i/N$. Inserting these estimates in the log-likelihood function for the multinomial distribution leads to the test statistic

$$T_{\text{LR}} = -2 \left(\sum_{i=1}^{6} n_i \cdot (\log \tilde{q}_i - \log \hat{q}_i) \right). \tag{7}$$

Note that $\tilde{q}_i$, $i = 1, \ldots, 6$, cannot be written in any comprehensible closed form, but can be found using an optimization routine or analytically by solving a Lagrangian system of equations. We do the latter using Maple 12, for details see Günther et al. (2008).

The difference tests are based on the estimator $g(\boldsymbol{N})$ for the difference δ in (3),

$$g(\boldsymbol{N}) = \frac{N_4 + N_5}{N_1 + N_2 + N_4 + N_5} - \frac{N_4 + N_6}{N_1 + N_3 + N_4 + N_6} \tag{8}$$

and the test statistic is derived by subtracting the expectation of $g(\boldsymbol{N})$ and dividing by its approximate standard deviation, which is found by taking the variance of the first order Taylor expansion of $g(\boldsymbol{N})$. Let $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{N})$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{N})$ as defined in Section 2. This yields

$$T_{\mathrm{g}} = \frac{(g(\boldsymbol{N}) - g(\boldsymbol{\mu}))^2}{G^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\, G(\boldsymbol{\mu})} \tag{9}$$

where G is a vector containing the first order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$ and G^T is the transpose of G. $G(\boldsymbol{\mu})$ is G with $\boldsymbol{\mu}$ inserted for $\boldsymbol{N}$.

Under the null hypothesis $g(\boldsymbol{\mu}) = 0$. $G(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ depend on the unknown parameters $\boldsymbol{q}$ which must be estimated when calculating the test statistic. We can either use the unrestricted maximum likelihood estimates $\hat{\boldsymbol{q}}$ for the multinomial distribution or the restricted maximum likelihood estimates $\tilde{\boldsymbol{q}}$ under H_0. In the first case we refer to the test as the *unrestricted* difference test (uDT) and denote the test statistic T_{uDT} and in the second case we refer to the test as the *restricted* difference test (rDT) and denote the test statistic T_{rDT}.

Leisenring et al. (2000) presented a score test, which we denote the LAP test, for testing equivalence of positive predictive values of two diagnostic tests, based on generalized estimating equations. They define three indicator variables. First D_{ij} indicates the disease status of subject i for diagnostic test j, i.e. $D_{ij} = 0$ if the subject does not have the disease and $D_{ij} = 1$ if it does have the disease. Z_{ij} indicates which test is used, it is 0 for test A and 1 for test B. X_{ij} indicates the test result, it is 0 if the test is negative and 1 if it is positive. Then the positive predictive value for test A can be written $\mathrm{PPV}_A = P(D_{ij} = 1 \mid Z_{ij} = 0, X_{ij} = 1)$ and the positive predictive value for test B is $\mathrm{PPV}_B = P(D_{ij} = 1 \mid Z_{ij} = 1, X_{ij} = 1)$. Leisenring et al. (2000) fit the generalized linear model

$$\mathrm{logit}(P(D_{ij} = 1 \mid Z_{ij}, X_{ij} = 1)) = \alpha_P + \beta_P Z_{ij},$$

and test whether $\beta_P = 0$ which is equivalent to testing whether $\mathrm{PPV}_A = \mathrm{PPV}_B$. We translate the test to the GO situation and in our notation the test statistic can be written as

$$T_{\mathrm{LAP}} = \frac{((N_1 + N_2 + N_4 + N_5)(N_4 + N_6) - (N_1 + N_3 + N_4 + N_6)(N_4 + N_5))^2}{f(N_1, N_2, N_3, N_4, N_5, N_6)}, \tag{10}$$

where

$$
\begin{aligned}
&f(N_1, N_2, N_3, N_4, N_5, N_6) \\
&\qquad = N_1(N_2 - N_3 + N_5 - N_6)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 \\
&\qquad + N_2(N_1 + N_3 + N_4 + N_6)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 \\
&\qquad + N_3(N_1 + N_2 + N_4 + N_5)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 \\
&\qquad + N_4(N_2 - N_3 + N_5 - N_6)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 \\
&\qquad + N_5(N_1 + N_3 + N_4 + N_6)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 \\
&\qquad + N_6(N_1 + N_2 + N_4 + N_5)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 .
\end{aligned}
$$

The numerator can be found from by setting the difference in (8) equal to 0 and rearranging the terms. In the denominator, the number of genes that do not belong to GO category D, N_1, N_2 and N_3, are each multiplied by the proportion of genes that belong to the category D and in this proportion, the genes that are present on both lists, N_1 and N_4, are given double weight. The number of genes that belong to GO category D, N_4, N_5 and N_6, are each multiplied by the proportion of genes that do not belong to the gene class D where the genes that are present on both lists are given double weight.

3.2 CALCULATION OF p-VALUES

We will use parametric bootstrapping to approximate the distribution of the test statistics under the null hypothesis and find approximate p-values. The test statistic of interest, is either T_{LAP}, T_{LR}, T_{uDT} or T_{rDT}. To calculate the p-values we use the following algorithm:

1. For a given sample of size N, find the maximum likelihood estimates of the parameters under H_0, $\tilde{\boldsymbol{q}}$, and calculate the test statistic t for this sample, denoted t_s.
2. Draw B samples from the multinomial distribution with parameters N and $\tilde{\boldsymbol{q}}$.
3. Calculate the test statistic t_k for each of these samples, $1 \leq k \leq B$.
4. The p-value is given as $\sum_{k=1}^{B} I(t_k \geq t_s)$, where $I(t_k \geq t_s) = \left\{ \begin{array}{ll} 1 & \text{if } t_k \geq t_s \\ 0 & \text{if } t_k < t_s \end{array} \right.$, thus the p-value is the proportion of simulated test statistics greater than or equal to the given test statistics.

4 ASSESSMENT OF METHOD

To assess the performance of the four tests in terms of test size, we perform a simulation study. The test size is the probability of making a type I error, i.e. for rejecting H_0 when H_0 is true. We consider different sample sizes to evaluate the effect of N on the test size, and we also use several parameter values of $\boldsymbol{q}$ in the multinomial distribution to explore different areas of the null hypothesis. All analysis are performed using the R language, R Development Core Team (2008), except finding the maximum likelihood estimates under H_0 which is done using Maple 12.

4.1 Simulation algorithm

Given $\boldsymbol{q}$ and N, we draw M datasets from the multinomial distribution with parameters $\boldsymbol{q}$ and N. For each of these datasets we find the p-value using parametric bootstrapping as described in Section 3.2.

4.2 Cases under study

The data sets are generated from the parameters $\boldsymbol{q}$ in the multinomial distribution and we choose six cases of parameters, given in Table 2 and depicted in Figure 2.

Case	q_1	q_2	q_3	q_4	q_5	q_6
1	0.068	0.135	0.135	0.527	0.068	0.068
2	0.043	0.130	0.130	0.348	0.174	0.174
3	0.267	0.267	0.267	0.067	0.067	0.067
4	0.300	0.267	0.267	0.033	0.067	0.067
5	0.400	0.200	0.200	0.100	0.050	0.050
6	0.450	0.200	0.200	0.050	0.050	0.050

TABLE 2: Specification of parameters in the simulation study.

The parameters in case 1 and 2 are motivated by the setting for diagnostic tests and chosen as described in the multinomial simulation experiment of Günther et al. (2008). In case 3–6, we first set the probabilities $o_1 = P(A \cap B)$, $o_2 = P(A \cap B^*)$ and $o_3 = P(A^* \cap B)$ and then $p_1 = P(D|A \cap B)$, $p_2 = P(D|A \cap B^*)$ and $p_3 = P(D|A^* \cap B)$. From these probabilities $\boldsymbol{q}$ are calculated as follows,

$$q_i = \begin{cases} o_i(1-p_i) & i = 1, 2, 3 \\ o_i p_i & i = 4, 5, 6. \end{cases}$$

In case 3 $o_1 = o_2 = o_3 = 1/3$ and $p_1 = p_2 = p_3 = 1/5$. In case 4 $o_1 = o_2 = o_3 = 1/3$ and $p_1 = 1/10$ while $p_2 = p_3 = 2/10$. The probabilities in case 5 are $o_1 = 1/2$, $o_3 = o_4 = 1/4$ and $p_1 = p_2 = p_3 = 1/5$. Finally, in case 6, $o_1 = 1/2$, $o_2 = o_3 = 1/4$, $p_1 = 1/10$ and $p_2 = p_3 = 2/10$.

The remaining parameter in the multinomial distribution, N, must also be chosen and since we are considering small sample sizes, we use $N = 10, 15, 20$ and 25. For each of the values of N all the cases given in Table 2 are run. In each of the six cases we draw $M = 10000$ samples and for each of these samples we draw $B = 10000$ bootstrap samples.

4.3 Results

The test size is estimated as the proportion of p-values being less than or equal to the chosen nominal level α. Let W be a random variable counting the number of p-values smaller than or equal to α.

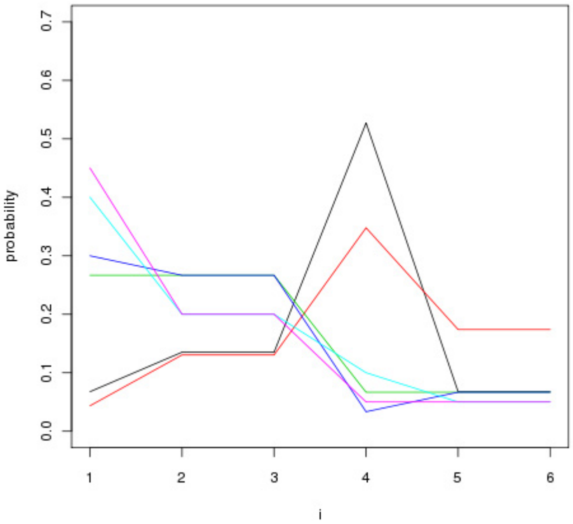


FIGURE 2: Values of q_i, $i = 1, \ldots, 6$, black: case 1, red: case 2, green: case 3, dark blue: case 4, turquoise: case 5, cyan: case 6.

Then W is binomially distributed with size M, the number of p-values generated, and probability α. The estimate of the test size of the test, $\hat{\alpha}$ is then

$$\hat{\alpha} = \frac{W}{M}. \tag{11}$$

We say that the test preserves its test size if $\hat{\alpha} \leq \alpha$. The smaller $\hat{\alpha}$ is, while less than α, the more conservative the test is. If $\hat{\alpha} > \alpha$ the test does not preserve its test size and we say that it is too optimistic. We choose $\alpha = 0.05$ and calculate $\hat{\alpha}$ for the four test statistics, six cases and four values of N.

Table 3 show the estimated test size for all the combinations of $\boldsymbol{q}$, N and test statistic. There is one table for each of the six cases. The likelihood ratio test has the largest test size in all the cases and for all values of N except for $N = 20$ and $N = 25$ in case 1 and $N = 25$ in case 2. The unrestricted difference test has the smallest test size in all the cases and for all values of N except for $N = 20$ and $N = 25$ in case 1 and $N = 25$ in case 2, which are the exceptions when the likelihood ratio test has the smallest test size. The test size of the LAP test and the unrestricted difference test lies somewhere in between, which one is the largest varies.

When $N = 10$, all the tests are conservative for all the cases, but when N increases, the test size also increases and in case 1 and 2 the tests do not preserve their test size for $N \geq 15$. This also happens in case 3 for $N \geq 20$ for the likelihood ratio test and for $N = 25$ for the restricted difference test. In case 4, 5 and 6, the tests are conservative for all values of N except the likelihood ratio test which test

(a) Case 1

N	LAP	LRT	rDT	uDT
10	0.042	0.046	0.039	0.035
15	0.059	0.061	0.060	0.057
20	0.063	0.061	0.062	0.062
25	0.055	0.052	0.054	0.055

(b) Case 2

N	LAP	LRT	rDT	uDT
10	0.038	0.043	0.040	0.018
15	0.056	0.058	0.057	0.047
20	0.056	0.057	0.057	0.054
25	0.058	0.057	0.058	0.057

(c) Case 3

N	LAP	LRT	rDT	uDT
10	0.014	0.028	0.026	0.007
15	0.029	0.044	0.041	0.024
20	0.041	0.055	0.051	0.038
25	0.047	0.056	0.053	0.046

(d) Case 4

N	LAP	LRT	rDT	uDT
10	0.012	0.023	0.021	0.004
15	0.026	0.041	0.036	0.020
20	0.035	0.047	0.044	0.029
25	0.044	0.051	0.047	0.041

(e) Case 5

N	LAP	LRT	rDT	uDT
10	0.010	0.022	0.020	0.008
15	0.021	0.037	0.033	0.020
20	0.032	0.042	0.039	0.031
25	0.040	0.049	0.046	0.040

(f) Case 6

N	LAP	LRT	rDT	uDT
10	0.007	0.013	0.012	0.004
15	0.014	0.024	0.022	0.011
20	0.029	0.037	0.034	0.026
25	0.039	0.045	0.041	0.037

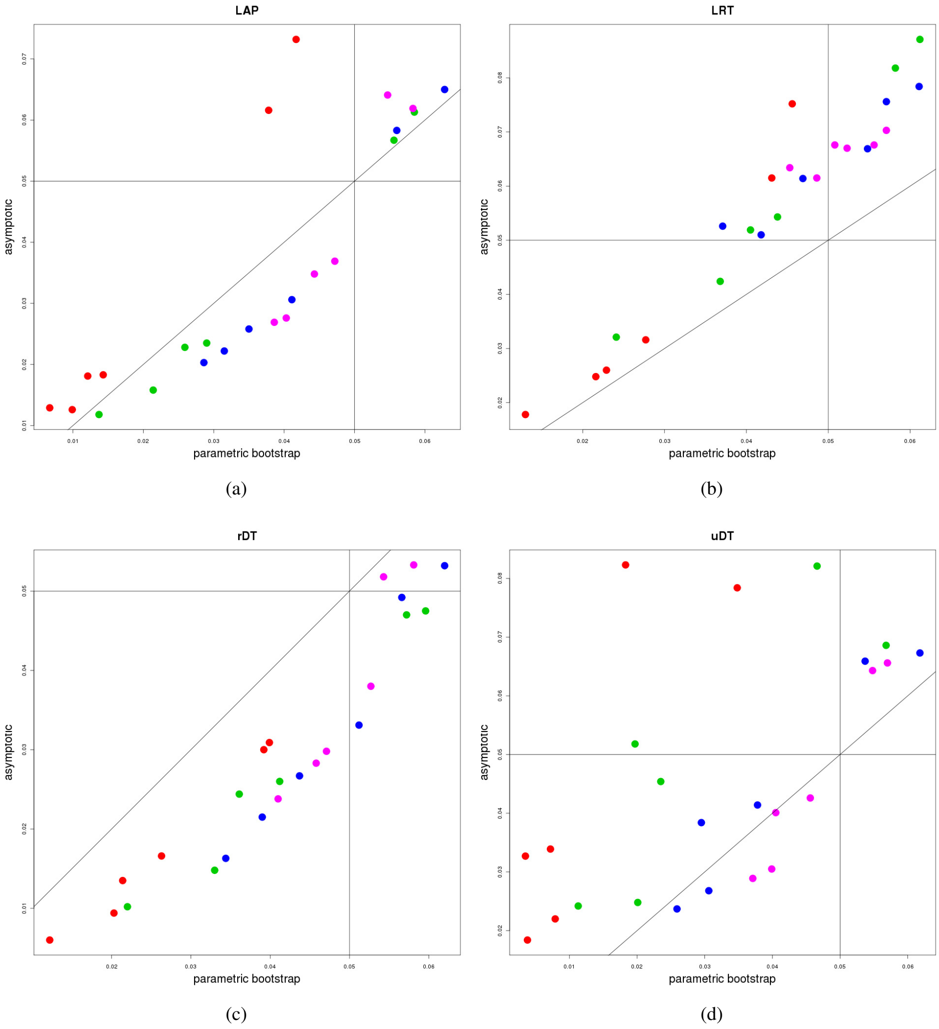
TABLE 3: Estimated test size, $\hat{\alpha}$, for $\alpha = 0.05$. LAP denotes the LAP test, LRT the likelihood ratio test and uDT and rDT denote the unrestricted and restricted difference test respectively.

size is 0.050 for $N = 25$ in case 4 and 6.

Figure 3 shows the estimated test size for the asymptotic methods plotted against the estimated test size for the parametric bootstrap methods, there is one plot for each method for $\alpha = 0.05$. If the points lie above the diagonal line, the test size of the asymptotic test is higher than the test size of the parametric bootstrap test, and lower if the points are below the line. If the points lie above the horizontal line the test size for the asymptotic test is greater than $\alpha = 0.05$ and smaller if they lie below the line. Similarly, for the points that lie to the right of the vertical line, the test size for the parametric bootstrap test is higher than 0.05 and it is lower than 0.05 if they lie to the left of this line.

We note in particular that for all the cases and for all values of N, the test size for the parametric bootstrap restricted difference test is greater than the test size for the large sample restricted difference test. For the likelihood ratio test, the opposite is true, the test size of the asymptotic likelihood ratio test is greater than the test size of the parametric bootstrap likelihood ratio test. This indicates that the parametric bootstrap test is an improvement compared to the asymptotic likelihood ratio test for small samples. However, the asymptotic likelihood ratio test does not preserve its test size in 15 of the 24 combinations of N and $\boldsymbol{q}$ and in six of those the parametric bootstrap test is still too optimistic. To use the parametric bootstrap restricted difference test does not yield an improvement compared to using the asymptotic restricted difference test.

Figure 4 shows the observed level, i.e. the test size, of the tests plotted against the nominal level for $N = 10, 15, 20, 25$ in case 3 for a chosen nominal level in the range from 0 to 0.10. We see that the test size increases when N increases and also that the tests yield more similar results for the higher values of N. The unrestricted difference test and the LAP test preserve the test size in all the cases while the likelihood ratio test is too optimistic when $N = 20$ or 25.



