Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

The effects of applying cell-suppression and perturbation to aggregated genetic data

Athos Antoniades*, John Keane¶, Aristos Aristodimou*, Christa Philipou*, Andreas Constantinou*,
Christos Georgousopoulos†, Federica Tozzi §, Kyriacos Kyriacou‡, Andreas Hadjisavvas‡, Maria Loizidou‡,
Christiana Demetriou‡ and Constantinos Pattichis*

*University of Cyprus, Nicosia, Cyprus. Email: athos@cs.ucy.ac.cy
¶University of Manchester, Manchester, United Kingdom. Email: john.keane@manchester.ac.uk
†INTRASOFT International, Luxembourg, Brussels. Email: christos.georgousopoulos@intrasoft-intl.com
‡Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. Email: kyriacos@cing.ac.cy
§School of Medicine, University of North Carolina, Chapel Hill, NC, USA. Email: federica.tozzi1@tin.it

Abstract—**The key test for confidence in any association discovered within the medical domain is replication testing. That is, the ability of the association to be detected in independent populations. At the same time, in order to increase the likelihood of discovering statistically significant associations there is a clear need to increase the statistical power of any given study. A key methodology for increasing statistical power is through the use of as many subjects as possible that match a study's inclusion criteria. Thus many have attempted to merge data from multiple independent sources/sites/studies that contain the same inclusion criteria for subjects as a way of creating a much larger study with significantly more statistical power. For these approaches to work though data from multiple sites need to be made available to a single analysis. This practice is significantly limited by the need to respect legal and ethical requirements that are often complicated, ambiguous and inconsistent across different countries. The common approach to achieve merging of data is by sharing aggregated data rather than subject's personal data. Aggregated data however may still in some cases be reverse engineered, therefore traditionally cells within the aggregated data with small values were suppressed, and some or all of the aggregated data were perturbed in order to add noise inhibiting any attempts at identifying personal information of a specific person or sub-group in the original data. In this paper we study the effects of cell-suppression and perturbation on the results of the data analysis. Each approach is looked at by itself as well as in combination using the typical settings documented in the literature. The tests are based on a real dataset that looks for associations between phenotypes and genetic markers. This work is part of the Linked2Safety project that aims to dynamically interconnect distributed patients' data to better enable medical research efforts, whilst respecting patients' anonymity, as well as European and national legislation.**

Index Terms—**Cell-suppression, Perturbation, Noise, Anonymisation, Aggregated Data**

I. Introduction

In recent years a considerable number of genetic studies have resulted in a plethora of large independent datasets for many diseases. Although it is clear that merging applicable datasets together will significantly increase the likelihood of discovering true positive associations, between genetic and phenotypic factors, this is not happening primarily due to legal and ethical hurdles put forth to ensure subjects' privacy and confidentiality rights are observed [1]. As a way of addressing legal and ethical issues and still be able to derive and publish results, as well as merging data from multiple studies together for analyses, scientists have been publishing and sharing aggregated data in various forms and formats. Hence, rather than publishing personal data of each subject in a study, only counts of subjects that match each phenotypic combination of interest are reported, typically in the form of contingency tables. This enables merging analyses from two or more independent studies as long as they all match the variables in the contingency table. By simply adding the respective values together from each independent study's contingency tables, a merged contingency table is created that is identical to the one that would have been generated if the entire set of personal medical records were analysed.

However this approach is known to be vulnerable to reverse engineering of the aggregated data potentially leading to revelation of some of the subject's personal information. These include the case where a single subject is extremely rare in a dataset based on a small number of variables (as an example a Cypriot 16 year old widow). If there is only one (or small number) in the general population that matches that combination then all who know of him/her will have identified that she has provided data to the study and could even identify more of the same subject's data by then examining supersets of the variables (Cypriot, 16 year old, widow, diabetic and obese). Another example of vulnerability is the case of dual release of information of the data using two different quantization techniques. As an example consider a publication that reports the number of female widows over the age of 16 in a study, and another on the same study that reports the number of widows over the age of 17. If the difference is 1 then again that subject's data could be re-identified. All of these potential vulnerabilities are magnified by mapping of the study dataset to other electronically available datasets that have or may in the future be published.

