Efficient Qualitative Method for Matching Subjects with Multiple Controls

Hung-Jui Chang
Department of Applied Mathematics
Chung Yuan Christian University,
Taoyung, Taiwan
Email: hjc@cycu.edu.tw

Yu-Hsuan Hsu, Chih-Wen Hsueh
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei, Taiwan
Email: {b99902110, cwhsueh}@csie.ntu.edu.tw

Tsan-sheng Hsu*
Institute of Information Science, Academia Sinica,
Taipei, Taiwan
Email: tshsu@iis.sinica.edu.tw
*Corresponding Author

Abstract*—In the era of learning healthcare systems and big data, observational studies play a vital role to discover hidden (causal) associations in the dataset. To control bias, a matching step is usually employed to match case subjects to control candidates in observational studies randomly. The matching ratio refers to the number of control candidates matched with one case subject, and the successful matching rate is the percentage a matching is found given a matching ratio. A good matching algorithm should be not only efficient but also have high successful matching rate and high quality of randomness which means that a control candidate has a roughly equal chance of being matched with any of the matchable study cases. In this paper, we propose a matching algorithm, which is efficient with above mentioned good properties, RandFlow, a high-quality matching algorithm, is proposed and compared with commonly used ones – Simple_Match, Matchit, and Optmatch. The benchmark testing shows the effectiveness of the new algorithm. In our experimental studies, we noticed that the variation of the estimated Relative Risk (RR) value is minimized at the maximum matching ratio. Thus, we propose a two-phase matching method to obtain more reliable study results. The first phase is to identify the maximum matching ratio, and followed by matching multiple times and then take an average.

Keywords–matching; observational study; relative entropy

I. INTRODUCTION

Observational studies are often used for investigating causal relationships. Given two events, α and β, researchers can analyze whether the occurrence probability of the event β is affected by the event α happening previously. In the medical field, an event can be a diagnosis, prescription or treatment. To control bias, several approaches have been applied, and one of them is matching [1]. Hence, the observational study process starts from identifying the study group G_α (those individuals with α), matching to the control candidates $G_{\neq\alpha}$ (those individuals without α), and then performing statistical analysis to draw a conclusion. For example, Relative Risk (RR) is used to estimate the relative risk of having β with and without the occurrence of α before. For example, in Table I, there are $a+b$ individuals with the event α, and a of them also with the event β. The conditional probability, R_1, which denotes the probability of having β under the condition of with the event α is therefore $a/(a+b)$. Also, there are $c+d$ individuals without the event α, and c of them with the event β. The conditional probability, R_2, which denotes the probability of having β under the condition of without α is therefore $c/(c+d)$. The RR value is defined as $RR = R_1/R_2$. RR values greater than, less than, or equal to 1 indicate positive, negative or no relationships, respectively. Other statistics, such as Odds Ratio (OR), may be used instead of RR depending on the study design.

TABLE I. EXAMPLE OF STUDY GROUP AND ITS MATCHED CONTROL GROUP

	α	$\neg\alpha$
β	a	c
$\neg\beta$	b	d
Sum	$a+b$	$c+d$

Matching is a critical step in the analysis of the observational study. Generally, a matching algorithm randomly permutes the order of the input of study case s, and control candidate c, and then checks whether the input s-c pair can be matched, and finally matches s with K-fold eligible controls one by one. The constant K is called the matching ratio. Some matching methods assign a propensity score to each pair [2] and return a matching with the best total score. However, if the distribution of cases is skewed, the study case may not be able to match with the required amount of controls successfully and would be dropped to avoid incurring further bias. Therefore, the output matching needs to satisfy some quality criteria, such as randomness and successful matching rate. In a good quality matching algorithm, a control candidate has a roughly equal chance of being matched with any of the matchable study cases. Keeping as many successful matchings as possible is also desired.