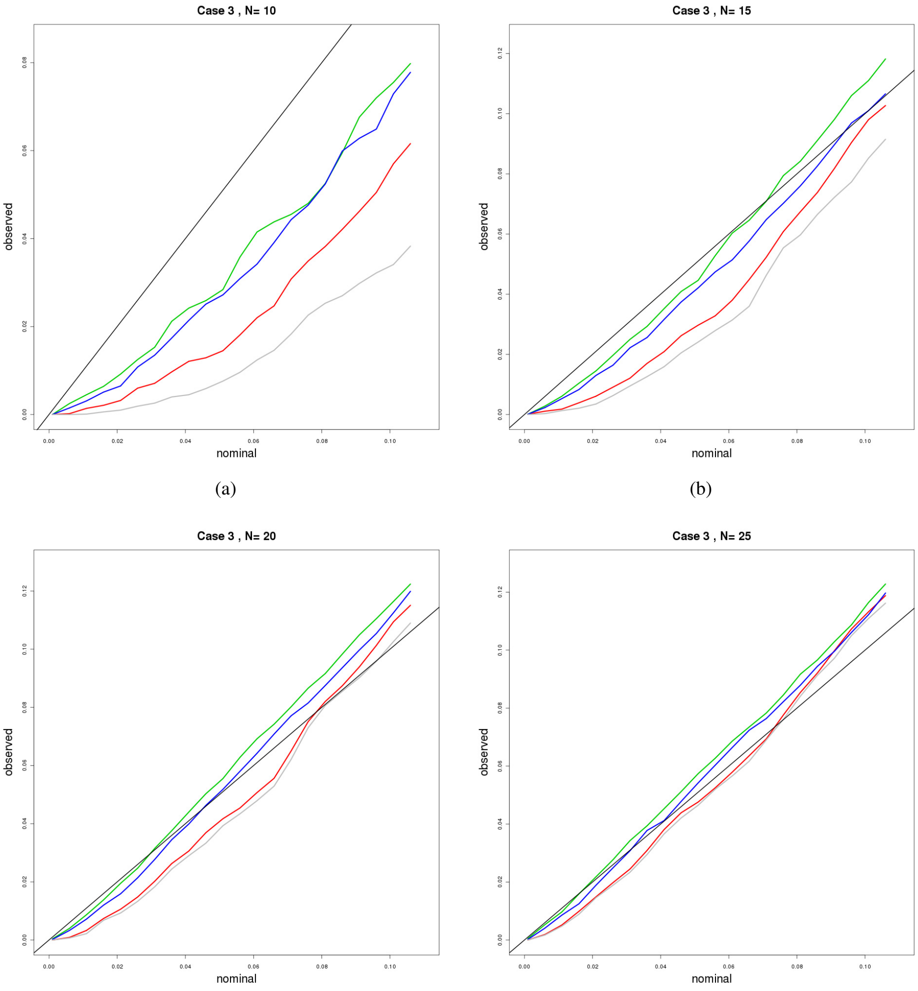
(a)

(b)

(c)

(d)

FIGURE 3: Estimated test size for the asymptotic versus the small sample tests, for different values of N: Red=10, green=15, dark blue=20, cyan=25.



(a)

(b)

(c)

(d)

FIGURE 4: Observed level versus nominal level for (a) $N = 10$, (b) $N = 15$, (c) $N = 20$, (d) $N = 25$, green = LRT, red = LAP, dark blue = rDT, grey = uDT.

5 Example from Gene Ontology

As an example of how the tests perform on a data set from literature, we use part of the data presented by Bye et al. (2008). To estimate the effect of running capacity of the trained rats, we compare the gene expression for the HCR trained rats with the LCR trained rats. This gives us a list of differentially expressed genes between these two groups, we call this list A. It may be of interest to estimate the joint effect of training and inbread running capacity by comparing the trained HCR rats versus the sedentary LCR rats. This gives us another list of differentially expressed genes which we call list B. To determine which genes are differentially expressed a cut-off must be chosen. For each gene, a p-value and an adjusted p-value are calculated. The adjusted p-values are adjusted using the Benjamini-Hochberg step-up procedure to control the false discovery rate (FDR), Benjamini and Hochberg (1995). The cut-off is chosen such that all the genes that have a p-value smaller than or equal to this value are said to be differentially expressed. We will use two different cut-offs and first we choose an FDR cut-off of 0.025 for both lists, which yields 12 genes on list A and 24 genes on list B. These lists are submitted to eGOn, Beisvåg, Jünge, Bergum, Jølsum, Lydersen, Günther, Ramampiaro, Langaas, Sandvik and Lægreid (2006), a web-based tool that automatically translates the lists to GO categories. We are interested in genes annotated to the main category molecular function. There are three genes from the first list and nine genes from the second list annotated to this category. Of these genes two are on both list A and B and therefore there are $N = 10$ unique genes on the two lists associated with molecular function.

The GO tree has several levels corresponding to the hierarchy of the GO categories. One gene can belong to more than one GO category, and given that it belongs to a subcategory it will also belong to the parent categories of this subcategory on the upper levels. After submitting the lists, one has to choose which main category to consider, i.e. either molecular function, biological process or cellular component. Level 1 is the main category itself with no subcategories, e.g. molecular function. The higher level number is chosen, the more subcategories are included, and they are all subcategories of the chosen main category. We choose to display the GO tree at level 3 for the main category molecular function and Table 4 shows the 11 GO categories that are represented on the lists, i.e. the categories which the genes on the lists belong to. A hypothesis test is performed for each category, testing whether it is over-represented or depleted on one of the lists compared to the other list.

If we use an FDR cut-off of 0.05 on differential expression instead we get two lists of 30 and 63 genes, 42 of these genes can be classified under the main category molecular function. Within this category, seven genes are present on both lists, seven genes are present only on list A and 21 genes are present only on list B, yielding $N = 35$ unique genes. Table 5 shows the GO categories for these genes along with their p-values.

Table 4 and 5 both include a column with the p-value calculated by eGOn. These p-values are calculated using the asymptotic LAP test. We calculate the p-values for the three other tests, i.e. the likelihood ratio test and the restricted and unrestricted difference test using parametric bootstrapping and compare them to the asymptotic p-values for all four tests. When performing the bootstrapping we draw $B = 10000$ bootstrap samples for each GO category.

Table 6 shows the results for the FDR cut-off of 0.025. The GO category ion binding (GO:0043167) is significant when using either the parametric bootstrap or asymptotic likelihood ratio or restricted difference test. It is also significant when using the asymptotic unrestricted difference test, while it is not significant when using the parametric bootstrap or asymptotic LAP test or the asymptotic unrestricted difference test. None of the other GO categories are significant for any of the tests.

GO identifier	Name	p-value	n_4	n_5	n_6
GO:0005488	binding	0.157	2	1	5
GO:0030246	carbohydrate binding	0.273	1	0	0
GO:0043167	ion binding	0.077	1	1	0
GO:0008289	lipid binding	0.279	0	1	0
GO:0003676	nucleic acid binding	0.317	0	0	1
GO:0005515	protein binding	0.705	2	0	5
GO:0046906	tetrapyrrole binding	0.279	0	1	0
GO:0003824	catalytic activity	0.245	1	1	2
GO:0016787	hydrolase activity	0.273	1	0	0
GO:0016874	ligase activity	0.317	0	0	1
GO:0016491	oxidoreductase activity	0.46	0	1	1

TABLE 4: GO categories within molecular function with their corresponding p-values and number of genes on the lists.

When considering only the parametric bootstrap tests, in general the likelihood ratio test and restricted difference test give similar p-values which in some cases are smaller than the p-values for the LAP test and the unrestricted difference test. One example is the GO category lipid binding (GO:0008289) where the p-values are 0.160, 0.065, 0.065 and 0.220 for the LAP, likelihood ratio, restricted difference and unrestricted difference tests respectively. Even though lipid binding is not significant for any of these tests, it is not far from being significant for the LRT and rDT tests which is not the case for the LAP and unrestricted difference tests. Together with the example ion binding, this indicates that a GO category may be declared significant more often for the likelihood ratio test and restricted difference tests than with the LAP and unrestricted difference tests. This coincide with the findings in the simulation experiments where the estimated test size in several cases were higher for the likelihood ratio and restricted difference tests than for the LAP and uDT tests.

Table 7 shows the results for the FDR cut-off of 0.05. The GO category catalytic activity (GO:0003824) is significant with a p-value <0.05 for all the tests, both the parametric bootstrap and asymptotic tests. The GO category hydrolase activity (GO:0016787) is significant when using the parametric bootstrap LRT or rDT tests and when using the asymptotic LRT, rDT and uDT tests. We see the same for the GO category substrate-specific transporter activity (GO:0022892), except that it is not significant using the asymptotic uDT test. We note that the category ion binding which is significant when we use an FDR cut-off of 0.025 is not significant now. In general, for the parametric bootstrap tests, the LRT and rDT tests yield similar p-values that are often smaller than the p-values for the LAP and uDT tests. The difference between the p-values can be quite large and for the GO term lipid binding (GO:0008289) the p-values are 0.179, 0.0345, 0.036 and 0.188 for the LAP, likelihood ratio, restricted difference and urestricted difference tests respectively. This example shows that the choice of test statistic is critical when finding GO categories that are significantly over-represented or depleted in one gene list compared to the other list. The differences between the parametric bootstrap and asymptotic tests do not follow a clear pattern, for some GO categories the parametric bootstrap p-values are smaller, for other GO categories they are greater.

The GO category chromatin binding (GO:0003682) was found to be significantly over-represented on the list of differentially expressed genes between HCR and LCR sedentary rats, using an FDR cut-off

GO identifier	Name	p-value	n_4	n_5	n_6
GO:0005488	binding	0.651	5	7	18
GO:0030246	carbohydrate binding	0.325	1	0	0
GO:0003682	chromatin binding	0.313	0	0	1
GO:0043167	ion binding	0.13	3	1	1
GO:0008289	lipid binding	0.161	1	1	0
GO:0003676	nucleic acid binding	0.485	1	1	5
GO:0000166	nucleotide binding	0.518	1	2	3
GO:0005515	protein binding	0.544	4	6	14
GO:0046906	tetrapyrrole bindin	0.308	0	1	0
GO:0003824	catalytic activity	0.016	5	5	6
GO:0016787	hydrolase activity	0.059	1	4	2
GO:0016874	ligase activity	0.56	1	0	2
GO:0016829	lyase activity	0.325	1	0	0
GO:0016491	oxidoreductase activity	0.226	2	1	1
GO:0016740	transferase activity	1	1	0	1
GO:0030234	enzyme regulator activity	0.325	1	0	0
GO:0030695	GTPase regulator activity	0.325	1	0	0
GO:0060089	molecular transducer activity	0.325	1	0	0
GO:0004871	signal transducer activity	0.325	1	0	0
GO:0005198	structural molecule activity	1	1	0	1
GO:0005201	extracellular matrix structural constituent	0.325	1	0	0
GO:0008307	structural constituent of muscle	0.313	0	0	1
GO:0030528	transcription regulator activity	0.388	1	1	1
GO:0003702	RNA polymerase II transcription factor activity	0.325	1	0	0
GO:0003700	transcription factor activity	0.388	1	1	1
GO:0016564	transcription repressor activity	0.325	1	0	0
GO:0005215	transporter activity	0.168	1	2	1
GO:0022892	substrate-specific transporter activity	0.073	1	2	0
GO:0022857	transmembrane transporter activity	0.168	1	2	1

TABLE 5: GO categories within molecular function with their corresponding p-values and number of genes on the lists.

	Parametric bootstrap				Asymptotic			
GO identifier	LAP	LRT	rDT	uDT	LAP	LRT	rDT	uDT
GO:0005488	0.198	0.315	0.344	0.204	0.157	0.249	0.354	0.109
GO:0030246	0.162	0.072	0.067	0.179	0.273	0.127	0.132	0.257
GO:0043167	0.054	0.004	0.007	0.072	0.077	0.010	0.012	0.021
GO:0008289	0.160	0.065	0.065	0.220	0.279	0.074	0.065	0.221
GO:0003676	0.383	0.371	0.401	0.446	0.317	0.433	0.540	0.289
GO:0005515	0.845	0.803	0.817	0.805	0.705	0.687	0.683	0.699
GO:0046906	0.146	0.062	0.062	0.206	0.279	0.074	0.065	0.221
GO:0003824	0.338	0.263	0.219	0.448	0.245	0.207	0.213	0.194
GO:0016787	0.164	0.073	0.065	0.181	0.273	0.127	0.132	0.257
GO:0016874	0.390	0.375	0.402	0.454	0.317	0.433	0.540	0.289
GO:0016491	0.678	0.454	0.299	0.714	0.460	0.376	0.353	0.431

TABLE 6: Parametric bootstrap and asymptotic p-values for the subcategories of molecular function.

of 0.05, compared to the list of all genes by Bye et al. (2008). Another GO category, nucleic acid binding (GO:0003676) was significantly over-represented on the list of genes that were significantly more expressed for HCR rats than for LCR rats compared to the list of genes that were significantly more expressed for LCR rats than for HCR rats, Bye et al. (2008). None of these GO-categories are over-represented in our two lists, but since we are not comparing the gene expression between sedentary HCR and LCR rats it is not surprising.

Instead of comparing the comparison of gene expression for the HCR trained rats and the LCR trained rats to the comparison of trained HCR rats and sedentary LCR rats, we could have compared the gene expression of trained LCR rats with the gene expression of the sedentary LCR rats directly. This is done in Bye et al. (2008). The list of differentially expressed genes is then submitted to eGOn and compared to the master list. However, with an FDR cut-off of 0.05 the comparison results in only one gene on the list and this gene is not annotated to any GO category.

With the first FDR cut-off of 0.025, we compared the two lists at 11 GO categories and with the second FDR cut-off at 0.05 we compared the lists at 29 GO categories. The problem thus involves multiple testing, and the p-values should be adjusted accordingly. This has not been done when comparing the methods and the p-values in Table 6 and 7 are therefore unadjusted.

6 DISCUSSION

To obtain list of differentially expressed genes, a cut-off on the differential expression must be set. The lists can then be submitted to a GO tool, e.g. eGOn, to discover GO categories that are over-represented or depleted. This approach has been criticised, see Goeman and Bühlman (2007) for an overview. Firstly, it is not clear where the cut-off should be set and secondly, one may argue that all the data should be used. Other proposed methods address this problem by either using all the p-values from the experiment or use raw expression data instead of p-values, see Goeman and Bühlman (2007).

The statistical tests in this report all treat the genes as the sampling units and are based on the assumption that the genes on the lists act independently under the null hypothesis. Statistically, it would

	Parametric bootstrap				Asymptotic			
GO identifier	LAP	LRT	rDT	uDT	LAP	LRT	rDT	uDT
GO:0005488	0.670	0.672	0.676	0.667	0.651	0.652	0.657	0.649
GO:0030246	0.369	0.180	0.171	0.371	0.325	0.240	0.308	0.326
GO:0003682	0.382	0.393	0.407	0.403	0.313	0.363	0.472	0.309
GO:0043167	0.124	0.113	0.075	0.128	0.130	0.095	0.091	0.127
GO:0008289	0.179	0.034	0.036	0.188	0.161	0.053	0.075	0.152
GO:0003676	0.511	0.528	0.535	0.512	0.485	0.500	0.510	0.483
GO:0000166	0.544	0.528	0.516	0.544	0.518	0.498	0.491	0.515
GO:0005515	0.561	0.564	0.568	0.559	0.544	0.547	0.552	0.541
GO:0046906	0.182	0.091	0.082	0.190	0.308	0.132	0.151	0.299
GO:0003824	0.017	0.017	0.016	0.020	0.016	0.015	0.016	0.011
GO:0016787	0.093	0.043	0.022	0.094	0.059	0.029	0.027	0.046
GO:0016874	0.606	0.612	0.609	0.606	0.56	0.559	0.568	0.557
GO:0016829	0.378	0.175	0.167	0.38	0.325	0.240	0.308	0.326
GO:0016491	0.262	0.227	0.166	0.262	0.226	0.180	0.175	0.222
GO:0016740	1.000	0.871	1.000	1.000	1.000	1.000	1.000	1.000
GO:0030234	0.372	0.182	0.172	0.375	0.325	0.240	0.308	0.326
GO:0030695	0.368	0.18	0.171	0.371	0.325	0.240	0.308	0.326
GO:0060089	0.374	0.178	0.168	0.378	0.325	0.240	0.308	0.326
GO:0004871	0.371	0.176	0.166	0.373	0.325	0.240	0.308	0.326
GO:0005198	1.000	0.876	1.000	1.000	1.000	1.000	1.000	1.000
GO:0005201	0.366	0.173	0.164	0.37	0.325	0.240	0.308	0.326
GO:0008307	0.377	0.382	0.393	0.394	0.313	0.363	0.472	0.309
GO:0030528	0.501	0.422	0.349	0.501	0.388	0.340	0.33	0.384
GO:0003702	0.364	0.180	0.168	0.366	0.325	0.240	0.308	0.326
GO:0003700	0.494	0.418	0.346	0.493	0.388	0.340	0.33	0.384
GO:0016564	0.373	0.180	0.169	0.376	0.325	0.240	0.308	0.326
GO:0005215	0.236	0.134	0.079	0.237	0.168	0.107	0.101	0.157
GO:0022892	0.064	0.008	0.009	0.055	0.073	0.012	0.019	0.062
GO:0022857	0.234	0.132	0.081	0.234	0.168	0.107	0.101	0.157

TABLE 7: Parametric bootstrap and asymptotic p-values for the subcategories of molecular function.

be more intuitive to use the subjects as the sampling units, as discussed by Goeman and Bühlman (2007). Indeed, when testing for equality of the positive predictive values of two diagnostic tests, the observational unit is the individual and the assumption of independence of test results between individuals is in most cases not seen to be problematic. But in the gene class setting the assumption does not hold, because genes act together in pathways and genes that are functionally related can be strongly correlated. If the gene expression measurements are correlated, the p-values tend to be positively correlated, see Goeman and Bühlman (2007). A possible extension of the methods developed in this report could be to look at different dependence structures between the observational units.

We have considered test statistics designed for comparing positive predictive values for diagnostic tests which translates to comparing association with GO categories for overlapping gene lists. Other possible approaches to handle overlapping gene lists include deleting the genes that are on both lists from each list or simply ignore the fact that there are genes that are on both lists and treat them as mutually exclusive lists. The deletion approach is implemented in the GO tool FatiGO, Al Shahrour, Diaz Uriarte and Dopazo (2004), in which Fisher's exact test is implemented. In the ignore approach Fisher's exact test or Pearson χ^2 test can be used. The asymptotic LAP test is implemented in eGOn which then handles the problem of overlapping gene lists more correctly than other GO-tools by not deleting the genes that are on both lists or ignore that the genes are overlapping.

In the simulation experiments in Section 4 and the example using data from the literature in Section 5, we see that the likelihood ratio and restricted difference tests yields similar results which differ from the LAP and unrestricted difference tests. The likelihood ratio and restricted difference test both use the maximum likelihood estimates for the parameters under the null hypothesis in addition to the general maximum likelihood estimates. The LAP and unrestricted difference also yield similar test results and these test statistics are functions of the observed data $\boldsymbol{n}$ and thereby the general maximum likelihood estimates only, and are thus not influenced by the maximum likelihood estimates under the null.

When considering small sample sizes, one or more of the cells in Table 1 have often zero counts which leads to non-computable test statistics for the LAP and difference tests. In these cases, we set the test statistics to 0 if the numerator is 0, implying that the null hypothesis will never be rejected for such tables. If only the denominator is 0, the test statistic is disregarded. For $N = 10$ there are 18 out of 3003 possible tables for which this will happen, while if $N = 15$ it will happen for 28 out of 15504 tables, for $N = 20$ for 38 of 53130 tables and for $N = 25$ for 48 of 142506 outcomes. For the likelihood ratio test statistic zero counts does not represent a problem, it can always be calculated. If any of the counts are 0, the summation term in (7) is also 0.

While the methodology in the present paper does not rely on asymptotic results, it is still approximative in the sense that it relies on simulations. Another shortcoming is that the test size is not preserved in general. Both these issues will be addressed in a forthcoming paper, Günther, Bakke, Rue and Langaas (2009), where enumeration rather than simulation will be applied to the testing method of the present paper and where it will be modified to yield p-values that preserve the test size. The test size and power will be calculated exactly.

7 Conclusions

In this report we look at the problem of testing the null hypothesis given in (3) when the sample size is small. The large sample tests using asymptotic distributions do not preserve their test size in this case and therefore small sample tests are needed. We suggest using parametric bootstrapping to approximate the distribution and to calculate the p-values. The likelihood ratio test and the restricted difference test are both functions of the maximum likelihood estimates $\boldsymbol{q}$ under the null hypothesis which may be difficult to find because of local optima. Especially zero counts causes problems, but our method that analytically solves the system of equations handles these problems well.

The simulation experiments show, at least based on the present six cases, that the small sample likelihood ratio test yields a smaller test size than the large sample likelihood ratio test, while for the restricted difference test the large sample test yields the smallest test size and is still conservative, thus for this test there was no improvement.