In the case of genetic data these problems become even greater as the nature of genetic markers makes it impossible to truly anonymise them. Therefore a safe and secure way of sharing aggregated data was needed. Many studies have attempted to do this by using either cell suppression values to remove aggregated data that were less than a pre-set threshold, or perturbation to randomly add or subtract a small value from each aggregated count (sometimes referred to as Barnardisation when limited to a perturbation range of $[-1, 1]$) as a way to introduce noise into the data and hence reduce the possibility of re-identification [1]–[4]. Most researchers however tend to prefer to use a combination of both techniques as a way to safeguard that legal and ethical concerns are sufficiently addressed when releasing aggregated data [5]–[7].

In this paper we focus on studying the effects of both cell suppression and perturbation when applied alone or together with various typically used parameters. In order to do so we use a real dataset from a breast cancer study that contains genetic data in the form of Single Nucleotide Polymorphisms (SNPs) and binary phenotypic variables.

II. Methodology and Resources

A. The Linked2Safety Project [8]

Electronic Health Records (EHRs) contain a wealth of medical information. They have the potential to help significantly advance medical research, as well as improve health policies, providing society with additional benefits. However, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models.

The vision of Linked2Safety is to advance clinical practice and accelerate medical research, by providing pharmaceutical companies, healthcare professionals, scientists and patients with an innovative secure semantic interoperability framework facilitating efficient and homogenized access to anonymised distributed EHRs.

The dataset described in the following section is part of the data that is considered for inclusion in the Linked2Safety project. This project aims to aggregate the data from each data provider's site, sharing only the aggregated data in the form of data cubes with the rest of the Linked2Safety platform. Therefore the question of how to address legal and ethical considerations associated with confidentiality issues in each data resource in the project is key. In this paper work is focused on a part of the data provided by the Cyprus Institute of Neurology and Genetics (CING). The goal is to identify the maximum amount of noise that can be added through perturbation and cell suppression whilst meeting an a-priori defined correlation to the analyses that would result from analysing the original data.

B. Description of Data and Resources

For this project we use published data from the "MASTOS", a case control study, from CING. The goal of "MASTOS" was to investigate the association of several SNPs in DNA repair genes, with breast cancer in Cypriot women [9]–[12]. The genetic markers underwent quality control testing with a threshold of 10% for the minor allele frequency. 19 SNPs passed the threshold and where included in the analyses.

For the purposes of this work it was important that the degrees of freedom of all tests were the same, therefore at the stage of quality control only binary phenotypic variables were selected with a frequency of the rarest value in each variable being greater than 5%. A total of 8 phenotypic variables passed this criterion resulting in a total number of phenotypic-genotypic association tests of 152.

The software used for this analyses was developed in c++ and can run on Unix, Linux or MS Windows operating systems. It is available for anyone wishing to replicate these analyses on independent datasets by emailing the primary author.

C. Perturbation

Perturbation for the purposes of this paper takes as input a parameter referred to as the perturbation range. Perturbation is implemented allowing for both negative and positive perturbation. Thus if the perturbation range is 3, then a random integer will be generated under a normal distribution within the range of $[-3, 3]$ and added to the original value. If the resulting value is negative then it will be replaced with 0.

D. Cell Suppression (Cut-off)

Cut-off is a straight forward approach to limit the risk of having aggregated values that are so small that it may be possible to identify the subjects to which they relate. Each cell is tested if it exceeds a pre-set threshold value referred to as the cut-off threshold. If it passes the cut-off threshold then it is reported as is, if not then it is replaced. For the purpose of this paper we replaced the aggregated values that failed to pass the cut-off threshold with the median between 0 and the threshold. Tests where cut-off is not applied will be reported as cut-off with a threshold of 0 as that would cause no changes.

