Compreendendo os métodos de escore de propensão na pesquisa clínica

Understanding the methods of propensity score in clinical research

Leonardo Roever[1], André Rodrigues Durães[2]

RESUMO

Os métodos de escore de propensão são a probabilidade de um sujeito receber um tratamento condicional em um conjunto de características de base (confundidores), sendo usado para comparar pacientes com distribuição similar de fatores de confusão, de modo que a diferença nos resultados forneça estimativa imparcial do efeito do tratamento. Esta revisão mostra os conceitos básicos dos escore de propensão e fornece orientação na implementação de métodos de propensão, além de outros, como estratificação, ponderação e ajuste de covariáveis, tornando-se uma guia prático para o clínico.

Descritores: Escore de propensão; Modelos estatísticos; Ensaios clínicos controlados aleatórios como assunto; Fatores de confusão (epidemiologia).

ABSTRACT

The propensity score methods are the probability of a subject receiving conditional treatment on a set of baseline characteristics (confounders), and are used to compare patients with similar confounding distributions, so that the difference in results provides an unbiased estimate of the treatment effect. This review shows the basic concepts of propensity scores, and provides guidelines for the implementation of propensity methods, and others based on it, such as stratification, weighting, and adjustment of covariables, becoming a practical guide for the clinician.

Keywords: Propensity score; Models, statistical; Randomized controlled trials as the subject; Confounding factors (epidemiology).

INTRODUÇÃO

Ensaios clínicos randomizados são considerados como o estudo mais cientificamente rigoroso para investigar a eficácia do tratamento, minimizando o viés sistemático. De fato, os sujeitos são aleatoriamente atribuídos ao tratamento ou ao grupo controle, permitindo, assim, a distribuição igualitária entre os dois grupos e balanceando os fatores de confusão medidos e não medidos (variáveis que influenciam tanto na variável dependente quanto a na variável independente, causando uma associação espúria, referida como covariável no contexto de regressão).[1]

No entanto, estudos randomizados podem ser difíceis de se conduzirem, e estudos de observação podem fornecer evidências importantes. Em estudos observacionais, os indivíduos nos grupos de tratamento e controle provavelmente diferem entre os fatores de confusão, e as diferenças nos resultados podem refletir variações nas condições de base, em vez de um efeito real de tratamento.[1]

Combinar cada variável, no grupo de tratamento, com sujeitos, no grupo controle, com confundidores de linha de base comparáveis é uma maneira intuitiva de minimizar confusões em estudos observacionais. No

[1] Universidade Federal de Uberlândia, Uberlândia, MG, Brasil.
[2] Universidade Federal da Bahia, Salvador, BA, Brasil.
Data de submissão: 12/11/2018. **Data de aceite:** 02/12/2018.
Fontes de auxílio à pesquisa: não há. **Conflito de interesse:** não há.
Autor correspondente: Leonardo Roever. Universidade Federal de Uberlândia – Avenida Pará, 1.720 – Umuarama
CEP 38400-902 – Uberlândia, MG, Brasil – Tel.: (34) 98803-9878 – E-mail: leonardoroever@hotmail.com

entanto, a correspondência simultânea em poucos confundidores é um processo muito complexo e geralmente resulta em um número muito limitado de correspondências semelhantes. Um método alternativo é a correspondência com base no escore de propensão (EP).[2]

O EP é a probabilidade de um sujeito receber um tratamento T condicional ao conjunto de fatores de confusão (X), sendo comumente estimado por meio de regressão logística. O objetivo de estimar o EP é simplificar o processo de correspondência, ao reduzir todos os fatores de confusão em um único valor. A correspondência de pacientes com um EP estimado similar cria um equilíbrio aproximado para todos os fatores de confusão, e a diferença nos resultados dentro dos grupos com um EP similar fornece estimativa imparcial do efeito do tratamento.[3,4]

A correspondência de EP pode contornar algumas limitações da modelagem de regressão multivariada padrão[5] (Quadro 1) e tem aparecido cada vez mais em pesquisas clínicas.[6] Este artigo fornece uma orientação para a implementação de métodos baseados em EP, para promover transparência e consistência, e facilitar a interpretação dos resultados do estudo.[1]

MÉTODOS DE ESCORE DE PROPENSÃO

Existem quatro métodos diferentes baseados em EP: (i) correspondência: corresponde a 1 ou mais casos de controle com EP, que é (quase) igual ao EP para cada caso de tratamento; (ii) estratificação (subclassificação): divide a amostra em estratos, ordenados por classificação, e as comparações entre grupos são realizadas dentro de cada estrato; (iii) ponderação: pondera casos pelo inverso do EP, é semelhante ao uso de amostragem de pesquisa, e pesos são usados para garantir que as amostras são representativas de populações específicas; e (iv) ajuste de regressão: isso inclui EP como covariável em modelo de regressão usado para estimar o efeito do tratamento. O método EP deve ser escolhido principalmente com base na estimativa de interesse, que depende da questão de pesquisa e da população-alvo.[1]

Quadro 1. Vantagens da correspondência de escore de propensão sobre a regressão multivariável padrão