For testing whether there is a difference in enrichment or depletion of genes belonging to a certain GO category between two list of genes from a microarray experiment, there are several test statistics to choose from, and depending on the sample size, one can use either parametric bootstrapping or the asymptotic χ^2_1 distribution to calculate p-values. The choice of test statistic can influence which GO categories that are found to be significant and because the small sample parametric bootstrap likelihood ratio and restricted difference tests are more optimistic than the small sample parametric bootstrap LAP and unrestricted difference tests, the first two will yield more significant GO categories than the other two. The smaller the sample size is, the more conservative all the tests are which means they will not reject the null hypothesis even when it is not true., i.e. the tests will not discover gene classes that are over-represented or depleted on one list compared to the other list. Therefore, parametric bootstrapping does not seem to be an optimal solution and a better approach would probably be to use an exact small sample test that preserves its test size without being conservative, which will be investigated further.

REFERENCES

Al Shahrour, F., Diaz Uriarte, R. and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, *Bioinformatics* 20(4): 578–580.

Beisvåg, V., Jünge, F. K., Bergum, H., Jølsum, L., Lydersen, S., Günther, C.-C., Ramampiaro, H., Langaas, M., Sandvik, A. K. and Lægreid, A. (2006). GeneTools - application for functional annotation and statistical hypothesis testing, *BMC Bioinformatics* 7(470).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society* 57: 289–300.

Bye, A., Langaas, M., Høydahl, M. A., Kemi, O. J., Heinrich, G., Koch, L. G., Britton, S. L., Najjar, S. M., Ellingsen, Ø. and Wisløff, U. (2008). Aerobic capacity-dependent differences in cardiac gene expression., *Physiol Genomics* 33: 100–109.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Goeman, J. J. and Bühlman, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues., *Bioinformatics* 23(8): 980–987.

Günther, C.-C., Bakke, Ø., Lydersen, S. and Langaas, M. (2008). Comparison of predictive values from two diagnostic tests in large samples. Preprint Statistics No. 9, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Günther, C.-C., Bakke, Ø., Rue, H. and Langaas, M. (2009). Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters. Preprint Statistics No. 4, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Moskowitz, C. S. and Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs, *Clinical Trials* 3: 272–279.

R Development Core Team (2008). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
`http://www.r-project.org`

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology, *Nature Genetics* 25: 25–29.

Wang, W., Davis, C. S. and Soong, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares, *Statistics in Medicine* 25: 2215–2229.

Paper VI

STATISTICAL HYPOTHESIS TESTING FOR CATEGORICAL DATA USING ENUMERATION IN THE PRESENCE OF NUISANCE PARAMETERS

CLARA-CECILIE GÜNTHER, ØYVIND BAKKE, HÅVARD RUE AND METTE LANGAAS
Department of Mathematical Sciences.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

MARCH 2009

SUMMARY

The existing asymptotic tests for comparing positive predictive values of two diagnostic tests do not preserve the test size when the sample is small. As an exact approach we suggest using enumeration for small sample spaces, i.e. to utilize the exact distribution of the test statistic by adding probabilities of each outcome. In the problem of comparing positive predictive values, there are nuisance parameters present which must be handled. We discuss different solutions, e.g. estimation, maximization, integration and combinations thereof. The methods presented in this report are general and can be applied to different discrete finite distributions. Further insight into the mechanisms behind the different approaches are given and the performance of various test statistics and p-values are compared systematically with respect to test size and power, both in the setting of positive predictive values and in an example from literature comparing independent binomial proportions. We find in general that a combination of estimation and maximization yields the highest test size and power among the valid p-values, and when comparing the positive predictive values, the test statistics involving maximum likelihood estimates under the null hypothesis perform the best in terms of test size and power.

1 INTRODUCTION

In many hypothesis testing problems, tests statistics with a known asymptotic distribution are available. When the sample size is small, however, the asymptotic distribution may approximate the exact distribution poorly and the exact distribution of the test statistics can be challenging or impossible to derive. For discrete models, one solution is to use enumeration, i.e. to find p-values by adding probabilities under the null hypothesis of all possible outcomes having a more extreme value of the test statistic than the observed outcome. If there are nuisance parameters in the model, this is however not straight forward, the unknown parameters must be handled appropriately.

We consider different approaches, in particular estimation, maximization and integration. Our main focus will be on the problem of comparing positive predictive values from two diagnostic tests where a multinomial distribution is assumed, but the methods are general and can be applied to other null hypotheses for other finite discrete distributions.

We start by defining important properties for p-values and different ways to handle the problem of nuisance parameters in Section 2. A trinomial situation is used as an example to explain how to calculate the various p-values. As a stepping stone to our main problem, comparing positive predictive values, in Section 3 we go through a fictitious example discussed and analyzed by Berger and Boos (1994) and by Lloyd (2008) that concerns testing independence in a 2×2 contingency table. We suggest alternative test statistics and compare their performance in terms of test size and power to the test statistics used by Lloyd (2008). In Section 4 we present the problem of comparing positive predictive values for two diagnostic tests, and evaluate a variety of test statistics and p-values for this problem. Some computational details are given in Section 5, we discuss further aspects of the presented problems in Section 6 and summarize the conclusions in Section 7.

2 Theory

Before applying the methods, the general framework should be set. We present the necessary notation, definitions and properties of p-values and explain different approaches on how to calculate p-values by enumeration in the presence of nuisance parameters.

2.1 Null hypothesis

In the general outline we assume that the random variables $Y_1, \ldots, Y_n$ are multinomially distributed with parameters $\boldsymbol{p} = (p_1, \ldots, p_n)$ and N, but other discrete distributions are possible (see e.g. Section 3). Let $\boldsymbol{Y}$ denote the vector of the random variables, i.e. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, and let $\mathcal{Y}$ be the sample space or reference set of $\boldsymbol{Y}$.

Our null hypothesis is that a function f of some or all the parameters $p_1, \ldots, p_n$ equals 0, i.e.

$$H_0 : \; f(\boldsymbol{p}) = 0. \tag{1}$$

The alternative hypothesis is

$$H_1 : \; f(\boldsymbol{p}) \neq 0.$$

Let $\mathcal{P}$ be the parameter space for $\boldsymbol{p}$ and $\mathcal{P}_0$ the subspace of $\mathcal{P}$ for which the null hypothesis (1) is satisfied, i.e $\mathcal{P}_0 = \{\boldsymbol{p} : f(\boldsymbol{p}) = 0\}$. For illustrative purposes, an example from the trinomial distribution will be studied throughout this section.

Trinomial example As an illustrative example we will use the trinomial model where $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ are multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, p_3)$ and N, or alternatively $\boldsymbol{Y} = (Y_1, Y_2, N - Y_1 - Y_2)$, are multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, 1 - p_1 - p_2)$. The joint probability function of $\boldsymbol{Y}$ is

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} (1 - p_1 - p_2)^{N - y_1 - y_2}.$$

We consider the null hypothesis,

$$H_0 : \; f(\boldsymbol{p}) = p_1 - p_2 = 0, \tag{2}$$

that is $\mathcal{P}_0 = \{(\phi, \phi, 1-2\phi) : 0 \leq \phi \leq 1/2\}$. So $p_1 = p_2 = \phi$ under the null hypothesis, which can be considered an unknown nuisance parameter. The probability function of $\boldsymbol{Y}$ simplifies to

$$\text{Pr}(Y_1 = y_1, Y_2 = y_2) = \frac{N!}{y_1!y_2!(N-y_1-y_2)!}\phi^{y_1+y_2}(1-2\phi)^{N-y_1-y_2} \tag{3}$$

under the null hypothesis. □

2.2 Properties of p-values

When testing whether a null hypothesis is true, one usually calculates a p-value and if this p-value is less than or equal to some chosen significance level α the null hypothesis is rejected.

A p-value may initially be defined as the probability of what has been observed or something more extreme, given that the null hypothesis is true. A p-value can also be considered a test statistic in its own right. We let $P(\boldsymbol{Y})$ denote our p-value statistic which is a function of the random variables $\boldsymbol{Y}$. For continuous models without nuisance parameters and for simple null hypotheses, i.e. when the parameter space under H_0 consists of only one point, the p-values are uniformly distributed under the null hypothesis and the test size of a test that rejects H_0 when $P(\boldsymbol{Y}) \leq \alpha$ is exactly equal to α, Bickel and Doksum (2001). Our sample space is discrete which means that not all p-values can possibly be obtained. Instead, is is usually demanded that the p-value is *valid*, i.e. the probability of rejecting the null hypothesis when it is true is less than or equal to the significance level α,

$$\text{Pr}(P(\boldsymbol{Y}) \leq \alpha; \boldsymbol{p}) \leq \alpha$$

for all $\boldsymbol{p}$ in $\mathcal{P}_0$ and all α, $0 \leq \alpha \leq 1$, Casella and Berger (2002). The valid p-values yield a valid test for any chosen significance level, although they are often conservative. If a p-value satisfy

$$\text{sup}_{\boldsymbol{p}\in\mathcal{P}_0}\text{Pr}(P(\boldsymbol{Y}) \leq P(\boldsymbol{y}); \boldsymbol{p}) = P(\boldsymbol{y}),$$

for all $\boldsymbol{y}$ in $\mathcal{Y}$, then Lloyd (2008) call it *exact*.

In general, p-values are found by means of a test statistic $T(\boldsymbol{Y})$ having the property that for all $\boldsymbol{y}$ in $\mathcal{Y}$ and for all $\boldsymbol{p}$ in $\mathcal{P}_0$, $\text{Pr}(P(\boldsymbol{Y}) \leq P(\boldsymbol{y}); \boldsymbol{p}) = \text{Pr}(T(\boldsymbol{Y} \geq T(\boldsymbol{y}); \boldsymbol{p})$, assuming without loss of generality that the null hypothesis is rejected for larges values of $T(\boldsymbol{y})$. We define the tail set of an outcome $\boldsymbol{y}_{\text{obs}}$ to be the set of all $\boldsymbol{y}$ for which $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$, i.e. the critical region for a significance level given $T(\boldsymbol{y}_{\text{obs}})$ as a critical value. For an observed outcome $\boldsymbol{y}_{\text{obs}}$, the reference set $\mathcal{Y}$ can be partitioned into the tail set $R(\boldsymbol{y}_{\text{obs}})$ of the observed outcome and the complement of the tail set $R^C(\boldsymbol{y}_{\text{obs}})$, so that $\mathcal{Y} = R \cup R^C$ where $R(\boldsymbol{y}_{\text{obs}}) = \{\boldsymbol{y} : T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})\}$ and $R^C(\boldsymbol{y}_{\text{obs}}) = \{\boldsymbol{y} : T(\boldsymbol{y}) < T(\boldsymbol{y}_{\text{obs}})\}$.

Trinomial example We set $N = 3$, and then the reference set is $\mathcal{Y} = \{(0,0,3),(0,1,2),(0,2,1),(0,3,0),(1,0,2),(1,1,1),(1,2,0),(2,0,1),(2,1,0),(3,0,0)\}$. One possible test statistic is

$$T(\boldsymbol{Y}) = |Y_1/N - Y_2/N|. \tag{4}$$

Table 1 shows the calculated test statistic for all the outcomes in the reference set. For example, $T(0,2,1) = 2/3$ and $R(0,2,1) = \{(0,2,1),(0,3,0),(2,0,1),(3,0,0)\}$. □

The test statistic $T(\boldsymbol{Y})$ used to define the tail set can be an ordinary test statistic like the likelihood ratio test statistic, a p-value originating from another test statistic, or even the multinomial probabilities of the outcomes themselves. If the tail sets are defined by the probabilities of the outcomes, they will

Outcome $\boldsymbol{y}$	y_1/N	y_2/N	$T(\boldsymbol{y})$
(0,0,3)	0	0	0
(0,1,2)	0	1/3	1/3
(0,2,1)	0	2/3	2/3
(0,3,0)	0	1	1
(1,0,2)	1/3	0	1/3
(1,1,1)	1/3	1/3	0
(1,2,0)	1/3	2/3	1/3
(2,0,1)	2/3	0	2/3
(2,1,0)	2/3	1/3	1/3
(3,0,0)	1	0	1

TABLE 1: The reference set in the trinomial example with associated test statistic, $T(\boldsymbol{y})$ given in (4).

depend on $\boldsymbol{p}$. This is not so if the tail sets are defined by either a p-value or some test statistic that does not depend on $\boldsymbol{p}$. A practical detail, when the multinomial probabilities or a p-value are used as the test statistic, actually the negative of the probabilities and the p-values will be applied since only the outcomes with probabilities or p-values smaller than or equal to the probability or p-value of the observed outcome will be in the tail set.

2.3 CALCULATING p-VALUES BY ENUMERATION

Let $\pi(\boldsymbol{y};\boldsymbol{p}) = \Pr(\boldsymbol{Y} = \boldsymbol{y};\boldsymbol{p})$ be the probability of an outcome $\boldsymbol{y}$. If $\pi(\boldsymbol{y};\boldsymbol{p})$ is known, the p-value for the observed outcome can be calculated using the following algorithm which is motivated by Fisher's exact test for 2×2 tables, Fisher (1935):

1. Generate all possible outcomes in the reference set $\mathcal{Y}$.
2. Calculate the probability of observing each outcome under the null hypothesis.
3. The p-value of an observed outcome is the sum of the probabilities of all outcomes that are in the tail set of the observed outcome.

Zelterman, Chan and Mielke (1995) tested mutual independence of all the three factors of a 2^3 contingency table using a multinomial distribution with eight parameters. Any outcome given N will then correspond to a specific table where the entries sum to N and the reference set $\mathcal{Y}$ will be all possible tables with grand total N. By conditioning on the set of one-way marginal totals, $\boldsymbol{M}$, the probability $\pi(\boldsymbol{y}|\boldsymbol{M})$ under H_0 can be derived. It does not depend on nuisance parameters, and therefore the second step in the algorithm is easily performed once the tables are generated.

With other null hypotheses it might be impossible to get rid of the nuisance parameters and conditioning only reduces the number of possible outcomes or the number of nuisance parameters. In this case, we must find a way to deal with the (remaining) nuisance parameters to be able to calculate the probability of each outcome. There are several ways to do this.

ESTIMATION The simplest approach to deal with nuisance parameters is to insert e.g. the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ under H_0 for $\boldsymbol{p}$. This is called the plug-in p-value by Bayarri and Berger (2000) and the estimation (E) p-value by Lloyd (2008). For an observed outcome $\boldsymbol{y}_{\text{obs}}$ we insert $\tilde{\boldsymbol{p}}_{\text{obs}}$ for $\boldsymbol{p}$ and the p-value is given as

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y}; \tilde{\boldsymbol{p}}_{\text{obs}}).$$

This p-value, however, is not valid as we will see numerically in Section 4.2.

Trinomial example Under H_0, given the outcome $\boldsymbol{y}_{\text{obs}} = (y_{1,\text{obs}}, y_{2,\text{obs}}, y_{3,\text{obs}})$ the maximum likelihood estimate of ϕ is $\widetilde{\phi}_{\text{obs}} = \frac{y_{1,\text{obs}}+y_{2,\text{obs}}}{2N}$. If we insert this estimate in the multinomial probability function, we obtain the estimation p-value

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y}; \widetilde{\phi}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \frac{N!}{y_1!y_2!(N-y_1-y_2)!} \widetilde{\phi}_{\text{obs}}^{y_1+y_2} (1-2\widetilde{\phi}_{\text{obs}})^{N-y_1-y_2}.$$

The third column of Table 2 shows the estimation p-values for all outcomes in the reference set when $N = 3$. To explain how the p-values are calculated, we consider the outcome $\boldsymbol{y} = (0,2,1)$. The maximum likelihood estimate under H_0 is $\widetilde{\phi} = 1/3$. We then calculate the probability for each outcome from (3) with $\widetilde{\phi}$ inserted for ϕ. The tail set consists of the four outcomes $\boldsymbol{y}$ of Table 1 for which $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$, where $T(\boldsymbol{y})$ is given in (4), and the estimation p-value of $\boldsymbol{y}_{\text{obs}}$ is the sum 0.30, of the four probabilities. □

Outcome $\boldsymbol{y}$	$\widetilde{\phi}$	$P_{\text{E}}(\boldsymbol{y})$
(0,0,3)	0	1.00
(0,1,2)	1/6	0.56
(0,2,1)	1/3	0.30
(0,3,0)	1/2	0.25
(1,0,2)	1/6	0.56
(1,1,1)	1/3	1.00
(1,2,0)	1/2	1.00
(2,0,1)	1/3	0.30
(2,1,0)	1/2	1.00
(3,0,0)	1/2	0.25

TABLE 2: P-values for the trinomial example when substituting $\widetilde{\phi}$ for ϕ.

CONDITIONING ON A SUFFICIENT STATISTIC Another solution to the problem of nuisance parameters is to condition on a sufficient statistic X for $\boldsymbol{p}$, Casella and Berger (2002), then the probability of the observed outcome given H_0 and the sufficient statistic can be calculated and the p-value is given by

$$P_{\text{suff}}(\boldsymbol{y}_{\text{obs}}) = \Pr(T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}}) \mid X; \boldsymbol{p} \in \mathcal{P}_0) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y} \mid X; \boldsymbol{p} \in \mathcal{P}_0).$$

Trinomial example Under H_0, $X = Y_1 + Y_2$ is a sufficient statistic for ϕ. The conditional probability distribution of (Y_1, Y_2) given X is

$$\Pr(Y_1 = y, Y_2 = y \mid X = x) = \frac{x!}{y_1!y_2!} \left(\frac{1}{2}\right)^x$$

and the p-value is then the sum of these probabilities over the outcomes in the tail set,

$$P_{\text{suff}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\boldsymbol{y}\in R(\boldsymbol{y}_{\text{obs}})} \frac{x!}{y_1!y_2!}\left(\frac{1}{2}\right)^x.$$

The p-value for outcome $\boldsymbol{y}_{\text{obs}} = (0,2,1)$ is found by considering only the outcomes with $X = 2$. They are (0,2,1), (1,1,1) and (2,0,1). Looking back at Table 1, we see that $T(0,2,1) = 2/3,\ T(1,1,1) = 0$ and $T(2,0,1) = 2/3$. The p-value for $(0,2,1)$ is then the sum of the probabilities $\text{Pr}(\boldsymbol{Y} = \boldsymbol{y}|X = 2)$ for outcome $\boldsymbol{y} = (0,2,1)$ and $\boldsymbol{y} = (2,0,1)$ which are both 0.25, so the p-value is 0.50. The conditional probabilities and p-values for all the outcomes are given in Table 3. □

Outcome $\boldsymbol{y}$	x	$P(\boldsymbol{Y} = \boldsymbol{y}; X = x)$	P_{suff}
(0,0,3)	0	1	1
(0,1,2)	1	0.5	1
(0,2,1)	2	0.25	0.5
(0,3,0)	3	0.125	0.25
(1,0,2)	1	0.5	0.25
(1,1,1)	2	0.5	1
(1,2,0)	3	0.375	1
(2,0,1)	2	0.25	0.5
(2,1,0)	3	0.375	1
(3,0,0)	3	0.125	0.25

TABLE 3: P-values obtained for the trinomial example by conditioning on the sufficient statistic $X = Y_1 + Y_2$.

However, an appropriate sufficient statistic does not always exist. Instead of conditioning on a sufficient statistic, we may condition on an ancillary statistic, Berger and Boos (1994). We will not pursue this approach here.