There are some commonly used matching methods, like Simple_Match [3], MatchIt [4][5], and Optmatch [6]. The former is based on a simple randomized greedy approach using SAS and the randomized algorithm has no proof of being able to deliver a matching in reasonable time in [3], and the latter two are variations of the well-known max flow algorithm [7] though with a performance guarantee, but does not consider any randomness. If the matching is only performed once with a small matching ratio, the result may not be stable in the sense that it is possible that different matchings may yield fluctuating statistics, such as RR or OR. To obtain a reliable result, it is better to match multiple times and take an average of all the outcomes. However, it is not practical to do repeated matching due to its heavy time consumption. Moreover, determining the matching ratio is also a cloudy issue in practice. In the past few decades, the case-control study has been suggested to match with four or five times of controls [8]. It was reported that ”beyond a ratio of about 4/1, little power improvement results from increasing the number of controls” [9]. However, a matching ratio of 10 or even 15 is also seen in some studies [10][11]. In Hennessy’s study [12], they indicated a higher matching ratio might be needed while the disease prevalence is low and hence, the implied matching ratio should be data dependent [13]. Up to date, few studies are investigating the issue of finding a good matching ratio.

Previous researches have focused on the impact of the matching ratio [13], and whether to use a matching or not [14]. But how to determine the matching ratio is less discussed. To resolve the above problems, we proposed a high-quality matching algorithm called *RandFlow*, which adopts the idea from maximum flow in graph theory. In RandFlow, we added some vital functions to raise the randomness and matching efficiency. Furthermore, we leveraged the high efficiency of RandFlow to determine the optimal matching ratio. By using RandFlow, the maximum matching ratio of each data set is calculated, and the range of the suitable matching ratio is also determined. The researcher can choose a preferred matching ratio according to the suggested range.

The remainders of this paper is organized as follows. In Section II, we describe our matching algorithm, the data source used in this study and the factors compared between different matching methods. In Section III, we show the experiment results of RandFlow and the comparison between RandFlow and the original methods. In Section IV, we discuss the comparison results and summarized our conclusions.

II. METHODS

The approach of our method is to formulate our problem in the well-known framework of flows in networks [7]. Hence, our methods come with performance and correctness guarantees. In this study, we used Taiwan’s National Health Insurance Research Database (NHIRD) [15] as a data source and examined the validity of RandFlow by three causal relations reported in the published papers. We then compared RandFlow with the above matching methods with regard to successful matching rates, RR values and quality of randomness.

A. RandFlow Algorithm

We transform the matching problem in Figure 1(a) to the well-known max flow problem [7] in Figure 1(b). In a max flow problem, we assign maximum integer weights, not exceeding the pre-assigned capacity, to the edges so that for each vertex other than the source and sink, the sum of weights on its incoming edges equals the sum of weights on its outgoing edges. A study case S_i is matched with those control candidates C_j so that the weight of the edge from S_i to C_j is 1. The outcome is called a *max flow*. We further require that each study case has the same sum of incoming edge weights, which is called the *maximum matching ratio*, denoted by r. Thus, each study is matched with exactly r candidates, and each candidate is matched at most once. Since a max flow is found, r is as large as possible. Note that the value of r is data dependent. Each data set has its own maximum matching ratio. Naturally, it is unreasonable to ask for a matching whose ratio is more than r. In addition to whether a matching of a specified size can be found efficiently or not, we also concern whether the resulting matching is random or not, i.e., whether each candidate has an equal chance of being selected by any case subject. Without considering constraints incurred from competitions between case subjects, we use the well-known *entropy* [16] E of the ideal distribution among all possible candidates that can be matched to a case subject. Then we measure the entropy E' of the actual distribution of candidates being found by applying the matching repeatedly says 1000 times. We define the *relative entropy* to be $\frac{E'}{E}$ to quantify the quality of randomness in the matching obtained.

There are known algorithms to find such a max flow in $O(|E||F|)$ time, where E is the set of edges and F, called *flow*, is set of edges with weight 1 between the study cases and candidates. The value of $|F|$ is the number of edges inside. The algorithm finds a maximum flow by finding successively what is called an *augmenting flow* $F^{'}$ so that each time $F^{'}$ increases the current flow value by 1 after canceling edges from S_i to C_j and from C_j to S_i at the same time. We extend the original algorithm by finding a random augmenting flow, instead of a fixed one using a Randomized version of Depth First Search (RDFS). We also use a merging technique so that given two candidates C_i and C_j are merged if they have incoming edges from the same set of study cases. We also randomly shuffle the ordering of study cases from the input to obtain better randomness quality. Our revised algorithm runs faster and uses less memory than the original one in practice. The technical details can refer to our technique report [17].