E. Analytical Approach

Pearson's Chi-Square test was used to perform a test of the main effect (the association between a SNP and a categorical phenotype) as described in [13]. All phenotypes were selected to be binary phenotypes resulting in tests that all had 2 degrees of freedom (SNPs have 3 categories commonly indicated as "aa", "aA" and "AA" with "a" representing the less frequent allele). Yates' correction for continuity was also applied in order to address a known issue with Pearson's Chi Square test of overestimating the statistical significance of contingency tables with cells that contain very small numbers [14].

The purpose of this paper is not to discover statistically significant associations, but rather to evaluate the likelihood of discovering such after applying perturbation and cell suppression. Therefore, as all tests had the same degrees of freedom, in order to better visualize the results rather than showing a graph from each phenotype analyses and cut-off/perturbation parameter the results from all phenotypes and genetic markers were merged together producing a single pool of results per cut-off and perturbation combination setting.

In order to evaluate the performance of each tested cut-off and perturbation parameter, these will be compared to the same analyses performed on the raw data with no perturbation or cut-off applied. Pearson's correlation is estimated between the analyses performed on the original data and each cut-off and perturbation parameter combination. The cut-off for Pearsons correlation is set a-priori to accept two tests as having a reasonable agreement of $r > 0,95$ [15].

III. Results

For cut-off the thresholds tested were 1, 3, 5, 10. These were selected based on the literature as being the most often used values for cut-off thresholds. For perturbation the ranges used were $[-1, 1], [-3, 3], [-5, 5], [-10, 10]$ indicated in all graphs henceforth by the absolute number of the integer at the edge of the range correspondingly 1, 3, 5, 10. For both procedures when they were not used the parameter was set to 0, thus when both the cut-off and perturbation range were set to 0 the outcome was the one received from analysing the original contingency tables with no perturbation or cut-off applied. All possible unordered combinations of the two procedure parameters were tested.

The correlation coefficient was estimated between each of the results of a perturbation and cutoff parameter combination and the same analyses on the original data with no cut-off and perturbation. Table I shows the correlation coefficient between the original aggregated data and each of the cut-off threshold and perturbation range parameters. In this table the rows represent the perturbation range set while the columns represent the cut-off threshold set. All Pearson's correlation coefficients that pass the a-priori defined 0.95 threshold are in bold while the rest are not.

TABLE I
Pearson's Correlation Coefficients

Perturbation	10	0.54	0.54	0.54	0.55	0.53
	5	0.76	0.75	0.76	0.75	0.78
	3	0.90	0.89	0.89	0.90	0.86
	1	**0.98**	**0.97**	**0.98**	**0.98**	0.90
	0		**1**	**0.99**	**0.98**	0.90
		0	1	3	5	10
		Cutoff				

In order to better demonstrate the output of the analyses a graph is created for each of the perturbation and cut-off parameter combinations (Fig. 1). In these graphs the X axis represents the $-log$(p-value) of the analyses on the original aggregated data with no perturbation or cut-off performed while the y-axis represents the $-log$(p-value) of the analyses of the aggregated data after perturbation and cut-off has been applied to it. The key goal in these type of analyses is to identify statistically significant results thus bold grid lines indicate in each axis the typical threshold of statistical significance used in genetic studies (p-value< 0.01). The grid lines serve a secondary purpose as well, they intersect at the $x = y$ point. Under the hypothesis that the perturbation and cut-off applied in a specific graph had no effect on the resulting p-values we expect to see the results forming a line across the X=Y diagonal. All data points deviating from X=Y indicate data points that were affected by the perturbation and cut-off parameters applied.

In order to provide a visualization that will easily allow comparison between the performance of different parameter settings a super-graph is created that on the X-axis holds the cut-off threshold parameter and on the Y-axis the Perturbation range. Within this graph all the corresponding graphs of the perturbation and cut-off parameters are presented in Fig. 1.