FULL MAXIMIZATION Another approach to deal with nuisance parameters is to maximize over the set of unknown parameters, Casella and Berger (2002). In this approach, called full maximization by Lloyd (2008), the p-value is calculated as the supremum of the probability of the tail set over the parameter space of $\boldsymbol{p}$ under H_0, i.e. over $\mathcal{P}_0$. This p-value is valid and exact (as we will explain later in this section) and is given as

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \text{sup}_{\boldsymbol{p}\in\mathcal{P}_0}\text{Pr}(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \boldsymbol{p}) = \text{sup}_{\boldsymbol{p}\in\mathcal{P}_0} \sum_{\boldsymbol{y}\in R(\boldsymbol{y}_{\text{obs}};\boldsymbol{p})} \pi(\boldsymbol{y};\boldsymbol{p}).$$

Trinomial example For each outcome we calculate the full maximization p-value by maximizing the sum of multinomial probabilities for the outcomes in the tail set over all values of $\phi,\ 0 \leq \phi \leq 1/2$. Thus, the full maximization p-value for each outcome is the maximum of the sums of multinomial probabilities,

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \text{sup}_{\phi\in[0,0.5]} \sum_{\boldsymbol{y}\in R(\boldsymbol{y}_{\text{obs}};\phi)} \frac{N!}{y_1!y_2(N-y_1-y_2)!}\phi^{y_1+y_2}(1-2\phi)^{N-y_1-y_2}.$$

In this example, numerically we used a grid for ϕ of 5001 points, $\{0, 0.0001, 0.0002, \ldots, 0.5000\}$ in the maximization. For the outcome $\boldsymbol{y}_{\text{obs}} = (0, 2, 1)$, in each grid point, the multinomial probabilities are calculated for the outcomes in the tail set defined by $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$ where $T(\boldsymbol{y})$ is given in (4), i.e. for the outcomes (0,2,1), (0,3,0), (2,0,1), (3,0,0), and added. Then the maximum of those sums is the full maximization p-value. For this outcome, the maximum p-value is obtained when $\phi = 0.4$, then $\pi((0, 2, 1); \phi = 0.4) = \pi((2, 0, 1); \phi = 0.4) = 0.096$ and $\pi((0, 3, 0); \phi = 0.4) = \pi((3, 0, 0); \phi = 0.4) = 0.064$. Adding these probabilities yields the p-value 0.32. The p-values for the other outcomes are given in the third column of Table 4 with the value of ϕ for which the maximum p-value is obtained in the second column. □

Partial maximization   Not all values of $\boldsymbol{p}$ are equally likely under the null hypothesis, therefore it might not be desirable to maximize over all possible values of $\boldsymbol{p}$. The set over which the supremum is found can be restricted to a confidence set for $\boldsymbol{p}$ as suggested by Berger and Boos (1994). This partial maximization p-value is valid when a penalty ζ is added, Berger and Boos (1994), but it is not exact by the definition of Lloyd (2008). It is given by

$$P_{\text{PM}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in C_\zeta} \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); f(\boldsymbol{p}) = 0) + \zeta = \sup_{\boldsymbol{p} \in C_\zeta} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \boldsymbol{p})} \pi(\boldsymbol{y}; \boldsymbol{p}) + \zeta,$$

where C_ζ is the $1 - \zeta$ confidence region for $\boldsymbol{p}$ under the null hypothesis.

Trinomial example   Under H_0, $Y_1 + Y_2$ is binomially distributed with parameters 2ϕ and N. We will use the Clopper–Pearson confidence interval which is an exact confidence interval for binomial proportions, Agresti (2002). The $1 - \zeta$ confidence interval for ϕ is given by its lower limit C_{L} and upper limit C_{U},

$$C_{\text{L}} = \frac{1}{2}\left(1 + \frac{N - y_1 - y_2 + 1}{(y_1 + y_2) F_{2(y_1+y_2), 2(N-y_1-y_2+1)}(1 - \zeta/2)}\right)^{-1}$$

$$C_{\text{U}} = \frac{1}{2}\left(1 + \frac{N - y_1 - y_2}{(y_1 + y_2) F_{2(y_1+y_2+1), 2(N-y_1-y_2)}(\zeta/2)}\right)^{-1}$$

where $F_{\nu_1, \nu_2}(c)$ denotes the $1 - c$ quantile from the F distribution with ν_1 and ν_2 degrees of freedom. When $y_1 + y_2 = 0$, the lower limit is 0 and when $y_1 + y_2 = N = 3$, the upper limit is 0.50. We choose $\zeta = 0.001$ as suggested by Berger and Boos (1994) and calculate the p-values from the formula

$$P_{\text{PM}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\phi \in [C_{\text{L}}, C_{\text{U}}]} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \phi)} \frac{N!}{y_1! y_2 (N - y_1 - y_2)!} \phi^{y_1+y_2} (1 - 2\phi)^{N-y_1-y_2} + \zeta.$$

The p-values are calculated the same way as the full maximization p-values except that the maximization is now done over the possible values of ϕ in the confidence interval. The tail set is unchanged. Table 4, column 4 and 5, show the partial maximization p-values and the value of ϕ for which the maximum p-value is found. We see that the partial maximization p-values are the same as the full maximization p-values, except for the outcomes (0,1,2) and (1,0,2) where the p-values are smaller because the value of $\boldsymbol{p}$ that maximizes the full maximization p-value is outside the confidence interval for ϕ used by the partial maximization p-value.□

ESTIMATION AND MAXIMIZATION Lloyd (2008) proposed the estimation followed by maximization (E+M) p-value, where the negative of the estimation p-values serve as the values of a new test statistic, followed by a full maximization step. In this way valid and exact p-values are obtained.

Since the test statistic is the estimation (E) p-values, performing the estimation step results in a different ordering of the outcomes before performing the maximization step than the ordering defined by the original test statistic. Thus the tail sets are changed and the E+M p-values may differ from the full maximization (M) p-values. The estimation step can be done more than once with or without a final maximization step, each time resulting in a different ordering of the outcomes. If two estimation steps are performed before a maximization step, the p-values are called E^{2}M p-values, again yielding valid p-values.

Another way to look at the difference and similarity between the E and M p-values, is that for the E p-values, first the probability of the observed outcome is maximized through the maximum likelihood estimate of $\boldsymbol{p}$ under H_0, and then the probability of the tail set is calculated. For the M p-values, the tail sets are defined first, and then the probability of the tail set is maximized over $\boldsymbol{p}$ in $\mathcal{P}_0$. That is,

$$P_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\mathrm{obs}})} \sup_{\boldsymbol{p} \in \mathcal{P}_0} \pi(\boldsymbol{y}; \boldsymbol{p}),$$

and

$$P_{\mathrm{M}}(\boldsymbol{y}_{\mathrm{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\mathrm{obs}})} \pi(\boldsymbol{y}; \boldsymbol{p}).$$

Performing more than one maximization step in a sequence has no effect on the p-values. The reason is that the tail sets remain the same. Assume some chosen test statistic defines the tail set to be used in the first maximization step. The outcome having the largest value of this test statistic will have the smallest M p-value, the outcome having the second largest value of the test statistic will have the second smallest M p-value and so on. In the second maximization step, the negative of the M p-values are the values of the test statistic that defines the new tail set. The outcome having the largest negative M p-value is the outcome that had the largest value of the first test statistic, the outcome having the second largest negative M p-value is the outcome that had the second largest value of the first test statistic and so on. Since the p-value is the maximum sum of probabilities of the outcomes in the tail set, maximized over $\boldsymbol{p}$, and the tail set is the same, the p-values are unchanged.

Regardless of the choice of test statistic the M p-values are always valid. Assume that the chosen significance level is α. We would then reject the null hypothesis for all outcomes for which the test statistic $T(\boldsymbol{Y})$ is greater than or equal to some critical value k, where k is chosen so that all values of the test statistic that are greater than or equal to k yield a p-value less than or equal to α. The probability that a random outcome $\boldsymbol{y}_{\mathrm{obs}}$ is rejected under the null hypothesis is the probability that the test statistic $T(\boldsymbol{y}_{\mathrm{obs}})$ is greater than or equal to the critical value and this probability is less than or equal to the p-value for an outcome $\boldsymbol{y}$ for which $T(\boldsymbol{y}) = k$, which is less than or equal α. Also exactness of the M p-value follows by construction.

The M p-values are often conservative, see Bayarri and Berger (2000), whereas the E p-values are not valid and thus generally smaller than or equal to the M p-values. It is desired when comparing different p-values to obtain p-values as small as possible while still valid.

Trinomial example We first find the E p-values for all the outcomes where each p-value is calculated by inserting the maximum likelihood estimate $\tilde{\boldsymbol{p}}$ for $\boldsymbol{p}$ for that particular outcome as described previously and given in Table 2. The E p-values are used to define the tail set of the observed outcome which is the set of outcomes for which $P_{\mathrm{E}}(\boldsymbol{y}) \leq P_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}})$. For the observed outcome we then calculate the E+M p-value as the sum of multinomial probabilities of the outcomes in the tail set maximized over ϕ. Let $R_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}})$ be the tail set of the observed outcome defined by the E p-values smaller than or equal to the E p-value of the observed outcome. The E+M p-value is then given by

$$P_{\mathrm{E+M}}(\boldsymbol{y}) = \sup\nolimits_{\phi \in [0,0.5]} \sum_{\boldsymbol{y} \in R_{\mathrm{E}}(\boldsymbol{y})} \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} \phi^{y_1+y_2} (1 - 2\phi)^{N-y_1-y_2}.$$

The E p-value of outcome (0,2,1) is 0.30. The tail set $R^E(0,2,1) = \{(0,2,1), (0,3,0), (2,0,1), (3,0,0)\}$. This is the same tail set as when we used the test statistic $T(\boldsymbol{Y}) = |Y_1/N - Y_2/N|$ and therefore the maximization over ϕ here yields the same maximum p-value as the full maximization approach. The E+M p-values for all the outcomes are given in Table 4, column 7, with the value of ϕ for which the maximum p-value is found, ϕ_M, in column 6. The p-values are maximized with respect to ϕ over the same grid as in the full maximization approach. We see that for all the outcomes, except (0,1,2) and (1,0,2), the E+M p-values are the same as the full maximization p-values. For those two outcomes, the p-values are reduced from 1 to 0.5982 if we use the E+M approach, and they are the same outcomes for which the p-values were reduced when performing partial maximization instead of full maximization. □

Outcome $\boldsymbol{y}$	ϕ_{M}	P_{M}	ϕ_{PM}	P_{PM}	$\phi_{\mathrm{E+M}}$	$P_{\mathrm{E+M}}$
(0,0,3)	0.4535	1	0.4591	1	0.4535	1
(0,1,2)	0.50	1	0.4935	0.982	0.2265	0.5982
(0,2,1)	0.40	0.32	0.40	0.32	0.40	0.32
(0,3,0)	0.50	0.25	0.50	0.25	0.50	0.25
(1,0,2)	0.50	1	0.4935	0.982	0.50	0.5982
(1,1,1)	0.4535	1	0.4726	1	0.4535	1
(1,2,0)	0.50	1	0.50	1	0.4535	1
(2,0,1)	0.40	0.32	0.40	0.32	0.40	0.32
(2,1,0)	0.50	1	0.50	1	0.4535	1
(3,0,0)	0.50	0.25	0.50	0.25	0.50	0.25

TABLE 4: P-values obtained for the trinomial example by full maximization, partial maximization and estimation plus maximization with values of ϕ for which the p-values are maximized.

INTEGRATION In the partial maximization approach, the points in $\mathcal{P}_0$, the parameter space for $\boldsymbol{p}$ under H_0, are given weights 0 or 1. If $\boldsymbol{p}$ lie within their confidence interval, they are given weight 1 and 0 otherwise. Instead of weighing the probabilities with 0 and 1, we want to apply Bayesian methodology and weigh the points in $\mathcal{P}_0$ according to a prior distribution $\pi(\boldsymbol{p})$. We integrate out $\boldsymbol{p}$ in order to be able to calculate $\pi(\boldsymbol{y} \mid H_0)$. Bayarri and Berger (2000) review several Bayesian p-values as well as suggesting two new p-values. First there is the prior predictive p-value where a prior $\pi(\boldsymbol{p} \mid H_0)$ is chosen, so that the probability of an outcome under the null hypothesis is

$$\pi(\boldsymbol{y} \mid H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p}) \pi(\boldsymbol{p} \mid H_0) \mathrm{d}\boldsymbol{p}.$$

The prior predictive (PP) p-value of an observed outcome $\boldsymbol{y}_{\mathrm{obs}}$ is the sum of the probabilities

$\pi(\boldsymbol{y} \mid H_0)$ that are less than or equal to the probability $\pi(\boldsymbol{y}_{\text{obs}} \mid H_0)$. As we will see numerically in Section 4.2, these PP p-values are not valid.

Trinomial example We choose the uniform Dirichlet prior as the joint distribution of p_1 and p_2, thus $\pi(\boldsymbol{p}) = 2$. Let $z = p_1 - p_2$. Under H_0, $z = 0$, so that $\pi(\boldsymbol{p} \mid H_0) = \pi(\boldsymbol{p} \mid z = 0)$. The joint density of the transformed variables is $\pi(p_1, z) = 2$. Then $\pi(\boldsymbol{p} \mid z = 0) = \pi(p_1, z = 0)/\pi(z = 0)$. The density of z can be found by integrating out p_1 from $\pi(p_1, z)$, giving $\pi(z = 0) = 1$, after having identified the triangular region to which (p_1, z) belongs. The probability of the trinomial outcome given H_0 is

$$\pi(\boldsymbol{y} \mid H_0) = \int_0^{1/2} \frac{N!}{y_1!y_2!} p_1^{y_1+y_2}(1-2p_1)^{N-y_1-y_2} 2\mathrm{d}p_1 = \frac{(y_1+y_2)!}{y_1!y_2!(N+1)} \left(\frac{1}{2}\right)^{y_1+y_2}.$$

The p-value of the observed outcome is the sum of the probabilities $\pi(\boldsymbol{y} \mid H_0)$ that are less than or equal to the probability of the observed outcome. Table 5 shows the calculated probabilities as well as the p-values for the possible outcomes in the trinomial example. □

Outcome $\boldsymbol{y}$	$\pi(\boldsymbol{p} \mid H_0)$	P_{PP}
(0,0,3)	0.25	1
(0,1,2)	0.125	0.625
(0,2,1)	0.0625	0.1875
(0,3,0)	0.03125	0.0625
(1,0,2)	0.125	0.625
(1,1,1)	0.125	0.625
(1,2,0)	0.09375	0.375
(2,0,1)	0.0625	0.1875
(2,1,0)	0.09375	0.375
(3,0,0)	0.03125	0.0625

TABLE 5: P-values for the trinomial example using the Bayesian approach and a uniform Dirichlet prior on $\boldsymbol{p}$.

One challenge with the prior predictive approach is that the resulting p-values depend on the prior. To make them less dependent on the choice of prior and more dependent on the data, one can use the posterior predictive p-value, Bayarri and Berger (2000), where the probability of the observed outcome is given in terms of the posterior probability,

$$\pi(\boldsymbol{y} \mid H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p})\pi(\boldsymbol{p} \mid \boldsymbol{y}_{\text{obs}})\mathrm{d}\boldsymbol{p}.$$

To calculate this probability, improper priors can be used, and the probability will be less influenced by the choice of prior. However, the data are used twice since it first is needed to determine the posterior distribution and then in computing the tail set.

As improvements to the posterior predictive p-value, Bayarri and Berger (2000) also suggested the partial posterior predictive p-value and the conditional predictive p-value. In this work, we will only consider the prior predictive p-values.

TEST SIZE AND POWER. The test size and test power are common evaluation measures on the performance of a statistical hypothesis test. The test size is the probability of making a type I error, i.e. to reject the null hypothesis, when it is true. The power is the probability of rejecting the null hypothesis when it is not true, which is one minus the probability of making a type II error, i.e. to not reject the null hypothesis when it is not true. Given the chosen significance level α, the test size for a test for which we reject the null hypothesis when the p-value $P(\boldsymbol{y})$ is less than α is

$$\Pr(P(\boldsymbol{Y}) \leq \alpha; \boldsymbol{p}) = \sum_{\boldsymbol{y};\, P(\boldsymbol{y}) \leq \alpha} \pi(\boldsymbol{y}; \boldsymbol{p}) \tag{5}$$

for a parameter $\boldsymbol{p}$ in $\mathcal{P}_0$.

The test power is

$$\Pr(P(\boldsymbol{y}) \leq \alpha; \boldsymbol{p}) = \sum_{\boldsymbol{y};\, P(\boldsymbol{y}) \leq \alpha} \pi(\boldsymbol{y}; \boldsymbol{p}) \tag{6}$$

for a parameter $\boldsymbol{p}$ in $\mathcal{P}$.

3 Independent binomial proportions

Before focusing on our main problem of comparing positive predictive values, we go through a fictitious example analyzed in Berger and Boos (1994) and Lloyd (2008), and present alternative test statistics and a more elaborate analysis of the p-values.

3.1 Presentation of the problem

There are $n = 330$ subjects in a clinical trial of which $n_1 = 47$ subjects receive treatment and $n_2 = 283$ subjects receive placebo. Let X_1 be the number of subjects that survive among those who received treatment and let X_2 be the number of subjects that survive among those who received placebo. If p_1 is the survival probability for the treatment group and p_2 is the survival probability for the placebo group, we assume that X_1 is binomially distributed with parameters n_1 and p_1 and X_2 is binomially distributed with parameters n_2 and p_2. Let $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{p} = (p_1, p_2)$. The two-sided null hypothesis is that the survival probabilities in the two groups are equal, i.e.

$$H_0 : f(\boldsymbol{p}) = p_1 - p_2 = 0 \tag{7}$$

versus the alternative that they are not equal,

$$H_1 : f(\boldsymbol{p}) = p_1 - p_2 \neq 0.$$

Lloyd (2008) also considered the one-sided null hypothesis that the survival probability of the treatment group is no better than the survival probability of the placebo group, i.e.

$$H_0 : f(\boldsymbol{p}) = p_1 - p_2 \leq 0 \tag{8}$$

versus the alternative that the survival probability of the treatment group is better than the probability of the placebo group,

$$H_1 : f(\boldsymbol{p}) = p_1 - p_2 > 0.$$

Assuming independence between the treatment and placebo group, the joint distribution of X_1 and X_2 is the product of the two binomial distributions,

$$\Pr(X_1 = x_1, X_2 = x_2) = \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1} \cdot \binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2}$$

In this situation, the reference set is all possible outcomes (x_1, x_2) given $n_1 = 47$ and $n_2 = 283$ which is a set of 13 682 outcomes. When calculating p-values various test statistics can be used to define the tail set. One of the test statistics used by Berger and Boos (1994) and Lloyd (2008) is

$$T_{\mathrm{T}}(x_1, x_2) = \frac{x_1/n_1 - x_2/n_2}{\sqrt{(x_1+x_2)(n-x_1-x_2)/(nn_1n_2)}}.$$

When testing the null hypothesis (7) the tail set of an observed outcome $(x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}})$ is $R(x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}}) = \{(x_1, x_2) : |T_{\mathrm{T}}(x_1, x_2)| \geq |T_{\mathrm{T}}(x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}})|\}$, and when testing the null hypothesis (8) the tail set is $R(x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}}) = \{(x_1, x_2) : T_{\mathrm{T}}(x_1, x_2) \geq T_{\mathrm{T}}(x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}})\}$.