B. Data source

The NHIRD is a nationwide database extracted from the claim data of the National Health Insurance (NHI) program in Taiwan for research purposes. In recent years, NHIRD has been widely used to identify potential causal relationships. This study also used NHIRD as the data source and which was reviewed by the Institutional Review Board of Academia Sinica, Taiwan (approval number: AS-IRB-BM-16043). As a benchmark, we selected three distinct causal relations from two published papers. One paper investigated the bidirectional relationship between Obstructive Sleep Apnea (OSA) and depression [18]. The study showed a positive relationship that patients with OSA have increased the risk of occurring depression, and vice versa. The other paper examined whether previous Statin use in patients with stroke affects the subsequent risk of dementia [19]. The study found a negative relationship in such a way that Statin use in patients with stroke decreases the risk of dementia. In this study, we define an event pair as the former event affects the occurrence of the following event. Hence, the relationship between depression and subsequent OSA is denoted as Event Pair I, and the reverse is Event Pair II. The relationship between

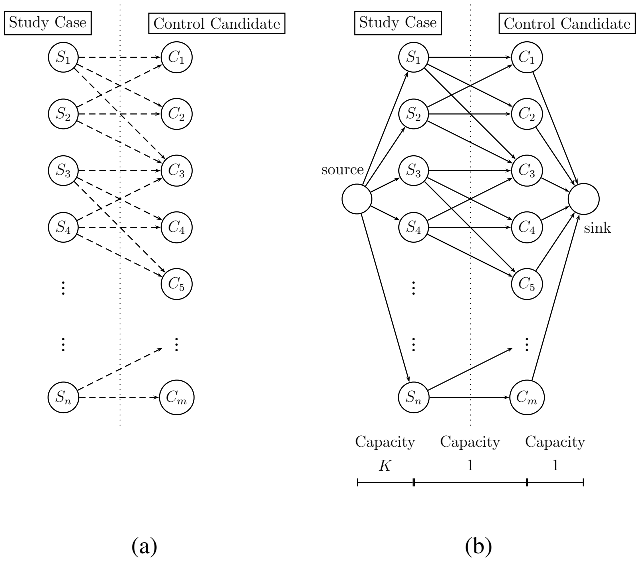


Figure 1. An example of transforming the matching problem into a flow problem.

Statin use in patients with stroke and following dementia is Event Pair III.

C. Comparisons between matching methods

In the original studies, event pair I and II were performed by exact match. Among these two event pairs, each study case was matched with five controls. Regarding event pair III, each study case was matched with one control by propensity score match [20] instead. In our study, all experiments were done by exact match. We used the ratio of control candidates to study cases to conjecture the maximum matching ratio.

We then compared the matching methods with regard to successful matching rates, RR values and quality of randomness. Successful matching rate is defined as the percentage of matched study cases that are not dropped. We assessed the average execution time, the corresponding successful matching rates and RR values with matching ratios from 1 to 30 (to 90 in the case of Event Pair II). To further understand the variation of RR values, we also examined the standard deviation of RR, R_1, and R_2. R_1 and R_2 represents the risk of having in the study group (G_α) and control group ($G_{\neq\alpha}$), respectively. The ratio of R_1/R_2 is RR. For the quality of randomness, we calculated the relative entropy of the matched control candidates with three different matching ratios: 70%, 100% and 110% of the maximum matching ratio. RR value and relative entropy were run 100 times and took the average. Because the programs implemented in C are more efficient and memory saving, we only compare C implementations in terms of successful matching rates, RR value and quality of randomness. All the experiments were performed on a Ubuntu 14.04 system with an Intel(R) Core(TM) i7-3770 CPU 3.40 GHz, and 16 Gbytes RAM.

III. Results

A. General result of the randomly sampled data

Figure 2 shows the result of the randomly generated data. The x-axis denotes the real RR value, and the y-axis denotes the estimated RR value which is calculated by RandFlow. Each point in the figure represents one data set. The results show RandFlow can get an estimated RR value very close to the real RR value. The statistic results are summarized in Table II. The first and second column denotes the real RR value and the estimated RR value. The third to fifth column denotes the absolute error between the real and the estimated RR value, the variance of the estimated RR value and the standard deviation of the estimated RR value. The experiment results show the absolute error RandFlow Algorithm is less than 0.06 and the variance and standard deviation is only 0.10 and 0.33, respectively.