The first observation from looking at Fig. 1 is that most of these graphs distort the results to a point were they would no longer be meaningful. Take for example all graphs with a perturbation range of $[-5, 5]$ and $[-10, 10]$ and to a lesser extent $[-3, 3]$. They are all distorted to such a level that many results that were not statistically significant before, adding noise to the data are now in some cases not just statistically significant (and thus false positive) but in many cases carry stronger associations than the results that were statistically significant based on the original data with no added noise. This observation is made clear also by the failure of these settings to pass the a-priori defined threshold for the Pearson's correlation coefficient as shown in Table I. It is therefore obvious to conclude that it is necessary to limit the perturbation range to $[-1, 1]$ in this dataset, in order to limit the noise to an acceptable level.

Cell suppression (cut-off) appears to be less intrusive to the end results. A cut-off threshold of 5 or below produced results that passed the a-priori defined Pearson's correlation with the noiseless data. Combining perturbation within the range of $[-1, 1]$ to cell suppression with a cut-off threshold of all tested values less than or equal to 5 also produced results that stood up to the added noise, passing he Pearson's correlation coefficient test and also not changing the results that were found to be statistically significant with the noiseless dataset. The cut-off threshold of 10 however failed Pearson's correlation threshold and resulted in a single test passing statistical significance even though it should not based on the analyses of the data with no added noise.

IV. Discussion

This analysis suggests that based on the phenotypes and genetic markers tested in the MASTOS dataset, the ideal settings for the cut-off threshold is 5 and a perturbation range of $[-1, 1]$ is recommended for adding the maximum amount of noise to the dataset whilst not altering the conclusions that will be derived from the resulting analyses. However, before actually sharing or publishing these results to the medical community outside the study organizing host (in this case CING), it is essential to determine if that amount of noise for that dataset is sufficient to satisfy legal and ethical requirements related to confidentiality.