Lloyd (2008) also uses the likelihood ratio test statistic

$$T_{\mathrm{LR}} = 2\sum_{i=1}^{2}\left(x_i \mathrm{log}\frac{\hat{p}_i}{\tilde{p}_i} + (n_i - x_i)\mathrm{log}\frac{1-\hat{p}_i}{1-\tilde{p}_i}\right)$$

where $\hat{p}_i = x_i/n_i$ is the general maximum likelihood estimate for p_i, $i = 1, 2$ and $\tilde{p}_i$ is the maximum likelihood estimate for p_i, $i = 1, 2$ under the null hypothesis. If we are testing the two-sided null hypothesis $\tilde{p}_1 = \tilde{p}_2 = (x_1 + x_2)/(2n)$, and if we are testing the one-sided null hypothesis, $\tilde{p}_1 = \tilde{p}_2 = (x_1 + x_2)/(2n)$ when $x_1/n_1 \geq x_2/n_2$ and $\tilde{p}_i = x_i/n_i, \ i = 1, 2$, when $x_1/n_2 < x_2/n_2$. These estimates were also used for the E step. For the maximization in the M step, 1001 equally spaced values of $p_1 = p_2$ in [0,1] were used for the two-sided test and 5151 equally spaced points in a rectangular grid in the triangular region $0 \leq p_1 \leq 1$, $p_1 \leq p_2 \leq 1$, were used for the one-sided test.

In addition to T_{T} and T_{LR} we propose three additional test statistics. Let $\pi(\boldsymbol{x}; \boldsymbol{p})$ denote $\Pr(X_1 = x_1, X_2 = x_2; \boldsymbol{p})$. First, we define a simplified version of the likelihood ratio test statistic,

$$T_{\pi_{\mathrm{e}}}(\boldsymbol{x}_{\mathrm{obs}}) = \pi(\boldsymbol{x}_{\mathrm{obs}}; \tilde{\boldsymbol{p}}_{\mathrm{obs}}),$$

which is simply the probability of the observed outcome $\boldsymbol{x}_{\mathrm{obs}} = (x_{1,\mathrm{obs}}, x_{2,\mathrm{obs}})$ with the maximum likelihood estimate of $\boldsymbol{p}$ under H_0 for this outcome, $\tilde{\boldsymbol{p}}_{\mathrm{obs}}$, inserted for $\boldsymbol{p}$.

In our second and third additional test statistic, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$, we let the probability $\pi(\boldsymbol{x}; \boldsymbol{p})$ of an outcome $\boldsymbol{x}$ play the role of a test statistic. It is of course dependent on the unknown parameters, and thus not a test statistic in the ordinary sense. It still makes sense to apply an E or M step to it, yielding the π_{E} p-value

$$T_{\pi_{\mathrm{E}}}(\boldsymbol{x}_{\mathrm{obs}}) = \Pr(\pi(\boldsymbol{X}; \tilde{\boldsymbol{p}}_{\mathrm{obs}}) \leq \pi(\boldsymbol{x}_{\mathrm{obs}}; \tilde{\boldsymbol{p}}_{\mathrm{obs}}); \tilde{\boldsymbol{p}}_{\mathrm{obs}}) = \sum_{\boldsymbol{x} \in R^*(\boldsymbol{x}_{\mathrm{obs}})} \pi(\boldsymbol{x}; \tilde{\boldsymbol{p}}_{\mathrm{obs}})$$

where $R^*(\boldsymbol{x}_{\mathrm{obs}})$ consists of those $\boldsymbol{x}$ for which $\pi(\boldsymbol{x}; \tilde{\boldsymbol{p}}_{\mathrm{obs}}) \leq \pi(\boldsymbol{x}_{\mathrm{obs}}; \tilde{\boldsymbol{p}}_{\mathrm{obs}})$, and the π_{M} p-value

$$T_{\pi_{\mathrm{M}}}(\boldsymbol{x}_{\mathrm{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(\pi(\boldsymbol{X}; \boldsymbol{p}) \leq \pi(\boldsymbol{x}_{\mathrm{obs}}; \boldsymbol{p}); \boldsymbol{p}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{\boldsymbol{x} \in R^*(\boldsymbol{x}_{\mathrm{obs}})} \pi(\boldsymbol{x}; \boldsymbol{p}),$$

where $R^*(\boldsymbol{x}_{\mathrm{obs}})$ consists of those $\boldsymbol{x}$ for which $\pi(\boldsymbol{x}; \boldsymbol{p}) \leq \pi(\boldsymbol{x}_{\mathrm{obs}}; \boldsymbol{p})$. Note that the sets $R^*(\boldsymbol{x})$ for these two statistics are dependent on the parameter $\boldsymbol{p}$, as opposed to the tail sets defined by an ordinary statistic. Although $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$ are constructed in a similar manner as p-values constructed by an E and an M step, respectively, their use will be as test statistics, rather than p-values.

3.2 COMPARISON OF TEST STATISTICS

Lloyd (2008) recommends using the E+M p-values in the problem of comparing independent binomial proportions and we want to compare the test size and power of the T_{T}, T_{LR}, $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$ test statistics for these p-values when testing both the one-sided and two-sided null hypotheses given in (7) and (8). To calculate the test size, i.e. the probability of rejecting the null hypothesis given that the null hypothesis is true, we generated 10001 equally spaced values of $p_1 = p_2$ in $[0, 1]$ and calculated the test size for each value $\boldsymbol{p}$ according to (5) by adding the probabilities of the outcomes that had a p-value less than or equal 0.05, which was the chosen significance level. To assess power, we used 9001 equally spaced points on the line $p_1 = p_2 + 0.1$, for which we calculated the power by adding the probabilities of the outcomes that had p-values less than or equal to 0.05, i.e. using (6).

Table 6 shows the mean test size and power for the five test statistics followed by an E and an M step. For the two-sided hypothesis, T_{T}, $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$ have similar mean test size, of which $T_{\pi_{\text{e}}}$ has the greatest. T_{LR} yields a smaller test size than the other test statistics. When testing the one-sided hypothesis, the test size of $T_{\pi_{\text{e}}}$ is 0.0434 which is greater than the test size for the other test statistics which ranges from 0.0382 to 0.0387. We see that the test statistics have similar power, lower for the two-sided test than for the one-sided test, and for the two-sided test, T_{LR} has the smallest power and $T_{\pi_{\text{e}}}$ has the largest power. For the one-sided test, $T_{\pi_{\text{E}}}$ has the lowest power and $T_{\pi_{\text{e}}}$ has the largest power. This indicates that for this problem, the $T_{\pi_{\text{e}}}$ statistic performs best and should be considered an alternative to T_{T} and T_{LR}.

Table 7 shows the E+M p-values for the observed outcome $(x_1, x_2) = (14, 48)$ which for T_{T} and T_{LR} agree with Lloyd (2008). For the two-sided test, the p-value for outcome (14,48), which is used as a test case by Berger and Boos (1994) and Lloyd (2008), is less than 0.05 for all the test statistics except T_{LR} so the null hypothesis would be rejected on a 5% significance level for four of the test statistics. T_{T} yields the smallest p-value. For T_{LR} we reject the one-sided null hypothesis. All test statistics yield the p-value 0.025 for the one-sided test and thus reject the null hypothesis.

Hypothesis	T_{T}	T_{LR}	$T_{\pi_{\text{e}}}$	$T_{\pi_{\text{E}}}$	$T_{\pi_{\text{M}}}$
Two-sided	0.0472	0.0435	0.0479	0.0478	0.0472
One-sided	0.0386	0.0384	0.0434	0.0382	0.0387
Two-sided	0.3629	0.3475	0.3649	0.3644	0.3622
One-sided	0.4421	0.4410	0.4639	0.4388	0.4427

TABLE 6: Mean test size in the two upper rows and mean test power in the two lower rows for the two-sided and one-sided hypothesis for the E+M p-values using different test statistics.

Hypothesis	T_{T}	T_{LR}	$T_{\pi_{\text{e}}}$	$T_{\pi_{\text{E}}}$	$T_{\pi_{\text{M}}}$
Two-sided	0.037	0.057	0.040	0.041	0.040
One-sided	0.025	0.025	0.025	0.025	0.025

TABLE 7: E+M p-values from the two-sided and one-sided tests for outcome (14,48).

Figure 1 shows the E+M p-values for $T_{\pi_{\text{e}}}$ test plotted against the E+M p-values for T_{T} for p-values less than or equal to 0.11. The plot shows that $T_{\pi_{\text{e}}}$ overall yields smaller p-values than T_{T} and as we want p-values that are as small as possible provided that they are valid, $T_{\pi_{\text{e}}}$ seems to be preferable over T_{T}.

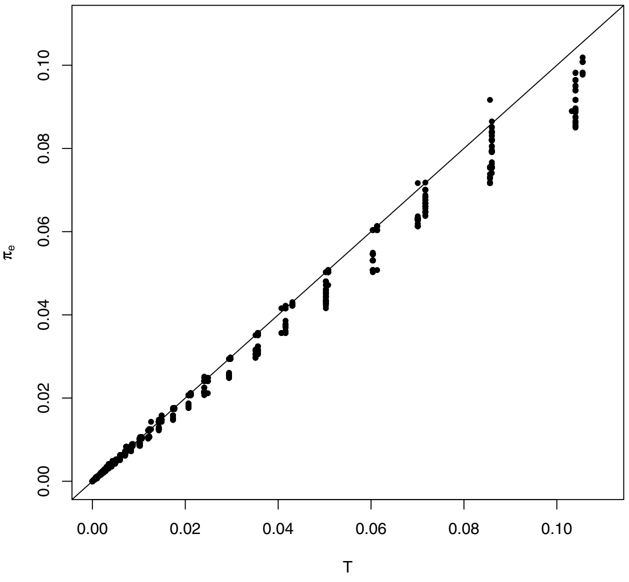


FIGURE 1: E+M p-values for T_{π_e} plotted against E+M p-values for T_{T}.

To explain what happens in the E and M steps, we look into the results for three particular outcomes, (16,52), (1,0) and (30,126), which are consecutive decreasing outcomes when ordering by T_{T}. Table 8 shows the value of T_{T} for these outcomes in the third column and the probabilities $\pi(\boldsymbol{x};\boldsymbol{p})$ of the outcomes inserted the maximum likelihood estimates under the null hypothesis in the second column. Note how much larger $\pi((1,0);\tilde{\boldsymbol{p}}_{(1,0)})$ is than $\pi((16,52);\tilde{\boldsymbol{p}}_{(16,52)})$ and $\pi((30,126);\tilde{\boldsymbol{p}}_{(30,126)})$. When performing the E step, this larger probability has a significant effect, as it is included in the sum of probabilities that yields the p-value for outcome (1,0). The fourth column of Table 8 shows the E p-values for the three outcomes. The p-value for outcome (1,0) is 0.05970 which is much greater than the two other p-values and H_0 is rejected on a 5% significance level, the other two p-values are both less than 0.01 and H_0 will not be rejected. Thus, the decision of whether to reject H_0 differ for these three outcomes, even though the values of T_{T} are almost the same. The effect of the large probability of outcome (1,0) also shows in the M p-values in the fifth column in Table 8. We note a large increase in the M p-value for outcome (30,126) as compared to the E p-value, the reason being that the M p-value is at least as large as the E p-value by construction, in particular for (1,0), and next, that the the M p-value is at least as large as the M p-value of (1,0), since (1,0) has a larger test statistic value. According to the M p-values, we would not reject H_0 for any of those outcomes as opposite to the E p-values where H_0 is rejected for outcome (30,126). To avoid the effect of outcome (1,0), we need a different ordering of the outcomes, in which (1,0) is placed further down on the list where the outcomes are sorted by decreasing value of a test statistic. This is obtained by treating the E p-value as

$\boldsymbol{x}$	$\pi(\boldsymbol{x};\tilde{\boldsymbol{p}})$	$T_{\mathrm{T}}(x_1,x_2)$	P_{E}	P_{M}	$P_{\mathrm{E+M}}$
(16,52)	0.00049	2.45928	0.00968	0.02458	0.01029
(1,0)	0.05247	2.45756	0.05970	0.06112	0.07229
(30,126)	0.00028	2.45512	0.00722	0.06112	0.00746

TABLE 8: The test statistic $T_{\mathrm{T}}(x_1, x_2)$ and corresponding E, M and E+M p-values for outcomes (16,52), (1,0) and (30,126).

a test statistic and then applying the M step. The outcomes that had unusually large p-values after the E step compared to their neighbours, e.g. as outcome (1,0) had, is then moved down on the list when sorting the outcomes by decreasing negative E p-values. The sixth column of Table 8 shows these E+M p-values and we see that outcome (30,126) now has a p-value of 0.00746 and is thus unaffected by the outcome (1,0). This example indicates why it is beneficial to perform an E step prior to the M step. The M step is necessary to obtain valid p-values and therefore the E step alone is not sufficient.

The M p-value for the outcome (14,48) is 0.06114, see Lloyd (2008), which is significantly greater than the E+M p-value of Table 7. Our investigation showed that the value 0.06114 also arises from outcome (1,0), because the value of T_{T} is less for outcome (14,48) than for outcome (1,0) which is thus in the tail set of (14,48). The E+M p-value is smaller (0.025) because the E step changed the ordering of the outcomes and (1,0) was placed behind (14,48) so that after performing the E step, (1,0) is no longer in the tail set of (14,48). It should also be noted that $\pi((14,48);\phi)$, as a function of $\phi = p_1 = p_2$ has a prominent and narrow peak near $\phi = 0$ (but $\phi > 0$), which explains the large value of $\pi((1,0);\tilde{\boldsymbol{p}})$ and thus of $P_{\mathrm{E}}(1,0)$. Partial maximization avoids this peak, explaining that the PM p-value is reasonable as reported by Berger and Boos (1994), though not as small as the E+M p-values as reported by Lloyd (2008).

4 COMPARING POSITIVE PREDICTIVE VALUES

We now present the main problem of comparing the positive predictive values of two diagnostic tests. The performance of various test statistics and p-values are compared in terms of test size and power.

4.1 PRESENTATION OF THE PROBLEM

Suppose that two diagnostic tests are available for a particular disease of interest. We want to compare the prediction abilities of the two tests, which can be quantified by the positive and negative predictive values. The positive predictive value is defined as the probability that a subject has the disease given that the test is positive and the negative predictive value is the probability that a subject does not have the disease given that the test is negative. Without loss of generality, in this work we will only consider the positive predictive values, as the tests for the negative predictive values can easily be derived along the same lines. We want to test whether the positive predictive value of test A is equal to the positive predictive value of test B against the alternative that they are not equal;

$$H_0\colon \mathrm{PPV_A} = \mathrm{PPV_B} \quad \text{vs} \quad H_1\colon \mathrm{PPV}_A \neq \mathrm{PPV}_B.$$

In this situation we define six random variables that are given in Table 9. Let $\boldsymbol{Y}$ be the vector of these random variables, i.e. $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$. We assume that $\boldsymbol{Y}$ is multinomially distributed with parameters N and $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6)$. Thus, the probability function of $\boldsymbol{Y}$ is

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = N! \prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!}.$$

Variable	Description
Y_1	Number of non-diseased subjects with positive test A and B.
Y_2	Number of non-diseased subjects with positive test A and negative test B.
Y_3	Number of non-diseased subjects with negative test A and positive test B.
Y_4	Number of diseased subjects with positive test A and B.
Y_5	Number of diseased subjects with positive test A and negative test B.
Y_6	Number of diseased subjects with negative test A and positive test B.

TABLE 9: Definition of the random variables $Y_1, \ldots, Y_6$.

The positive predictive value of test A is

$$\text{PPV}_A = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5}$$

and the positive predictive value of test B is

$$\text{PPV}_B = \frac{p_4 + p_6}{p_1 + p_3 + p_4 + p_6}.$$

The null hypothesis is then

$$H_0 : f_{\text{PPV}}(\boldsymbol{p}) = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5} - \frac{p_4 + p_6}{p_1 + p_3 + p_4 + p_6} = 0. \qquad (9)$$

The parameters $\boldsymbol{p}$ are not completely determined by the null hypothesis, we only know that $f_{\text{PPV}}(\boldsymbol{p}) = 0$ and that $\sum_{i=1}^{6} p_i = 1$. Thus, from these two constraints, two of the parameters can be expressed in terms of the four other remaining parameters, but these four parameters will be unknown nuisance parameters.

In order to test the null hypothesis (9) we will calculate p-values by enumeration as described by the algorithm in Section 2.3. The first step is to find the reference set.

4.1.1 FINDING THE REFERENCE SET

In the first step in the algorithm for calculating p-values, we enumerate using five nested for-loops to find all possible outcomes having the value of N, which is the number of observations. It can be distributed among six non-negative integer random variables having sum N, and the number of possible outcomes is $\binom{N+5}{5}$. This is a general result for the number of distinct unordered selections of N elements from six elements drawn with replacement. Figure 2 shows the number of outcomes on a log10 scale plotted against N. The number of outcomes grows very quickly when N increases. When $N = 25$ there are 142506 possible outcomes and when $N = 50$, there are nearly 3.5 million possible outcomes.

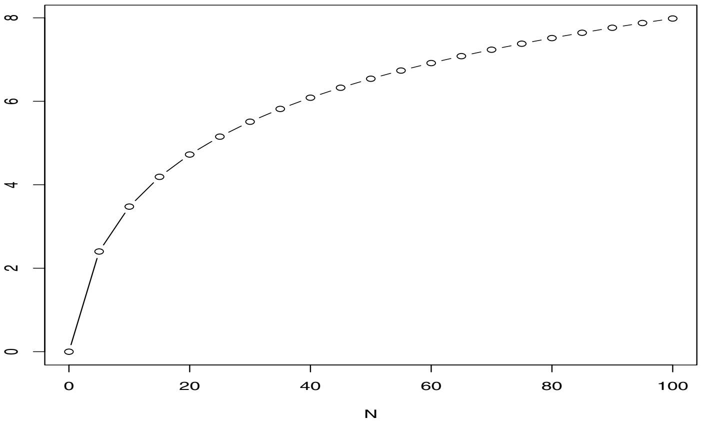


FIGURE 2: Number of possible outcomes on $\log_{10}$ scale as a function of N .

4.1.2 CALCULATING THE PROBABILITY OF THE OBSERVED OUTCOME

In the setting of comparing positive predictive values we will focus on calculating the probability of an outcome either by substituting maximum likelihood estimates for $\boldsymbol{p}$, maximize over the parameter space for $\boldsymbol{p}$ or integrate out $\boldsymbol{p}$ by a Bayesian approach. This results in the estimation (E), maximization (M) or combinations of these like the estimation and maximization (E+M) p-values, and the Bayesian prior predictive p-values. As far as we know, there is no sufficient or ancillary statistic for $\boldsymbol{p}$ in this problem. To calculate the p-values, a test statistic $T(\boldsymbol{Y})$ must be chosen. There are several possible test statistics for this problem, and they will be presented in Section 4.1.3.

ESTIMATION AND MAXIMIZATION If we substitute the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ for $\boldsymbol{p}$ under H_0, the E p-value for an outcome $\boldsymbol{y}_{\text{obs}}$ is

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \text{Pr}(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} N! \prod_{i=1}^{6} \frac{\tilde{p}_{i,\text{obs}}^{y_i}}{y_i!} \tag{10}$$

where the tail set $R(\boldsymbol{y})$ is defined by the chosen test statistic, $T(\boldsymbol{Y})$, and $\tilde{p}_{i,\text{obs}}$ is the maximum likelihood estimate under H_0 for p_i, $i = 1, \ldots, 6$ for the outcome $\boldsymbol{y}_{\text{obs}}$.