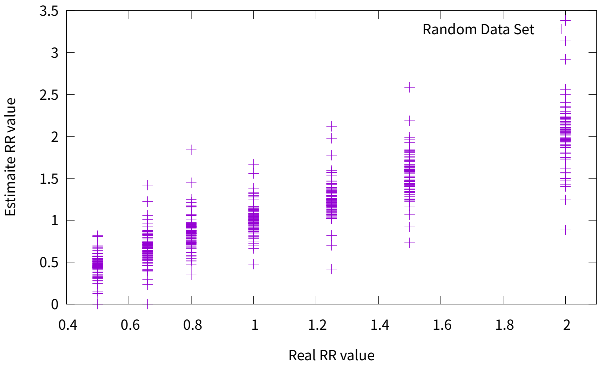


Figure 2. The distribution of real RR value and the estimated RR value of RandFlow.

TABLE II. THE STATISTIC RESULTS OF REAL RR VALUE AND THE ESTIMATED RR VALUE OF RANDFLOW.

Real RR	Estimated RR	Δ	Variance	STD
0.50 (1/2)	0.462	0.038	0.021	0.144
0.66 (2/3)	0.658	0.002	0.033	0.182
0.80 (4/5)	0.854	0.054	0.041	0.202
1.00 (1/1)	1.016	0.016	0.029	0.169
1.25 (5/4)	1.259	0.009	0.049	0.209
1.50 (3/2)	1.542	0.042	0.056	0.236
2.00 (2/1)	2.049	0.049	0.105	0.325

B. General information of the selected event pairs

Table III shows the general information of the selected event pairs from the original papers and our results, including the number of controls/control candidates, the ratio of control candidates to study cases, and maximum matching ratio.

TABLE III. GENERAL INFORMATION OF THE SELECTED EVENT PAIRS.

	Event Pair I	Event Pair II	Event Pair III
Original results			
No. study cases	27,073	6,427	5,527
No. control cases	135,365	32,135	5,527
Matching ratio	5	5	1
Our results			
No. control candidates	562,707	619,904	9,102
Control candidates/Study cases	≈21	≈97	≈2
Maximum matching ratio	11	51	0
Total edge	149,676,628	38,629,676	404,835

Among these event pairs, the greatest number of study cases was found in Event Pair I. With such a great amount of study cases, there were a total of more than 149 million edges generated while matching by RandFlow. We speculated the maximum matching ratio would be different among the event pairs as it turned out to be the ratio of 11, 51 and zero for Event Pair I, II and III, respectively. Additionally, these event pairs covered both positive and negative relationships. As a result, we believed that they could be representatives for testing matching quality.

C. RR values and Successful matching rates

Flow-based matching methods keep on matching until they use up all the matchable control candidates. They are expected to have the same traits in terms of RR value variation and successful matching rate. Hence, we only show the comparisons between Simple_Match and RandFlow in this section.

Overall, the average RR values of Simple_Match are higher than the values of RandFlow. In both methods, the average RR values are fairly stable while the matching ratio is small and gradually decrease when the matching ratio exceeds a certain value. In RandFlow, the decline occurs at the maximum matching ratio. By contrast, the decline of Simple_Match happens earlier than that (Figure 3(a) and 3(b)). In the case of a negative relationship in Event Pair III, the average RR values increase instead of decrease (Figure 3(c)).

Generally speaking, the variation of RR values of Simple_Match are more unstable than that of RandFlow. In both methods, the variation of RR values steadily decrease and then turn up at a certain matching ratio. The least variation of RR values of RandFlow occurs right at the maximum matching ratio. That of Simple_Match happens before the maximum matching ratio (Figure 3(d)-3(f)).

Since RR is calculated as R_1 divided by R_2, we examined the variation of R_1 and R_2 in RandFlow to further survey where the RR variation comes from. When the matching ratio is less than the maximum matching ratio, no study cases are dropped; thus, the standard deviation of R_1 remains zero. On the other hand, the standard deviation of R_2 decreases with matching ratio until it reaches the maximum. When the size of the control group increases to a certain number, the standard deviation of R_2 becomes relatively small and steady. Beyond the maximum, the standard deviation of R_1 surges because study cases are dropped rapidly (Figure 3(g)-3(i)).