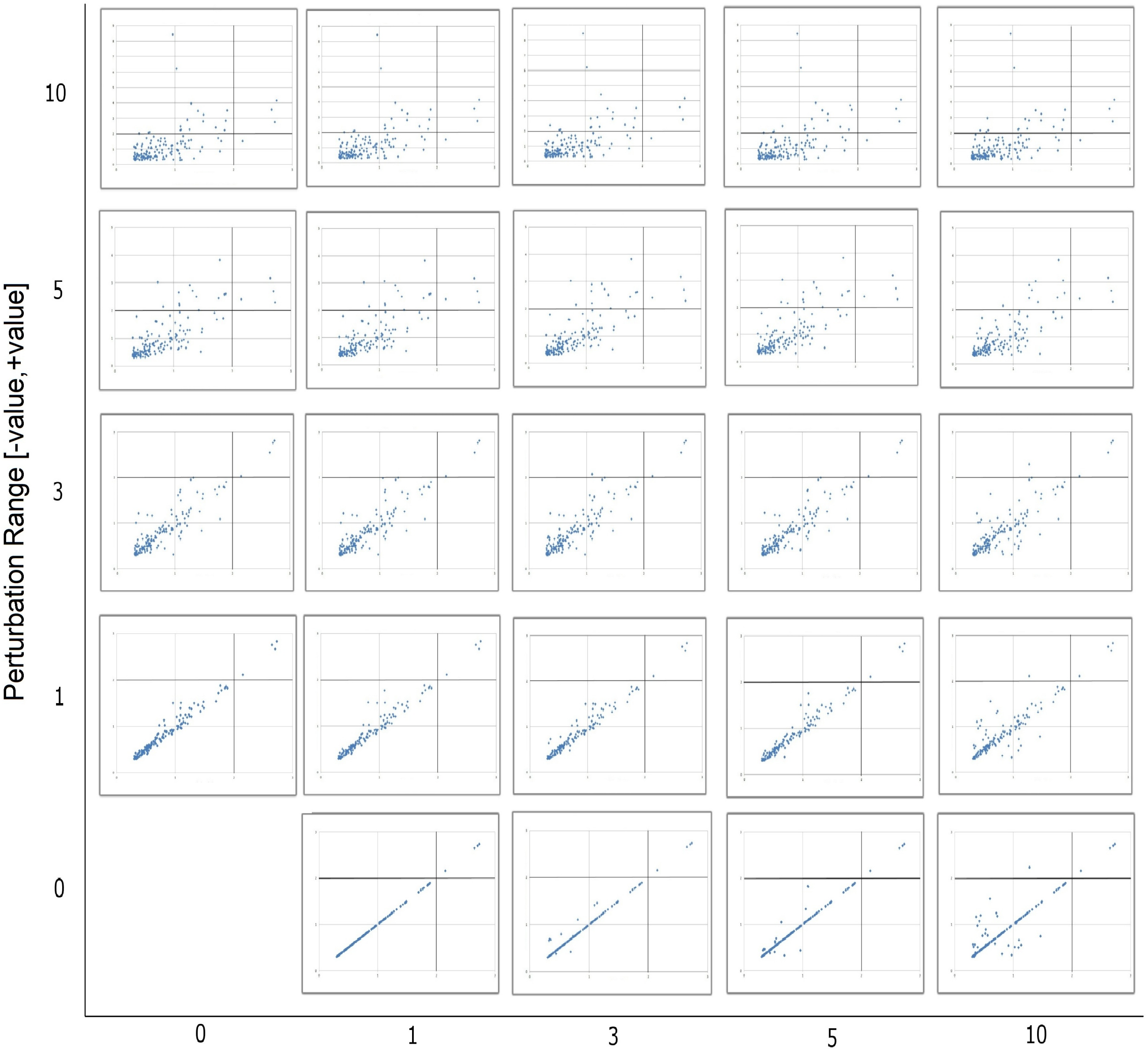

Fig. 1. Cut-off and perturbation evaluation graphs. Each cut-off and perturbation parameter combination is represented by a graph at the relevant cell in the figure. The resulting $-log$(p-value) are plotted against the results of the corresponding tests with no perturbation or cut-off applied (no added noise). The bold vertical and horizontal lines indicate in each graph the a-priory statistical significance threshold

The statistical analyses applied in this study, Pearson's Chi Square to detect main effects, although a typical and traditionally used measure for association tests between categorical phenotypes and genetic markers in genetic analyses, is not the only type of analysis. There are other types of analyses that may be more or less susceptible to noise. Unfortunately it is impossible to know a-priori all of the analytical techniques that may be applied on the aggregated, noise induced data after their publication. This is typically the case for most published aggregated data even if they are presented as a simple contingency table in a publication [8]. However, the choice of test for this specific work was one that is widely applied to such studies and data. This therefore allows us to draw a conclusion based on the use of these noise adding techniques and their effects on a standard analytical technique.

In addition, when estimating the statistical power of detecting true positive effects in a dataset one needs to consider many parameters. These same parameters are likely to affect the tolerance of noise of each individual dataset. Examples of such parameters are the size of the general population sampled, the size of the sample, the frequency of the phenotype tested, number of categories in the variables etc. All these parameters were kept constant in all tests performed in this paper in order to be able to reach a conclusion on the ideal parameters for adding the maximum noise with an acceptable impact on results of Pearson's Chi Square test for main effects. Therefore, the final conclusion of this paper is that although it was possible to identify the ideal parameters for adding maximum noise while maintaining the result's quality to an acceptable level in this specific study, parameters and tests, these should not be interpreted as being transferable to other datasets unless similar work is performed for those datasets to establish that the same observations apply for them as well.

V. Conclusion and Future Work

In this study the effects of adding noise to data through cell suppression and perturbation were evaluated by applying a traditional statistical approach to test for association between genetic markers and binary phenotypes. By testing multiple cut-off thresholds, perturbation ranges and their combinations and comparing the end results of each analysis to the original noiseless analyses it was possible to identify the combination of parameters that induced the most noise, whilst preserving the findings of the original noiseless data. The aim of this work was to identify these parameters so that in future work establish if these would satisfy the legal and ethical requirements for confidentiality in this study and in it's intended application in the Linked2Safety project [8]. Although this work has demonstrated that it is possible to identify these parameters, there is no inference that these can be expanded further to other studies without repeating this type of analyses or at least taking into consideration parameters that were kept constant in this work. Furthermore, considerable research is taking place in parallel to this work to determine what constitutes sufficient levels of noise in the MASTOS dataset so as to successfully address legal and ethical requirements for subject confidentiality.