By maximizing the probability of the outcome $\boldsymbol{y}_{\text{obs}}$ over the parameter space $\mathcal{P}_0$ where $f_{\text{PPV}}(\boldsymbol{p}) = 0$, the M p-value is given by

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \text{sup}_{\boldsymbol{p} \in \mathcal{P}_0} \text{Pr}(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \boldsymbol{p}) = \text{sup}_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{y \in R(\boldsymbol{y}_{\text{obs}})} N! \prod_{i=1}^{6} \frac{p_{i,\text{obs}}^{y_i}}{y_i!}, \tag{11}$$

where $R(\boldsymbol{y})$ is defined by the chosen test statistic. When we calculate the E+M p-value, the expression is the same as in (11), but the tail set is then defined by the p-values calculated from (10). We will also consider the double estimation (E^2) p-values, where we first calculate p-values from (10) and then use these p-values as test statistics to define the tail set when calculating p-values from (10) once more. Finally we will maximize the E^2 p-values by using these as test statistics to define the tail set in (11), which results in E^{2}M p-values.

INTEGRATION We also consider the Bayesian prior predictive p-values, which requires a different approach. The starting point is still that the probability of an outcome under the null hypothesis is unknown because the parameters $\boldsymbol{p}$ are not completely specified. Instead of estimating $\boldsymbol{p}$ or maximizing the p-values over $\boldsymbol{p}$, we weigh them according to how likely they are under the null hypothesis.

We start out by conditioning on the parameter $\boldsymbol{p}$. Let $\pi(\boldsymbol{y}|H_0)$ be the probability of the outcome $\boldsymbol{y}$ under the null hypothesis (9). Then

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y}|\boldsymbol{p}) \cdot \pi(\boldsymbol{p}|H_0)\mathrm{d}\boldsymbol{p} \tag{12}$$

The first factor of the integrand, the probability of $\boldsymbol{y}$ given $\boldsymbol{p}$ is simply the multinomial distribution, i.e.,

$$\pi(\boldsymbol{y}|\boldsymbol{p}) = N! \prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!},$$

and $\pi(\boldsymbol{p}|H_0)$ is the probability density function for $\boldsymbol{p}$ under the null hypothesis.

Since $p_6 = 1 - \sum_{i=1}^5 p_i$ we first reduce the problem to five unknown parameters. Let

$$z = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5} - \frac{1 - p_1 - p_2 - p_3 - p_5}{1 - p_2 - p_5}. \tag{13}$$

which is $f_{\text{PPV}}(\boldsymbol{p})$ (9) with $1 - \sum_{i=1}^5 p_i$ inserted for p_6.

Under the null hypothesis $z = 0$ and from this an expression for p_4 can be derived, yielding four unknown parameters. We change variables from p_1, p_2, p_3, p_4, p_5 to p_1, p_2, p_3, z, p_5. The vector of the new parameters is denoted $\boldsymbol{p}^*$ in the following. Then $\pi(\boldsymbol{p}^*|z = 0) \propto \pi(\boldsymbol{p}^*, z = 0)$ so therefore we start by finding $\pi(\boldsymbol{p}^*, z)$. We use the formula for change of variables, $\pi(\boldsymbol{p}) = \pi(\boldsymbol{p}^*, z) \cdot |J|$, where J is the Jacobian determinant,

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{\partial z}{\partial p_1} & \frac{\partial z}{\partial p_2} & \frac{\partial z}{\partial p_3} & \frac{\partial z}{\partial p_4} & \frac{\partial z}{\partial p_5} \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix} = \frac{\partial z}{\partial p_4}$$

The absolute value $|J|$ is $\frac{p_1+p_2}{(p_1+p_2+p_4+p_5)^2}$. As a prior distribution for $\boldsymbol{p}$ we first apply the Dirichlet distribution with parameters $\alpha_1 = \alpha_2 = \ldots \alpha_5 = 1$, thus $\pi_1(\boldsymbol{p})$ is constant. Then

$$\pi(\boldsymbol{p}^*|z = 0) \propto \pi(\boldsymbol{p}^*, z = 0) = \pi_1(\boldsymbol{p}) \cdot |J|^{-1} \propto \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2},$$

and

$$\pi(\boldsymbol{p}|H_0) = \frac{1}{k_1} \cdot \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2},$$

where k_1 is a normalizing constant which can be found from $\int_{\mathcal{P}_0} \pi(\boldsymbol{p}|H_0) = 1$.

This leads to the following expression for the probability under H_0 of an outcome $\boldsymbol{y}$,

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_1} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2} \mathrm{d}\boldsymbol{p}. \tag{14}$$

For details on how to numerically compute the integral, see Section 5.

The p-value is the sum of these probabilities for outcomes in the tail set of the observed outcome. In this setting we use the probability of an outcome under H_0 as the test statistic and the tail set for an outcome $\boldsymbol{y}_{\text{obs}}$ is defined as the outcomes for which $\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)$, so the p-value is given as

$$\begin{aligned} P_{\text{PP},1}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_1} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{15}$$

To assess the effect of the choice of prior, we choose the non-uniform prior $\pi_2(\boldsymbol{p}) \propto p_1$ as an alternative prior, which leads to

$$\pi_2(\boldsymbol{p}|H_0) = \frac{p_1(p_1 + p_2 + p_4 + p_5)^2}{k_2(p_1 + p_2)},$$

where k_2 is a normalizing constant and the probability under H_0 of an outcome $\boldsymbol{y}$ is

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_2} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{p_1(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2} \mathrm{d}\boldsymbol{p}. \tag{16}$$

We see that the probability of the outcome depends on the chosen prior $\pi_2(\boldsymbol{p})$ as expected. The p-value, which is the sum of the probabilities in (16) for the outcomes that are in the tail set of the one observed, is denoted $P_{\text{PP},2}$ and given by

$$\begin{aligned} P_{\text{PP},2}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_2} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{p_1(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{17}$$

An alternative formulation of the null hypothesis (9) is

$$f^*_{\text{PPV}}(\boldsymbol{p}) = p_1(p_1 + p_2 + p_3 + p_4 + 2p_5 - 1) - p_2(1 - p_1 - p_2 - p_3 - p_5) + p_3(p_4 + p_5) = 0. \tag{18}$$

In this case the absolute value of the Jacobi determinant will be $p_1 + p_3$ and if we assume the uniform Dirichlet prior $\pi_1(\boldsymbol{p})$,

$$\pi(\boldsymbol{p}|H_0) = \frac{1}{k_3} \cdot \frac{1}{p_1 + p_3},$$

where k_3 is a normalizing constant and the probability under H_0 of an outcome is

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_3} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \cdot \frac{1}{p_1 + p_3} \mathrm{d}\boldsymbol{p}. \tag{19}$$

This probability is clearly not equal to the probability (14) and this is an example of Borel's paradox. The p-value is the sum of the probabilities in (19) for the outcomes in the tail set of the observed outcome. It is denoted $P_{\text{PP},3}$ and given by

$$\begin{aligned} P_{\text{PP},3}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0)\leq\pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_3} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{1}{p_1+p_3} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{20}$$

4.1.3 Defining the tail set

The tail set of an outcome $\boldsymbol{y}$ is defined by a test statistic $T(\boldsymbol{y})$. To test the null hypothesis in (9) there are several tests available that are used for large samples, see Günther, Bakke, Lydersen and Langaas (2008) for a detailed description of four possible test statistics. In this work we will use these test statistics to define the tail set while ignoring their asymptotic distribution.

The first test statistic is the likelihood ratio test statistic which is the ratio between the maximum likelihood under the null hypothesis and the general maximum likelihood, of which by convenience the logarithm is taken and which is multiplied by -2, Casella and Berger (2002). In our multinomial situation, it is given as

$$T_{\text{LR}} = -2 \cdot \log \frac{\sup_{\boldsymbol{p}\in\mathcal{P}_0} L(\boldsymbol{p}|\boldsymbol{y})}{\sup_{\boldsymbol{p}\in\mathcal{P}} L(\boldsymbol{p}|\boldsymbol{y})} = -2 \sum_{i=1}^{6} y_i \cdot (\log \tilde{p}_i - \log \hat{p}_i), \tag{21}$$

where $\tilde{p}_i$ is the restricted maximum likelihood estimates of p_i, i.e. under H_0, $i = 1, \ldots, 6$, and $\hat{p}_i$ is the unrestricted general maximum likelihood estimates for the multinomial distribution, i.e., $\hat{p}_i = n_i/N$, $i = 1, \ldots, 6$, Johnson, Kotz and Balakrishan (1997). The maximum likelihood estimates under H_0, $\tilde{p}_i$, $i = 1, \ldots, 6$ cannot be written in closed form, but can be found analytically by solving a system of equations arising from the method of Lagrange multipliers, which we did using Maple 12. More details can be found in Günther et al. (2008), Section 3.1.2.

The difference test statistic is given by

$$T_g(\boldsymbol{y}) = \frac{(g(\boldsymbol{Y}) - g(\boldsymbol{\mu}))^2}{G^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\, G(\boldsymbol{\mu})} \tag{22}$$

where $g(\boldsymbol{Y})$ is an estimator for the difference $f_{\text{PPV}}(\boldsymbol{p})$ in (9), i.e.

$$g(\boldsymbol{Y}) = \frac{Y_4+Y_5}{Y_1+Y_2+Y_4+Y_5} - \frac{Y_4+Y_6}{Y_1+Y_3+Y_4+Y_6},$$

and $\boldsymbol{\mu} = \text{E}(\boldsymbol{Y}) = N \cdot \boldsymbol{p}$, $\Sigma = \text{Cov}(\boldsymbol{Y}) = N(\text{Diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p})$, G is a vector containing the first order partial derivatives of $g(\boldsymbol{Y})$ with respect to the components of $\boldsymbol{Y}$, G^T is the transpose of G, and $G(\boldsymbol{\mu})$ is G with $\boldsymbol{\mu}$ inserted for $\boldsymbol{Y}$. Under the null hypothesis $g(\boldsymbol{\mu}) = 0$. $G(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ depend on the unknown parameters $\boldsymbol{p}$ which must be estimated when calculating the test statistic. We can either insert the unrestricted maximum likelihood estimates for the multinomial distribution $\hat{\boldsymbol{p}}$ and then we refer to the test as the *unrestricted* difference test (uDT) and denote the test statistic T_{uDT}, or insert restricted maximum likelihood estimates under H_0, $\tilde{\boldsymbol{p}}$. Then the test is referred to as the *restricted* difference test (rDT) and the test statistic is denoted T_{rDT} .

Leisenring, Alonzo and Pepe (2000) presented a score test based on generalized estimating equations. We denote this test the LAP test. The test statistic can be written as

$$T_{LAP} = \frac{((Y_1 + Y_2 + Y_4 + Y_5)(Y_4 + Y_6) - (Y_1 + Y_3 + Y_4 + Y_6)(Y_4 + Y_5))^2}{h(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)}, \tag{23}$$

where

$$\begin{aligned} &h(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) \\ &\qquad = Y_1(Y_2 - Y_3 + Y_5 - Y_6)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\ &\qquad + Y_2(Y_1 + Y_3 + Y_4 + Y_6)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\ &\qquad + Y_3(Y_1 + Y_2 + Y_4 + Y_5)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\ &\qquad + Y_4(Y_2 - Y_3 + Y_5 - Y_6)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\ &\qquad + Y_5(Y_1 + Y_3 + Y_4 + Y_6)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\ &\qquad + Y_6(Y_1 + Y_2 + Y_4 + Y_5)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 . \end{aligned}$$

These four test statistics will be used to define the tail set for the E and M p-values. Other test statistics are possible and we suggest three, $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$, which are defined in the same way as they were for the independent binomial proportions (Section 3), but with the multinomial distribution with six parameters substituted for the joint distribution of two independent binomial distributions, that is,

$$T_{\pi_{\text{e}}}(\boldsymbol{y}_{\text{obs}}) = \pi(\boldsymbol{y}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}), \tag{24}$$

$$T_{\pi_{\text{E}}}(\boldsymbol{y}_{\text{obs}}) = \Pr(\pi(\boldsymbol{Y}; \tilde{\boldsymbol{p}}_{\text{obs}}) \leq \pi(\boldsymbol{y}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) \tag{25}$$

and

$$T_{\pi_{\text{M}}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(\pi(\boldsymbol{Y}; \boldsymbol{p}) \leq \pi(\boldsymbol{y}_{\text{obs}}; \boldsymbol{p}); \boldsymbol{p}). \tag{26}$$

Finally, we also consider the Bayesian prior predictive p-values, that in addition to being p-values in their own right, can be used as test statistics to define the critical region for the E and M p-values, and we denote them T_{PP} where

$$T_{\text{PP}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p}) \cdot \pi(\boldsymbol{p} \mid H_0) \mathrm{d}\boldsymbol{p}. \tag{27}$$

4.2 Results

In the PPV setting, we have studied the performance of the different types of p-values with respect to test statistics and the parameters in the multinomial distribution. The performance will be evaluated in terms of test size and test power, which are calculated as given by (5) and (6). We choose the significance level $\alpha = 0.05$.

4.2.1 Evaluation of test size

The test statistics considered were the LAP, likelihood ratio, unrestricted difference and restricted difference test statistics, T_{LAP}, T_{LR}, T_{uDT} and T_{rDT}. In addition we used the π_{e}-probabilities and the π_{E}, π_{M} and Bayesian p-values as test statistics, i.e. $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{M}}}$ and T_{PP}. For each of these test statistics and for the chosen values of N we calculated the E, M, E+M, E^2 and E^{2}M p-values. We also considered the performance of the Bayesian p-values as p-values in itself.

The performance of the test statistics can depend highly on the parameters $\boldsymbol{p}$ in the multinomial distribution. Both the overall mean performance as well as the performance for specific values are evaluated. For the mean performance a set of 10385 values of $\boldsymbol{p}$ in $\mathcal{P}_0$ is used, which are obtained by using a four dimensional grid for the four free parameters where each side in the grid is divided into 30 subintervals, and the 10385 values of $\boldsymbol{p}$ in the grid that belong to $\mathcal{P}_0$ are then the cases we consider. For this set of cases we calculate the mean test size, i.e. we calculate the test size from (5) for each case and then find the average for the 10385 cases. In addition, six specific cases of $\boldsymbol{p}$ in $\mathcal{P}_0$ are evaluated, see Table 10. These are the same cases as in Günther, Bakke and Langaas (2009) where the reasoning for choosing these values can be found.

Case	p_1	p_2	p_3	p_4	p_5	p_6
1	0.068	0.135	0.135	0.527	0.068	0.068
2	0.043	0.130	0.130	0.348	0.174	0.174
3	0.267	0.267	0.267	0.067	0.067	0.067
4	0.300	0.267	0.267	0.033	0.067	0.067
5	0.400	0.200	0.200	0.100	0.050	0.050
6	0.450	0.200	0.200	0.050	0.050	0.050

Table 10: Specification of multinomial parameters under H_0.

The size of the multinomial sample determines how many possible outcomes there are and is an interesting factor to consider. We want to investigate whether the performance of the test statistics depends on sample size, in particular for small sample sizes, so we use $N = 10, 15, 20, 25$.

When maximizing the p-values over $\boldsymbol{p}$ in $\mathcal{P}_0$ we used a four-dimensional grid since there are four free parameters, with 50 points on each side. In addition, the maximum likelihood estimates for all possible outcomes given N were included in the grid. The Bayesian p-values were calculated on the grid with 50 points in each side, for further details see Section 5.

Table 11 shows the mean test size for all the test statistics, values of N and type of p-values investigated. We first compare the performance of the different types of p-values, E, M, E+M, E^2 and E^{2}M.

The M p-values yield the smallest test size for all values of N for all the test statistics except for the likelihood ratio test when $N = 20$, there the E^{2}M p-values yields the smallest test size. In general the E and E^2 p-values result in larger test sizes than the E+M and E^{2}M p-values which we would expect since the E and E^2 p-values are not valid, whereas the E+M and E^{2}M p-values are. The exception is T_{uDT}, when $N = 10$, the E p-values yield smaller test size than the E+M and E^{2}M p-values, and when $N = 15$, the E p-values yield smaller test size than the E^{2}M p-values. The E^2 p-values yield larger test sizes than the E p-values, except for $T_{\pi_{\text{e}}}$.

Next we compare the performance of the different test statistics. First we consider the test statistics that originated from large samples where their asymptotic distributions were utilized, i.e. the LAP,

N	p-value	LRT	LAP	uDT	rDT	π_{M}	π_{E}	π_{e}	PP$_1$	PP$_3$
10	M	0.0164	0.0092	0.0019	0.0130	0.0177	0.0088	0.0145	0.0104	0.0190
10	E	0.0381	0.0282	0.0179	0.0352	0.0388	0.0359	0.0643	0.0330	0.0366
10	E+M	0.0285	0.0198	0.0181	0.0285	0.0286	0.0242	0.0207	0.0257	0.0297
10	E^2	0.0456	0.0420	0.0395	0.0439	0.0435	0.0441	0.0549	0.0441	0.0435
10	E^{2}M	0.0273	0.0247	0.0220	0.0268	0.0279	0.0213	0.0260	0.0297	0.0299
15	M	0.0256	0.0122	0.0006	0.0242	0.0227	0.0085	0.0138	0.0061	0.0150
15	E	0.0451	0.0395	0.0297	0.0447	0.0438	0.0376	0.0650	0.0395	0.0428
15	E+M	0.0354	0.0243	0.0234	0.0354	0.0364	0.0317	0.0274	0.0339	0.0360
15	E^2	0.0470	0.0466	0.0462	0.0470	0.0479	0.0453	0.0538	0.0472	0.0480
15	E^{2}M	0.0332	0.0303	0.0302	0.0350	0.0355	0.0305	0.0311	0.0374	0.0370
20	M	0.0333	0.0140	0.0002	0.0277	0.0239	0.0091	0.0130	0.0024	0.0088
20	E	0.0469	0.0430	0.0365	0.0475	0.0468	0.0394	0.0640	0.0433	0.0458
20	E+M	0.0395	0.0308	0.0283	0.0386	0.0392	0.0337	0.0317	0.0369	0.0378
20	E^2	0.0488	0.0477	0.0492	0.0482	0.0490	0.0465	0.0544	0.0491	0.0491
20	E^{2}M	0.0319	0.0331	0.0329	0.0339	0.0365	0.0341	0.0324	0.0381	0.0388
25	M	0.0335	0.0124	0.0001	0.0136	0.0214	0.0089	0.0131	0.0009	0.0032
25	E	0.0483	0.0449	0.0403	0.0486	0.0479	0.0405	0.0636	0.0449	0.0469
25	E+M	0.0385	0.0324	0.0313	0.0403	0.0403	0.0314	0.0343	0.0402	0.0398
25	E^2	0.0492	0.0479	0.0496	0.0490	0.0497	0.0482	0.0540	0.0494	0.0494
25	E^{2}M	0.0359	0.0353	0.0343	0.0360	0.0362	0.0374	0.0349	0.0381	0.0379

TABLE 11: Mean test size for the 10385 values of $\boldsymbol{p}$ in $\mathcal{P}_0$, for all the test statistics and M, E, E+M, E^2 and E^{2}M p-values when the chosen significance level is $\alpha = 0.05$. The top row denotes the test statistics, LRT is given in (21), LAP in (23), uDT and rDT in (22) with unrestricted and restricted maximum likelihood estimates inserted for $\boldsymbol{p}$ respectively, π_{M} in (26), π_{E} in (25), π_{e} in (24), PP$_1$ in (15) and PP$_3$ in (20). N is the sample size.

likelihood ratio, unrestricted difference and restricted difference test statistics. In general, for all types of p-values, T_{LR} and T_{rDT} have the largest test size, T_{uDT} and T_{LAP} have the smallest test size and T_{uDT} has mostly smaller test size than T_{LAP}. T_{LR} has the largest test size for the M p-values. When $N = 20$, T_{uDT} has the largest test size for the E^2 p-values, a result for which we have found no apparent reason.