Figure 4 shows the comparison of successful matching rates between Simple_Match and RandFlow. Because Simple_Match is based on a simple greedy algorithm, the matching results from it may vary. We used both the minimal (Simple_min) and the maximal (Simple_max) results from the 100 trials for comparison. Whether the minimal or the maximal result from Simple_Match, the successful matching rates drop before the maximum matching ratio, whereas that of RandFlow remains 100%. At any fixed matching ratio, RandFlow has the highest successful matching rates. Although Simple_Match runs faster than RandFlow, when the execution time is fixed, it cannot achieve the successful matching rate of RandFlow.

D. Quality of randomness

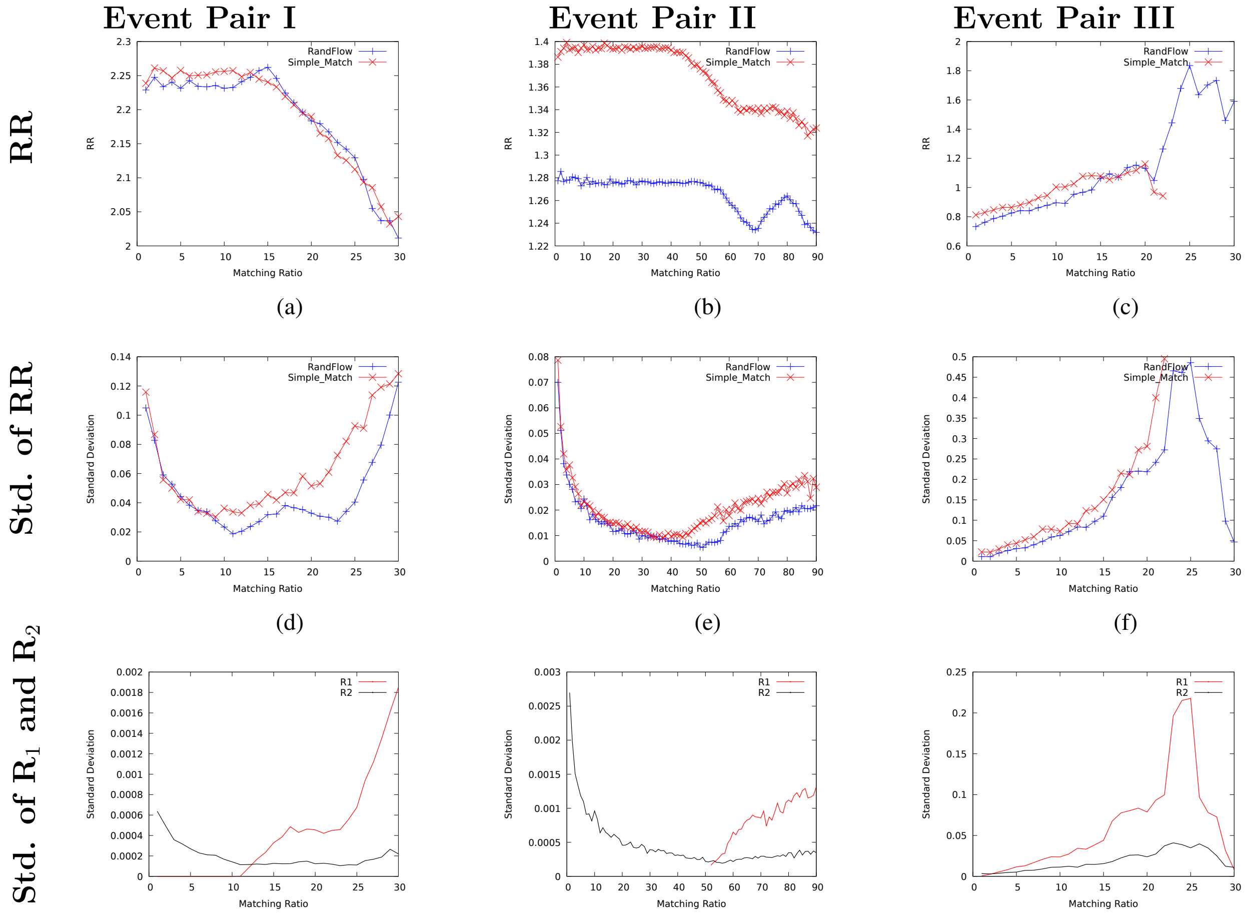
Optmatch and Matchit are both flow-based matching methods without randomly shuffling the input graph. In other words, their matched results remain unchangeable and no randomness at all. By contrast, we implemented RandFlow with inputting random graph and RDFS to enhance the quality of randomness. In this section, we show the comparison of the quality of randomness between RandFlow and Simple_Match.