In future work it would be beneficial to have multiple study owners repeat this same analysis in their own datasets, preferably studies that come from diverse independent populations, of different sizes, with different frequencies of the effects studied in each, with different types of variables etc. This would help better understand the effects of the differences between each study parameters to the amount of noise that can be added with each methodology. This is made feasible for any study as the software used as part of this work is available to other researchers. Testing the effects of each cell suppression and/or perturbation parameter in different datasets is feasible as the only data that needs to be shared is the same data presented in this paper, the p-values of the resulting tests, and Pearson's correlation coefficients; these are analyses results and are not aggregated, or personal data and can't be used to re-identify subjects.

Acknowledgements

The research leading to these results was conducted as part of the project *A next-generation secure linked data medical information space for semantically-interconnecting electronic health records and clinical trials systems advancing patients safety in clinical research (Linked2Safety)* that received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No 288328.

The data used in this research was created through the Research Promotion Foundation of Cyprus with projects 0104/13 and 0104/17 and through funding from the Cyprus Institute of Neurology and Genetics.

References

[1] J. Ellenberger and S. Muir, "National statistics offices and the prosumer challenge," in *NTTS 2009 Conf. Proc.*, 2009.

[2] Y. Li and H. Shen, "Anonymizing graphs against weight-based attacks with community preservation," *J. of Comput. Sci. and Eng.*, vol. 5, no. 3, pp. 197–209, 2011.

[3] N. Adam and J. Worthmann, "Security-control methods for statistical databases: a comparative study," *ACM Comput. Surveys (CSUR)*, vol. 21, no. 4, pp. 515–556, 1989.

[4] P. Williamson, "The impact of cell adjustment on the analysis of aggregate census data," *Environment and Planning A*, vol. 39, no. 5, p. 1058, 2007.

[5] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 92–106, 2006.

[6] C. Giannella, K. Liu, and H. Kargupta, "On the privacy of euclidean distance preserving data perturbation," *Arxiv preprint arXiv:0911.2942*, pp. 2, 5–9, November 2009.

[7] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Rec.*, vol. 29, no. 2, 2000, pp. 439–450.

[8] A. Antoniades *et al.*, "Linked2Safety: a secure linked data medical information space for semantically-interconnecting EHRs advancing patients safety in medical research," in *12th IEEE Int. Conf. on Bioinformatics and Bioengineering*, Cyprus, Nov. 2012.

[9] M. Loizidou *et al.*, "DNA-repair genetic polymorphisms and risk of breast cancer in Cyprus," *Breast cancer research and treatment*, vol. 115, no. 3, pp. 623–627, 2009.

[10] A. Hadjisavvas *et al.*, "An investigation of breast cancer risk factors in Cyprus: a case control study," *BMC cancer*, vol. 10, no. 1, p. 447, 2010.

[11] M. Loizidou *et al.*, “Genetic polymorphisms in the DNA repair genes XRCC1, XRCC2 and XRCC3 and risk of breast cancer in Cyprus,” *Breast cancer research and treatment*, vol. 112, no. 3, pp. 575–579, 2008.

[12] M. Loizidou, Cariolou *et al.*, “Genetic variation in genes interacting with BRCA1/2 and risk of breast cancer in the Cypriot population,” *Breast cancer research and treatment*, vol. 121, no. 1, pp. 147–156, 2010.

[13] A. Antoniades *et al.*, “A computationally fast measure of epistasis for 2 snps and a categorical phenotype,” in *Eng. in Medicine and Biology Soc. (EMBC), 2010 Annu. Int. Conf. of the IEEE*, 2010, pp. 6194–6197.

[14] F. Yates, “Contingency tables involving small numbers and the χ 2 test,” *Suppl. to the J. of the Roy. Statistical Soc.*, vol. 1, no. 2, pp. 217–235, 1934.

[15] B. Everitt, *The analysis of contingency tables*. Chapman & Hall/CRC, 1992.