If we look into the performance of $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$ we see that $T_{\pi_{\text{e}}}$ has the largest test size compared to all the other test statistics for the E and E^2 p-values for all N. $T_{\pi_{\text{M}}}$ has the largest test size for the M, E+M and E^{2}M p-values for $N = 10$, for the E+M and E^{2}M p-values when $N = 15$ and for the E^{2}M p-values when $N = 20$, whereas $T_{\pi_{\text{E}}}$ has the largest test size for the E^{2}M p-values when $N = 25$. We note that when N increases the likelihood ratio or restricted difference test performs better than the π_{M} statistic for the M, E+M and E^{2}M p-values, therefore the π_{M} test statistic is probably a better choice only when N is small.

What is worth noting, is that for the test statistics that are most conservative with respect to test size for the M p-values, the gain is greater when performing one or more E step(s) before the M step compared to the test statistics for which the test size for the M p-values is less conservative. This is particularly evident for T_{uDT}, T_{LAP} and $T_{\pi_{\text{E}}}$ compared to T_{LR}. The test size for T_{LR} increases less than the test size for the other three test statistics when comparing the M and E+M p-values. For T_{LR} the test size is also reduced if two E steps instead of one are applied before the M step, whereas for T_{LAP} and T_{uDT} the test size increases when two E steps are applied before the M step.

The mean test size increases when N increases. Comparing the test sizes for $N = 10$ to the test sizes for $N = 25$ reveals an increase for all test statistics and type of p-values except for some of the M p-values which have test size that is approximately 0. As an illustration, the mean test size for the M p-value for the likelihood ratio test statistic is 0.0164 when $N = 10$ and 0.0335 when $N = 25$.

In addition to the test statistics discussed so far, the Bayesian prior predictive p-values $P_{\text{PP},1}$ and $P_{\text{PP},3}$, originated from using the same prior $\pi_1(\boldsymbol{p})$, but different formulations of the null hypothesis, were used as test statistics to compute E, M, E+M, E^2 and E^{2}M p-values. When $N = 10$, the PP$_3$ test statistic yields larger test size than all the other test statistics for the M, E+M and E^{2}M p-values. Otherwise the test size of these two test statistics lies between the test size of the other test statistics, not following a clear pattern, except that the PP$_3$ yields larger test size than the PP$_1$ in general.

We also evaluated the performance of all the test statistics, values of N and types of p-values for the six multinomial cases of Table 10. Table 12 shows the test size for T_{LR} for $N = 10$ and $N = 25$. We see that the E and E^2 p-values yield a test size greater than 0.05 in case 1–5 for $N = 25$ and thus proves that these p-values are not valid. We also see that the test size is greater when $N = 25$ compared to $N = 10$. The results for the other test statistics and values of N are omitted in this report since the findings in respect to test statistics and p-values in the six specific multinomial cases were similar to the overall findings, however the test size for all test statistics was clearly dependent of the chosen multinomial cases, i.e. the parameter $\boldsymbol{p}$ in the multinomial distribution. In general, which can also be seen in Table 12, case 1 and 2 have larger test size than case 3–6. This trend was consistent through the different test statistics, types of p-values and N and indicates that when comparing test sizes the multinomial case chosen will have a large influence the test size, but it will not change the conclusions with respect to which test statistic or which p-value results in the largest or smallest test size.

Figure 3 shows histograms for the test size in the 10385 cases under H_0 for the M, E, E+M, E^2 and E^{2}M p-values for each of the test statistics T_{LAP}, T_{LR}, T_{uDT}, T_{rDT}, $T_{\pi_{\text{M}}}$, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{e}}}$ and $T_{\text{PP},1}$ for $N = 10$. We see that for T_{LR}, T_{rDT}, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{M}}}$,and $T_{\text{PP},1}$ the distribution of the test size for the E

N	p-value	case 1	case 2	case 3	case 4	case 5	case 6
10	M	0.0199	0.0193	0.0097	0.0071	0.0061	0.0035
10	E	0.0412	0.0426	0.0281	0.0218	0.0197	0.0122
10	EM	0.0281	0.0366	0.0242	0.0193	0.0170	0.0111
10	E^2	0.0475	0.0500	0.0379	0.0302	0.0333	0.0220
10	E^{2}M	0.0322	0.0294	0.0200	0.0157	0.0166	0.0101
25	M	0.0431	0.0428	0.0386	0.0391	0.0313	0.0287
25	E	0.0528	0.0529	0.0573	0.0563	0.0495	0.0441
25	EM	0.0431	0.0435	0.0458	0.0448	0.0389	0.0339
25	E^2	0.0510	0.0514	0.0573	0.0569	0.0523	0.0469
25	E^{2}M	0.0401	0.0387	0.0415	0.0404	0.0367	0.0322

TABLE 12: Test size for the likelihood ratio test statistic for the six multinomical cases.

p-values is skewed towards the right compared to the distribution for the M p-values, and we note that the test size is sometimes larger than 0.05, showing that the E p-values are not valid. The E+M p-values preserve the skewed distribution while shifting it to the left so that no test size is greater than 0.05. For the LAP and uDT test statistics, we note that the distribution of test size for the E p-values is not skewed in the same way, but the E^2 p-values are, so apparently it is necessary to do two E steps before maximization for the LAP and uDT statistics.

Figure 3 illustrates what happens under the E and M steps. To obtain an even better understanding of the effect of the E and M steps, we consider two possible outcomes when $N = 10$, $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$. Table 13 shows the p-values for these outcomes using the likelihood ratio and LAP test statistics.

Outcome	Test statistic	M	E	E+M	E^2	E^{2}M
$\boldsymbol{y}_1$	$T_{\text{LRT}} = 4.159$	0.1025	0.0940	0.1108	0.0770	0.1062
$\boldsymbol{y}_2$	$T_{\text{LRT}} = 4.077$	0.1025	0.0349	0.0450	0.0279	0.0408
$\boldsymbol{y}_1$	$T_{\text{LAP}} = 3.932$	0.1048	0.0805	0.1056	0.0768	0.1064
$\boldsymbol{y}_2$	$T_{\text{LAP}} = 2.492$	0.2297	0.0978	0.1329	0.0654	0.0855

TABLE 13: P-values for the likelihood ratio and LAP test statistics for the outcomes $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$.

Let us first consider the p-values for the likelihood ratio test statistic. We note that for $\boldsymbol{y}_2$ with a 5% significance level, we would reject the null hypothesis based on the E p-value and not reject it based on the M p-value. Since the likelihood ratio test statistic is greater for $\boldsymbol{y}_1$ than for $\boldsymbol{y}_2$, the M p-value for $\boldsymbol{y}_2$ will necessarily be greater than for $\boldsymbol{y}_1$, which will be greater than the E p-value for $\boldsymbol{y}_1$. Since the E p-value is less for $\boldsymbol{y}_2$ than for $\boldsymbol{y}_1$, $\boldsymbol{y}_1$ will not be in the tail set for $\boldsymbol{y}_2$ when performing the M step after the E step and the E+M p-value results in rejection of the null hypothesis on a 5% significance level for $\boldsymbol{y}_2$. Since the E and E^2 steps alone do not result in valid p-values, we should perform an M step afterwards. But as we see, the E step(s) are means to avoid certain outcomes having a large p-value because of other outcomes having greater test statistics and artifically large E and M p-values compared to other outcomes with similar magnitude of the test statistics.

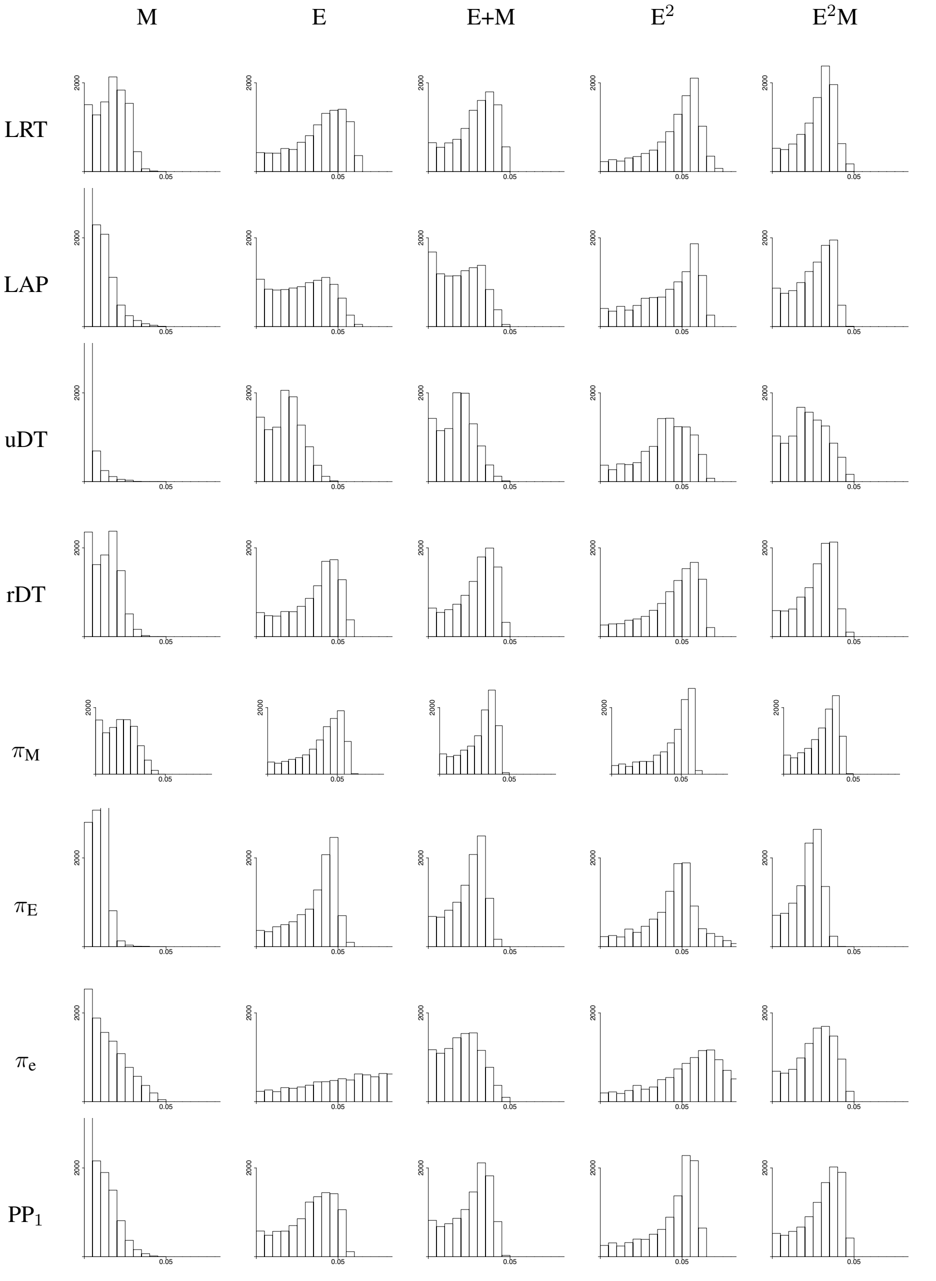


FIGURE 3: Distribution of test size for the various test statistics and p-values, $N = 10$, the x-axis is cut at 0.08 and the y-axis at 3000.

For the LAP test statistic, we do not see the same effect, even though the p-value with the smallest test statistic, $\boldsymbol{y}_2$, has an M p-value greater than the M p-value for $\boldsymbol{y}_1$. Here the E p-value for $\boldsymbol{y}_2$ is also greater than the one for $\boldsymbol{y}_1$, and thus this ordering is preserved when performing an M step after the E step. The E^2 p-value however, is smaller for $\boldsymbol{y}_2$ than $\boldsymbol{y}_1$, and performing the M step afterwards does not change this.

Here $\boldsymbol{y}_1$ is an example of an outcome for which the decision of rejecting the null hypothesis does not only depend on the type of p-value, but also of the chosen test statistic. Using the likelihood ratio test statistic, we reject the null hypothesis on a 5% confidence level for all of the p-values E, E+M, E^2 and E^{2}M. If we use the LAP test statistic instead, we do not reject it for any of the p-values.

Table 14 shows the mean test size for the Bayesian prior predictive p-values for $N = 10, 15, 20, 25$ using a grid with 50 points in each direction for both formulations of H_0, i.e. $f_{\text{PPV}}(\boldsymbol{p})$ and $f^*_{\text{PPV}}(\boldsymbol{p})$ and both priors for $\boldsymbol{p}$. We see that the test size depends highly on the choice of prior and formulation of H_0. The test size is smallest using the uniform Dirichlet prior and $f^*_{\text{PPV}}(\boldsymbol{p}) = 0$ as H_0, it increases to around 0.055 with $f_{\text{PPV}}(\boldsymbol{p}) = 0$ as H_0 and if we choose the non-uniform Dirichlet prior $\pi_2(\boldsymbol{p})$, the test size becomes very high. Clearly, the non-uniform prior is not a good choice and it also indicates that the choice of prior has a larger effect than how we choose to formulate the null hypothesis. Comparing these results to the results when using the prior predictive p-values as test statistics to define the tail sets for the M, E, E+M and E^{2}M p-values in Table 11 shows that the M step reduces the test size in all cases for all values of N and for both formulations of H_0 as expected. The test size for the E^2 p-values is higher than for the Bayesian p-values in many of the cases and in some cases, e.g. case 5 for $N = 15, 20, 25$ for $H_0\colon f_{\text{PPV}}(\boldsymbol{p}) = 0$ and for $N = 20, 25$ for $H_0\colon f^*_{\text{PPV}}(\boldsymbol{p}) = 0$, the test size increases for all the p-values except the M p-values. Table 14 also shows that the Bayesian prior predictive p-values are not valid since the test size is larger than the significance level.

N	PP$_1$	PP$_2$	PP$_3$
10	0.0561	0.1272	0.0489
15	0.0557	0.1465	0.0491
20	0.0556	0.1533	0.0494
25	0.0552	0.1556	0.0494

TABLE 14: Test size for the Bayesian positive predictive p-values using different priors, formulation of H_0 and values of N, PP$_1$ is given in (15), PP$_2$ is given in (17) and PP$_3$ is given in (20).

The prior predictive p-values for the two outcomes $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$ are given in Table 15. We see that the three Bayesian p-values are quite different for both outcomes. For $\boldsymbol{y}_2$ we reject the null hypothesis, whereas for $\boldsymbol{y}_1$ we do not reject the null hypothesis. The two p-values both found from the model with uniform Dirichlet prior are similar for $\boldsymbol{y}_2$, but for $\boldsymbol{y}_1$ it is the two p-values that are based on the same formulation of H_0 that are similar. The null hypothesis is rejected for $\boldsymbol{y}_2$, but not for $\boldsymbol{y}_1$ for any of the p-values.

4.2.2 EVALUATION OF TEST POWER

We would like to compare the test power of T_{LR}, T_{LAP}, T_{uDT}, T_{rDT} and $T_{\pi_{\text{M}}}$. Since the results of the test size comparisons showed that the M p-values have the smallest test sizes and since the E and E^2 p-values are not valid, we consider only the E+M and E^{2}M p-values when comparing test power. We

Outcome	PP_1	PP_2	PP_3
$\boldsymbol{y}_1$	0.1086	0.1084	0.0506
$\boldsymbol{y}_2$	0.0382	0.0171	0.0323

TABLE 15: Bayesian prior predictive p-values for the outcomes $\boldsymbol{y}_1 = (1,3,0,6,0,0)$ and $\boldsymbol{y}_2 = (1,0,1,3,5,0)$.

expect that the power increases with N and we used $N = 10$ and $N = 25$ to investigate the magnitude of the increase. The power was calculated the same way as the test size except that the values of $\boldsymbol{p}$ are chosen so that $\boldsymbol{p}$ does not satisfy the null hypothesis (9).

We wanted to compare the test power in specific multinomial cases and we chose six sets of the parameters $\boldsymbol{p}$, these were denoted case 7–12 and are given in Table 16. They were chosen because of their decreasing distance from H_0 which is measured by the magnitude of $f_{\text{PPV}}(\boldsymbol{p})$. If $f_{\text{PPV}}(\boldsymbol{p})$ is close to 0, then $\boldsymbol{p}$ nearly satisfies H_0 while the greater $|f_{\text{PPV}}(\boldsymbol{p})|$ is, the further away from H_0 $\boldsymbol{p}$ is. Since the power in our chosen cases may not be representative for a randomly chosen case, we also generate 10385 random cases under H_1, by drawing 10385 vectors of length 6 from the uniform distribution and scaling each vector to sum to 1.

Case	p_1	p_2	p_3	p_4	p_5	p_6	$f_{\text{PPV}}(\boldsymbol{p})$
7	0.06	0.01	0.44	0.26	0.22	0.01	0.52
8	0.01	0.10	0.44	0.01	0.43	0.01	0.76
9	0.20	0.05	0.24	0.28	0.22	0.01	0.27
10	0.01	0.07	0.27	0.28	0.26	0.11	0.29
11	0.06	0.12	0.18	0.14	0.35	0.15	0.18
12	0.17	0.12	0.18	0.21	0.16	0.16	0.05

TABLE 16: Specification of cases under H_1.

Table 17 and 18 shows the test power for the chosen cases, test statistics and p-values when $N = 10$ and $N = 25$ respectively. As expected the power increases when N increases. The test statistics T_{uDT} and T_{LAP} have the smallest power except for the E^{2}M p-values in case 6 when $N = 25$. When $N = 25$ the $T_{\pi_{\text{M}}}$ statistic has the highest power except in case 6 for the E^{2}M p-values. For $N = 10$, the T_{LR} statistic has highest power for the E+M p-values in four of six cases, while only in one case for the E^{2}M p-values. The E+M p-values yields in general higher power than the E^{2}M p-values, except for the T_{LAP} and T_{uDT} statistics when $N = 10$. If we compare these results to the calculated mean power for all the power cases, given in the last column of Table 17 and 18, we see that T_{LAP} and T_{uDT} have smaller power for the E+M than the E^{2}M p-values when $N = 10$ and also when $N = 25$ for T_{uDT}. T_{LR} has the largest power for the E+M p-values when $N = 10$, otherwise the π_{M} has the largest test power.

When comparing the power in each of the six cases by considering the value of $f_{\text{PPV}}(\boldsymbol{p})$ in Table 16 we see that the power seems to decrease when $f_{\text{PPV}}(\boldsymbol{p})$ decreases which we would expect since in cases that are far from H_0 the test should have higher power than in cases closer to H_0. However, in case 7 and 8 $f_{\text{PPV}}(\boldsymbol{p})$ is 0.52 and 0.76 respectively and yet case 7 has the highest power, particularly

Test statistic	p-value	case 7	case 8	case 9	case 10	case 11	case 12	mean
T_{LRT}	E+M	0.8344	0.7210	0.4125	0.2332	0.1161	0.0547	0.1064
T_{LAP}	E+M	0.7165	0.7203	0.2909	0.1863	0.0801	0.0360	0.0810
T_{uDT}	E+M	0.6997	0.3815	0.3269	0.1446	0.0556	0.0345	0.0697
T_{rDT}	E+M	0.8372	0.7173	0.4096	0.2286	0.1108	0.0529	0.1044
$T_{\pi_{\text{M}}}$	E+M	0.8271	0.7568	0.3996	0.2119	0.1080	0.0472	0.1016
T_{LRT}	E^{2}M	0.8219	0.7159	0.4140	0.2020	0.1006	0.0463	0.0967
T_{LAP}	E^{2}M	0.7492	0.7044	0.3498	0.1857	0.0848	0.0416	0.0881
T_{uDT}	E^{2}M	0.7640	0.3903	0.3712	0.1923	0.0755	0.0442	0.0844
T_{rDT}	E^{2}M	0.8240	0.7160	0.4081	0.2162	0.1028	0.0476	0.0977
$T_{\pi_{\text{M}}}$	E^{2}M	0.8274	0.7500	0.4112	0.2031	0.1060	0.0467	0.0994

TABLE 17: Test power for the E and E^{2}M p-values in case 7–12 and mean over 10385 cases for $N = 10$.