Figure 5 shows that Randflow has a better quality of randomness than Simple_Match. Relative entropies of Event Pair I and II were tested at 70%, 100% and 110% of the maximum matching ratio in 100 trials. The relative entropy of RandFlow was estimated to be around 1 and generally higher than that of Simple_Match. Additionally, RandFlow has consistently stable entropy at any matching ratio and study case. Even if the ratio was set at 110% of the maximum matching ratio, the relative entropy of Randflow slightly decrease. For those study cases having small sets of control candidates, that of Randflow remains high. By contrast, the relative entropy of Simple_Match fluctuates widely as the matching ratio increases. For those study cases having less matchable control candidates, that of Simple_Match plunges.

IV. Discussion

In this study, we adopted maximum flow theory to develop a highly efficient and good-quality matching method, RandFlow, for matching subjects with multiple controls. This method can accomplish difficult matching tasks, like matching 20 thousand study cases to 30 times controls within a few seconds. Comparing with the most popular matching method, RandFlow has a good quality of randomness and finds a matching rather than drops study cases as long as such a matching exists. Matching is used to make the study cases and controls to have similar distributions across confounding variables. During the matching process, the controls are expected to be randomly selected from the control candidates. Anything that may affect the sampling design like dropping cases should be avoided. Our study used relative entropy to quantify randomness and then verified that RandFlow has a good quality of randomness. The randomness of RandFlow does not vary with the chosen matching ratios as it is no more than the maximum ratio. With regards to successful matching rate, RandFlow outperforms simple greedy algorithms due to the nature of algorithms. Overall, RandFlow surpasses those commonly used matching methods.

Matching ratio is data dependent and should be differentially set at the maximum matching ratio to obtain consistent results. In the past few decades, the case-control study has been suggested to match with four or five times of controls. Previous studies had indicated a higher matching ratio may be desired [9][12][13]. Beyond the previous studies, we tested three distinct data sets and performed matching multiple times at a range of matching ratios. In our experiments, we found that the maximum matching ratio varies with the input data set and the least variation of RR values always happens when we set the matching ratio to be the maximum. This can be explained from the perspective of graph theory. If the matching ratio requested h is no more than the maximum matching ratio w, then we have many possible different matchings. From the law of large number, the RR value calculated from many instances is stable and close to the real average case. If h is more than w, then we do not have many choices in selecting the pairings. The deviation of RR computed tends to be higher than the formal case. Therefore, rather than using an empirical



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3. RR values and standard deviation of RR, R_1 and R_2 of Simple_Match and RandFlow.

fixed matching ratio, we suggest matching each study at its maximum matching ratio multiple times and taking an average for consistent results.

RandFlow being an exact matching has an inherent limitation. Of being unable to match some study cases with the required amount of controls while the distribution of the confounding variable is skewed. In the extreme case, even 1:1 match cannot be reached; thus, the RR values will be unstable at any matching ratios. In these circumstances, other matching methods should be considered in order to obtain reliable results.

In this study, we developed a highly efficient matching method and demonstrated its good quality of randomness. From our experiments, we further conclude that the matching ratio is data dependent and should be differentially set at the maximum matching ratio. For future study, we suggest that matching should be done in two phases. The first phase is to identify the maximum matching ratio. Then, the second phase is to carry out matching using the maximum matching ratio several times and take an average statistics. Using a two-phase matching, researchers can obtain stable results and draw unbiased study conclusions accordingly.

ACKNOWLEDGMENT

This work was supported in part by MOST, Taiwan, Grant No. 104-2221-E-001-021-MY3 and Multidisciplinary Health Cloud Research Program: Technology Development and Application of Big Health Data. Academia Sinica, Taipei, Taiwan. The authors would like to thank Mei-Lien Pan, Hsiao-Mei Tsao and Da-Wei Wang for their useful comments for the paper writing and proofreading.

REFERENCES

[1] E. A. Stuart, “Matching methods for causal inference: A review and a look forward,” Statistical science: a review journal of the Institute of Mathematical Statistics, vol. 25, no. 1, 2010, p. 1.

[2] P. R. Rosenbaum and D. B. Rubin, “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score,” The American Statistician, vol. 39, no. 1, 1985, pp. 33–38.

[3] H. Kawabata, M. Tran, and P. Hines, “Using SAS® to match cases for case control studies,” in Proceeding of the Twenty-Ninth Annual SAS® Users Group International Conference, vol. 29, 2004, pp. 173–29.

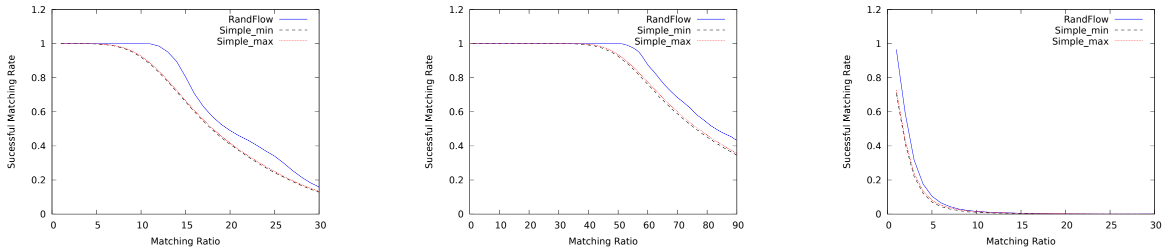
[4] D. E. Ho, K. Imai, G. King, and E. A. Stuart, “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference,” Political analysis, vol. 15, no. 3, 2007, pp. 199–236.

[5] ——, “Matchit: Nonparametric preprocessing for parametric causal inference,” Journal of Statistical Software, 2007, pp. 1–28.

[6] B. B. Hansen, “Full matching in an observational study of coaching for the SAT,” Journal of the American Statistical Association, vol. 99, no. 467, 2004, pp. 609–618.

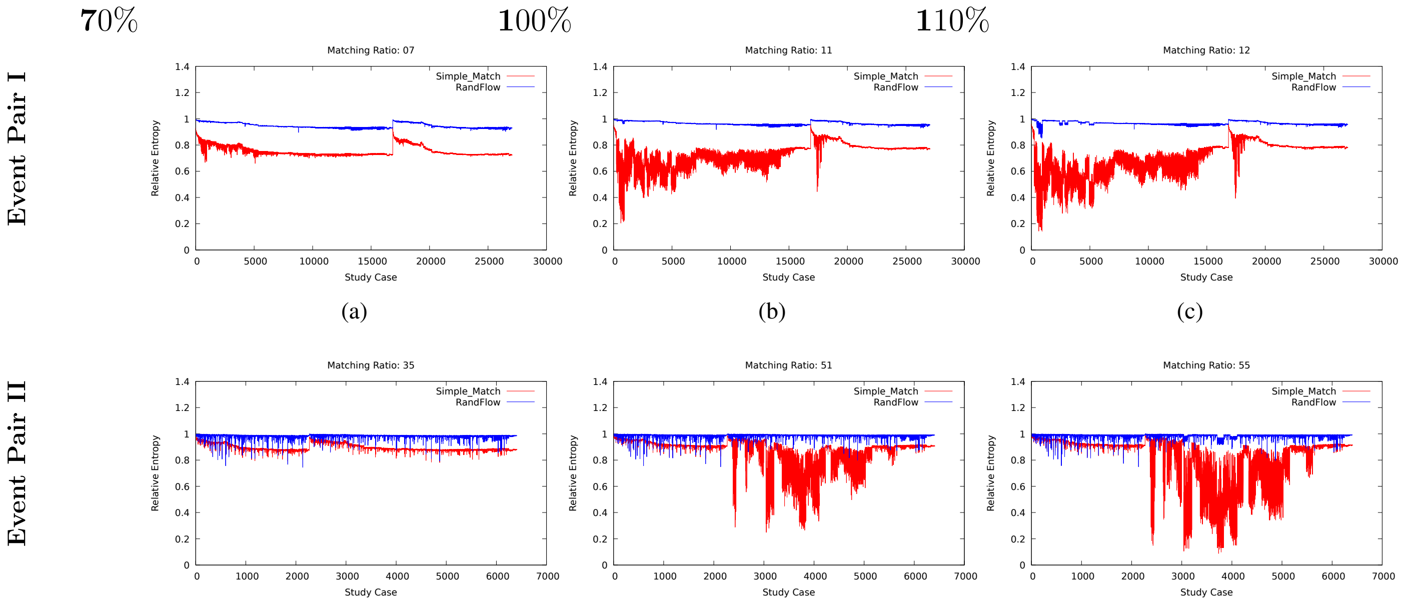
[7] L. R. Ford Jr. and D. R. Fulkerson, “Maximal flow through a network,” Canadian journal of Mathematics, vol. 8, no. 3, 1956, pp. 399–404.