Test statistic	p-value	case 7	case 8	case 9	case 10	case 11	case 12	mean
T_{LRT}	E+M	0.9979	0.9967	0.8014	0.4974	0.1936	0.0537	0.2163
T_{LAP}	E+M	0.9967	0.9935	0.7897	0.4713	0.1686	0.0509	0.2038
T_{uDT}	E+M	0.9970	0.9592	0.8056	0.4788	0.1699	0.0547	0.2075
T_{rDT}	E+M	0.9985	0.9976	0.8179	0.5275	0.2193	0.0585	0.2241
$T_{\pi_{\text{M}}}$	E+M	0.9987	0.9978	0.8296	0.5378	0.2257	0.0586	0.2242
T_{LRT}	E^{2}M	0.9976	0.9963	0.7908	0.4793	0.1835	0.0499	0.2080
T_{LAP}	E^{2}M	0.9961	0.9935	0.7866	0.4609	0.1691	0.0515	0.2063
T_{uDT}	E^{2}M	0.9969	0.9622	0.7874	0.4660	0.1704	0.0504	0.2068
T_{rDT}	E^{2}M	0.9980	0.9969	0.7938	0.5017	0.2005	0.0501	0.2087
$T_{\pi_{\text{M}}}$	E^{2}M	0.9983	0.9969	0.8091	0.5080	0.2070	0.0510	0.2101

TABLE 18: Test power for the E and E^{2}M p-values in case 7–12 and mean over 10385 cases for $N = 25$.

when $N = 10$. We see the same in case 9 and 10, $f_{\text{PPV}}(\boldsymbol{p})$ is then 0.27 and 0.29, and the power in case 9 is a lot higher than in case 10.

The mean value of $|f_{\text{PPV}}(\boldsymbol{p})|$ for the 10385 cases is 0.13, which explains the small overall power, since the mean value is not as far from H_0 as e.g. case 7 or 8. The mean power is comparable to case 11 where the distance from H_0 is 0.18.

It is not surprising that the likelihood ratio, restricted difference and π_M test statistics perform similarly, considering they are all functions of the maximum likelihood estimates for $\boldsymbol{p}$ under H_0, $\tilde{\boldsymbol{p}}$. The LAP and unrestricted difference test statistics however, do not depend on these estimates and this can be the reason their performance is poorer.

5 Computational details

To compute the integral (14), we used the midpoint rule on a 4-dimensional grid. The four dimensions correspond to p_1, p_2, p_3 and p_5. Each side in the grid is divided into a number of subintervals of equal length, and the midpoint in each subinterval is calculated. For each point (p_1, p_2, p_3, p_5) in the grid, we set p_4,

$$p_4 = \frac{p_1(1-p_1-2p_2-p_3-2p_5)+p_2(1-p_2-p_3-p_5)-p_3p_5}{p_1+p_3}$$

which is derived from (13).

If $0 < p_4 < 1$, we set $p_6 = 1 - \sum_{i=1}^{5} p_i$ and if $0 < p_6 < 1$ then the value

$$N!\left(\prod_{i=1}^{6}\frac{p_i^{y_i}}{y_i!}\right)\frac{(p_1+p_2+p_4+p_5)^2}{p_1+p_2}, \tag{28}$$

which is $\pi(\boldsymbol{y}|\boldsymbol{p})$ multiplied with the non-normalized density of $\boldsymbol{p}$, is added to the present value of the integral. If either p_4 or p_6 are less than 0 or greater than 1, the current point in the grid is discarded. The total non-normalized integral is the sum of (28) over the $\boldsymbol{p}$'s satisfying the constraints for p_4 and p_6. The integrals (16) and (19) are computed similarly.

The number of points in the grid has to be chosen and in the results presented in this report, a grid where each side is divided into 50 subintervals was used. This resulted in 79876 points after discarding those with p_4 or p_6 outside [0,1]. Table 19 shows the test size for the Bayesian prior predictive values using the uniform Dirichlet prior and original formulation of H_0 (9) for the six values of $\boldsymbol{p}$ given in Table 10 and $N = 10$ when the number of subintervals, n_{int}, on each side in the grid is 30, 35, 40, 45 and 50. We see that the test size varies with the grid size to some extent, the largest difference is in case 1 between $n_{\text{int}} = 45$ and $n_{\text{int}} = 50$.

n_{int}	case 1	case 2	case 3	case 4	case 5	case 6
30	0.0395	0.0642	0.0401	0.0365	0.0186	0.0157
35	0.0389	0.0632	0.0405	0.0367	0.0191	0.0159
40	0.0384	0.0620	0.0405	0.0367	0.0191	0.0159
45	0.0384	0.0620	0.0405	0.0367	0.0191	0.0159
50	0.0407	0.0625	0.0408	0.0372	0.0193	0.0162

TABLE 19: Test size in case 1–6 for the Bayesian prior predictive p-value in (15) for $N = 10$ using different grid sizes, n_{int} is the number of sub intervals on each of the four sides in the grid.

A grid for $\boldsymbol{p}$ is also needed for the p-values that include a maximization step, i.e. the M, E+M and E^{2}M p-values. In the positive predictive value setting, we used the same grid as for the Bayesian prior predictive p-values with 50 possible subintervals for each of the four sides in the grid, but in addition we included the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ of $\boldsymbol{p}$ under H_0 for all possible outcomes given N. Table 20 shows the number of possible outcomes in the positive predictive value situation for a given value of N and the size of the grid with the maximum likelihood estimates included. Thus, the size of this grid increased with N, for $N = 10$, the grid consisted of $3003 + 79876 = 82879$ points and when $N = 25$, it consisted of $142506 + 79876 = 222382$ points. Comparisons of the test size for different grid sizes showed that the grid did not have a great influence on the test size when the

N	Number of outcomes	Size of grid
10	3003	82879
15	15504	95380
20	53130	133006
25	142506	222382

TABLE 20: Number of possible outcomes given N and size of grid used when calculating p-values in the problem of comparing positive predictive values.

grid is used for maximization. We also investigated how often the maximum p-value was obtained in one of maximum likelihood points compared to the other points. The percentage increased with N and decreased with the size of the grid without maximum likelihood estimates. When N increases the number of maximum likelihood estimates increases, and it is not surprising that more of these points will give the maximum p-value and similarly, when the number of grid points in the grid without maximum likelihood estimates increases, more of the points that are not maximum likelihood estimates will give the maximum p-value.

The p-value computations for a sequence of E and M steps are quite computer intensive, as p-values for all outcomes (except in the last step), not only the one of interest in a specific study, must be computed for further use as a test statistic in the next step. The test statistic giving the original ordering of outcomes, e.g. the likelihood ratio test statistic, should be computed only once, as should the maximum likelihood estimates of $\boldsymbol{p}$ under the null hypothesis. The grid used for the numerical maximization in the M step and for calculation of the π_{M} statistic was also calculated in advance.

In both the E and the M step, the outcomes should be sorted according to the test statistic (original test statistic or negative output of a previous E or M step). In the E step, the p-values are then accumulated, starting with the probability of the outcome having the most extreme value of the test statistic, and the probabilities (with the maximum likelihood estimates of $\boldsymbol{p}$ under the null hypothesis of the outcome of interest as parameters) of the forthcoming outcomes successively being added until the outcome of interest is reached. Special care must be taken to include possible outcomes having an equal test statistic value ("draws"), and because of possible numerical inaccuracies also a threshold for when two values are counted as equal should be specified. In order to compute all possible p-values, this should be repeated for all outcomes – we have chosen to accumulate probabilities for all outcomes in parallel. Taking care when dealing with draws also applies to the M step and calculation of the π_{E} and π_{M} test statistics.

In the M step we accumulated probabilities given by the grid points as parameters in parallel while going through the sorted outcomes. As the number of grid points times the number of outcomes may be huge, only the accumulated probabilities for each outcome were saved, and for each outcome reached, the maximum of the accumulated probabilities were saved as the p-value of that outcome.

Calculation of the π_{E} and π_{M} test statistics, based on the probabilities of the outcomes themselves instead of on an external test statistic, are more computer intensive, as the ordering of the outcomes is specific for each outcome of interest, and not to a given test statistic. For π_{E}, the p-value for an outcome of interest is found by adding probabilities of all outcomes having a probability that is not greater, using the maximum likelihood estimate of $\boldsymbol{p}$ under the null hypothesis of the outcome of interest as parameters, thus the probability of each outcome has to be calculated for each outcome of interest. We found some gain in computation speed by sorting the outcomes before adding.

N	E	M	π_{e}	π_{E}	π_{M}
10	0m0.33s	0m17.69s	0m0.06s	0m2.84s	1m29.94s
25	16m17.13s	14m4.51s	0m0.96s	155m13.68s	101m20.51s
10	0m0.34s	0m18.65s	0m0.01s	0m2.86s	1m33.01s
25	15m51.13s	39m17.33s	0m0.97s	156m40.3s	262m6.08s

TABLE 21: Running time for E and M p-values and for calculating the test statistics $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi\mathrm{M}}$ in the positive predictive values setting for samples sizes $N = 10, 25$ (3003 and 142506 outcomes, respectively). The two upper rows show the time when using a grid with $n_{\text{int}} = 50$ without maximum likelihood estimates and the time in the two lower rows is the time when using the grid with $n_{\text{int}} = 50$ including maximum likelihood estimates (79876 points without estimates, 82879 including estimates for $N = 10$, and 222382 points including estimates for $N = 25$).

For π_{M}, the grid points rather than the outcomes were gone through in an outer loop. For each grid point, the probability of each outcome was calculated, the outcomes sorted accordingly, and probabilities accumulated from the smallest to the greatest. If the cumulative probability of an outcome was greater than some earlier maximum for that outcome, the maximum was replaced by the current sum.

In contrast, calculation of π_{e} is trivial, this is simply the probability of an outcome taking its maximum likelihood estimate of $\boldsymbol{p}$ under the null hypothesis as the parameter vector.

Power and size calculations for a given parameter vector are simply a matter of adding probabilities of outcomes having p-values not exceeding the significance level (in our case 0.05).

The code was written in C++, implemented in GCC and the calculations were performed with the Standard Template Library, using one of eight processors on a Dell PowerEdge 2950 with two Quad-core Xeon X5365 3.0 GHz processors, 4 MB cache, 16 GB RAM. The running time for calculating E and M p-values for any test statistic, along with the running time for calculating the π_{e}, π_{E} and π_{M} test statistics when comparing positive predictive values for $N = 10$ and $N = 25$ are given in Table 21 for the grid with $n_{\text{int}} = 50$, without and with the maximum likelihood estimates of $\boldsymbol{p}$ included. When $N = 10$, all the calculations are performed rather fast, except calculating the values of the π_{M} test statistic which takes one and a half minute. When N increases, the running time naturally increases severely since all calculations must be performed for all possible outcomes. We note that calculating the π_{E} test statistic takes longer than calculating the π_{M} test statistic when $N = 25$ for the grid without maximum likelihood estimates. This is because the number of possible outcomes is less than the number of grid points in this case. If the number of grid points is larger than the number of outcomes, as in the grid where the maximum likelihood estimates are included, calculating the π_{M} statistic takes much longer than calculating the π_{E} statistic.

6 DISCUSSION

The enumeration idea is not new as it goes back to Fisher (1935), but it has often been overlooked. We have demonstrated how to apply the idea for testing independent binomial proportions and comparing positive predictive values. Another recent application of the idea is in genome-wide association studies, in which single nucleotide polymorphisms (SNP) across the human genome are studied. When the mode of inheritance is unknown, the MAX test statistic, which is the maximum of the three Cochrane–

Armitage trend statistics for dominant, recessive and additive inheritance modes, see Freidlin, Zheng, Li and Gastwirth (2002), tests the association between the genotype and phenotype. The exact distribution of the MAX test statistic is unknown and calculating p-values based on proposed asymptotic distributions involves numerical integration. Another common approach is to use permutations tests, but both solutions leads to possible random errors in the calculated p-values. Moldovan, Langaas and Bahlo (2009) instead calculate exact p-values using the enumeration approach and thereby avoid this uncertainty.

When the sample size increases and enumeration will be too time consuming, the parametric bootstrapping approach can be used instead. Günther et al. (2009) used parametric bootstrapping to approximate the distribution of the likelihood ratio, LAP and restricted and unrestricted difference test statistics. The p-values obtained from this distribution are approximately the same as the E p-values we find by enumeration in this report, and the parametric bootstrap approach involving simulated outcomes is actually a numerical approximation that calculates the tail without using enumeration. This is seen if the test size for case 1–6 in Table 12 is compared to the test size for the small sample parametric bootstrap likelihood ratio test in Table 3 of Günther et al. (2009) – the values are almost the same. It may be of use for larger sample sizes when calculating maximum likelihood estimates and p-values for the bootstrap samples is less time consuming than calculating the maximum likelihood estimates and p-values for all possible outcomes. When using the formulas for calculating exact test size and power, i.e., (5) and (6), drawing outcomes from the multinomial distribution under H_0 or H_1 and estimating the test size or power by the proportion of these outcomes having p-values less than or equal to the significance level as was done in Günther et al. (2009) is not necessary, and therefore the uncertainty in the estimates are removed. This is however, only possible when the sample size is small enough so that the p-values for all possible outcomes can be calculated.

Another option when the sample size increases is to condition on sums of $N_i,\ i = 1, \ldots, 6$, which in a contingency table setting corresponds to conditioning on the marginals. This reduces the number of possible outcomes and makes it possible to use exact tests for higher values of N. The usability of this approach depends on the actual problem. In the example from Lloyd (2008), n_1 and n_2 are fixed as the number of subjects who receives treatment and placebo respectively. In the setting of positive predictive values, it is not clear which values that should be fixed. It could be the number of diseased and non-diseased subjects, if the disease status is decided before the two tests are applied, or it could be the number of subjects with positive test A, positive test B and positive tests A and B, but in practise, these numbers will usually not be fixed in advance.

As Table 12 showed, the test size of a test statistic for any p-value depends on the chosen value of $\boldsymbol{p}$, the parameter in the multinomial distribution. When the chosen significance level is 0.05, some cases have test size close to 0.05, whereas other cases have smaller test sizes. A further investigation reveals what the cases for which the test size is close to 0.05 have in common. Assume the outcomes are sorted by decreasing value of some chosen test statistic. The M step will result in rejection of the null hypothesis for outcomes that are above a certain limit, where the limit is the p-value closest to 0.05 (but not greater than 0.05). The null hypothesis is not rejected for any of the outcomes below the limit. Assume that the last outcome for which H_0 is rejected, $\boldsymbol{y}_0$ has a maximum tail probability $P_{\text{M},0}$, i.e. p-value, in the point $\boldsymbol{p}_0$. If the true value of $\boldsymbol{p}$ is in fact $\boldsymbol{p}_0$, then the probability of rejecting H_0 is the sum of the probabilities of this outcome and the outcomes above, which is $P_{\text{M},0}$. Thus a test size of almost 0.05 is always obtained for a particular $\boldsymbol{p}$, it is only the discreteness that prevents it from exactly being obtained for a specific value of $\boldsymbol{p}$. This value is the value of $\boldsymbol{p}$ that maximizes the p-value for the outcome that has the largest p-value less than or equal to 0.05. If one wants to report the test size in a certain multinomial case, choosing this value of $\boldsymbol{p}$ will ensure that the test size is close to 0.05 unlike the six multinomial cases we chose.

7 Conclusions

In this work we have provided an in-depth effort of using enumeration and exact p-values to address the problem of comparing positive predictive values. The existing tests for this situation rely on asymptotic distributions and have previously been shown not to preserve the test size when the sample size was moderate. The test size and power of nine test statistics in combination with five types of p-values have been thoroughly evaluated for different sample sizes. As demonstrated, the M step yields valid p-values, although these are often conservative. The E step provides a reordering of the reference set in contrast to the M step and one or two E steps before the M step increases the test size while yielding valid p-values.

We have presented three new test statistics, $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$, that can be applied to any problem. In the problem of comparing binomial proportions, the π_{e} test statistic performed better than the test statistics analyzed by Lloyd (2008) in terms of test size and power for the E+M p-values.

For comparing the positive predictive values from two diagnostic tests, we recommend using either the likelihood ratio, restricted difference or π_{M} test statistic and to calculate the E+M p-values. These p-values are valid, and for these test statistics the results have indicated that there is no need to do more than one E step before the final M step. However, the importance of one or more E steps before maximization is greater for e.g. the LAP and unrestricted difference test than for the likelihood ratio test as it increases the test size more significantly, suggesting that the ordering provided by the LAP and unrestricted difference test is not optimal with respect to test size and power.

We do not recommend using the prior predictive p-values, as these are very sensitive to the choice of prior and on the null hypothesis formulation.

This report gives further general insight into the mechanisms behind the E, M and E+M p-values in general and in the example discussed by Lloyd (2008). We describe how the E p-value changes the ordering of outcomes and why this reduces the conservativeness of the M p-values if the E p-values are applied before the M step.

In further work, it would be of interest to find a test statistic that in some sense provides an optimal ordering of the outcomes with respect to test size and power and in particular, the π_{e}, π_{E} and π_{M} should be studied in greater detail and compared to other test statistics. We would also like to investigate if ordering of the outcomes converges after a certain number of E steps, and also the effect of performing two or more consecutive sequences of the form $\mathrm{E}^k\mathrm{M}$.

References

Agresti, A. (2002). *Categorical data analysis*, second edn, John Wiley & Sons, Inc., Hoboken, NJ, chapter 1.4.4.

Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models, *Journal of the American Statistical Association* 95(452): 1127–1142.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter, *Journal of the American Statistical Association* 89(427): 1012–1016.

Bickel, J. and Doksum, K. A. (2001). *Mathematical statistics*, second edn, Prentice Hall, Inc., chapter 4.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Fisher, R. A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society* 98: 39–82.

Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Human Heredity* 53: 146–152.

Günther, C.-C., Bakke, Ø. and Langaas, M. (2009). Comparing positive predictive values for small samples with application to gene ontology testing. Preprint Statistics No. 3, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Günther, C.-C., Bakke, Ø., Lydersen, S. and Langaas, M. (2008). Comparison of predictive values from two diagnostic tests in large samples. Preprint Statistics No. 9, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Lloyd, C. J. (2008). Exact p-values for discrete models obtained by estimation and maximization, *Australian & New Zealand Journal of Statistics* 50(4): 329–345.

Moldovan, M., Langaas, M. and Bahlo, M. (2009). Efficient error-free computation of MAX p-values with an application to genome-wide association studies. Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, submitted.

Zelterman, D., Chan, I. S.-F. and Mielke, P. W. (1995). Exact tests of significance in higher dimensional tables, *The American Statistician* 49(4): 357–361.