[8] S. Wacholder, D. T. Silverman, J. K. McLaughlin, and J. S. Mandel, “Selection of controls in case-control studies: Iii. design options,” American journal of epidemiology, vol. 135, no. 9, 1992, pp. 1042–1050.



(a) Event Pair I (b) Event Pair II (c) Event Pair III

Figure 4. Successful matching rate of Simple_Match and RandFlow. Simple_min and Simple_max represent the minimal and maximal matching rate from the 100 trials run by Simple_Match.



(d) (e) (f)

Figure 5. Relative entropy of Simple_Match and RandFlow in Event Pair I, II.

[9] D. A. Grimes and K. F. Schulz, “Compared to what? finding controls for case-control studies,” The Lancet, vol. 365, no. 9468, 2005, pp. 1429–1433.

[10] M.-L. Pan, L.-R. Chen, H.-M. Tsao, and K.-H. Chen, “Relationship between polycystic ovarian syndrome and subsequent gestational diabetes mellitus: a nationwide population-based study,” PloS one, vol. 10, no. 10, 2015, p. e0140544.

[11] K.-J. Tien et al., “Obstructive sleep apnea and the risk of atopic dermatitis: A population-based case control study,” PloS one, vol. 9, no. 2, 2014, p. e89656.

[12] S. Hennessy, W. B. Bilker, J. A. Berlin, and B. L. Strom, “Factors influencing the optimal control-to-case ratio in matched case-control studies,” American Journal of Epidemiology, vol. 149, no. 2, 1999, pp. 195–197.

[13] K. J. Rothman, S. Greenland, and T. L. Lash, Modern epidemiology. Lippincott Williams & Wilkins, 2008.

[14] T. Faresjö and Å. Faresjö, “To match or not to match in epidemiological studiesxsame outcome but less power,” International journal of environmental research and public health, vol. 7, no. 1, 2010, pp. 325–332.

[15] “National health insurance research database, Taiwan , National Health Insurance Administration, Ministry of Health and Welfare, Taiwan, R.O.C.” retrieved: February, 2019. [Online]. Available: http://nhird.nhri.org.tw/en/index.htm

[16] C. E. Shannon, “A mathematical theory of communication,” ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, 2001, pp. 3–55.

[17] H.-J. Chang, Y.-H. Hsu, C.-W. Hsueh, and T.-S. Hsu, “Efficient randomized algorithms for large-scaled exact matching with multiple controls: Implementation and applications,” Institution of Information Science, Academia Sinica, Taiwan, Tech. Rep. TR-IIS-17-005, 2017.

[18] M.-L. Pan et al., “Bidirectional association between obstructive sleep apnea and depression: A population-based longitudinal study,” Medicine, vol. 95, no. 37, 2016, p. e4833.

[19] M.-L. Pan, C.-C. Hsu, Y.-M. Chen, H.-K. Yu, and G.-C. Hu, “Statin use and the risk of dementia in patients with stroke: A nationwide population-based cohort study,” Journal of Stroke and Cerebrovascular Diseases, 2018.

[20] L. S. Parsons, “Reducing bias in a propensity score matched-pair sample using greedy matching techniques,” vol. 26, 01 2001, pp. 214–226.