Problema com regressão: multivariável	Comentários	Vantagens da correspondência de EP
Número restrito de fatores de confusão no modelo	No modelo MV, o número de confundidores é limitado pelo número de eventos. Uma regra comum é uma covariável para cada 8 a 10 eventos. Isso limita a aplicação do modelo multivariável, particularmente em caso de grande número de confundidores e relativamente baixo número de eventos	Para o cálculo do EP, o número de confundidores usado no modelo EP não é limitado pelo número de eventos resultantes. O escore permite que o investigador inclua todos os possíveis fatores de confusão, que, de outra forma, pode não ter sido possível incluir e melhorar a eficiência estatística. Portanto, o uso do EP pode ser garantido quando o número de confundidores é grande ou o número de resultados é pequeno
Invalidação do estudo por confundir	Pacientes com contraindicações ao tratamento experimental (ou aqueles com indicações absolutas) podem não ter Sujeitos expostos (ou sujeitos não expostos) para estimativa válida de diferenças relativas ou absolutas nos resultados. Esses assuntos geralmente não são reconhecidos com modelagem de resposta convencional e podem ser influenciados devido à modificação da medida de efeito ou ao modelo incorreto	A correspondência no EP se concentra diretamente nas indicações para o tratamento experimental. Faz a comparação gráfica dos EP em sujeitos expostos vs. não expostos. Podem identificar estas áreas de não sobreposição, que são de outro modo difíceis de descrever em uma configuração multivariada, com muitos fatores que influenciam nas decisões de tratamento
Suposição de modelagem	O modelo de regressão MV baseia-se nos pressupostos de modelagem de linearidade entre as covariáveis e o logaritmo natural das chances do resultado	Correspondência pelo EP elimina a hipótese de linearidade entre o EP e os resultados
Design de modelo não separado do resultado da análise	Modelos de regressão MV ajustam-se para confundidores modelando relacionamento entre covariáveis e desfecho e, assim, a especificação do modelo pode ser influenciada pela expectativa do pesquisador para provar a hipótese original	Correspondência de EP estima o efeito do tratamento por modelagem das covariáveis e do tratamento. Espelhos de correspondência EP são utilizados em um experimento randomizado, porque o desenho do estudo (modelo EP e correspondência) é separado do resultado da análise. Isso protege de suspeitas reais ou viés. por parte do pesquisador

MV: multivariável; EP: escore de propensão.

As estimativas mais comuns são o *average treatment effect on the treated* (ATT), que é o efeito para aqueles no grupo de tratamento, e o efeito de tratamento médio (ATE, do inglês *average treatment effect*), que é o efeito em todos os indivíduos (tratamento e controle). A ATE é mais interessante se todo tratamento for potencialmente oferecido a todos os sujeitos, enquanto a ATT é preferível quando as características do paciente são mais propensas a determinar o tratamento recebido. A correspondência pode estimar apenas o ATT, a ponderação pode estimar os efeitos com base em como os pesos são definidos, a estratificação pode estimar os efeitos com base em como os estratos são ponderados e, por fim, o ajuste de covariáveis pode estimar apenas o efeito marginal, mas nem o ATT nem o ATE.[1]

Ao estimar o efeito do tratamento sobre desfechos binários (*odds ratio*), os resultados da comparação em estimativas apresentam menos viés do que a estratificação ou ajuste de covariável. A probabilidade inversa de ponderação de tratamento (PIPT) deve ser usada para estimar as diferenças de risco, particularmente quando o interesse é estimar o ATE. Ao estimar o efeito do tratamento nos resultados de tempo até o evento, a correspondência e o PIPT resultam em estimativas menos enviesadas do que a estratificação ou o ajuste de covariável. Ao estimar o efeito do tratamento nos resultados de tempo até o evento, a correspondência e o PIPT resultam em estimativas menos enviesadas do que a estratificação ou o ajuste de covariável. Recomendamos o uso dos seguintes pacotes R: MatchIt ou não aleatório, não aleatório para estratificação e twang para ponderação.[1]

Etapas na análise baseada na pontuação de propensão

A seguir, estão as etapas básicas para remover os efeitos de confusão do efeito do tratamento: decidir sobre confundidores, para que equilíbrio seja alcançado; medida (por exemplo, o EP; condição na medida de distância (e usando correspondência, ponderação ou subclassificação); avaliar o equilíbrio nas covariáveis de interesse, pois a a análise baseada em EP é um processo iterativo, e processos e métodos alternativos baseados em EP devem ser tentados, até que uma amostra bem equilibrada seja obtida; estimar efeito do tratamento na amostra condicionada.[1]

Seleção de confundidores

Confundidores (X) utilizados para o modelo EP não devem ser influenciados pelo tratamento (T) e precisam ser medidos (observados antes que T seja dado). Uma possível explicação para a atribuição do tratamento deve ser fornecida, incluindo a preferência do médico, políticas ou mudanças temporais na prática. Condições preexistentes em unidades de controle para as quais um determinado tratamento não é aplicável são removidas da população do estudo. Uma seleção não parcimoniosa de confundidores é recomendada para reduzir o viés residual. No entanto, a inclusão de muitos fatores de confusão pode reduzir o número de boas correspondências e diminui a precisão.[1]

Uma abordagem razoável é incluir os fatores de confusão relacionados ao resultado e à atribuição do tratamento, se o tamanho da amostra for grande, e se concentrar em variáveis em que se acredita serem fortemente relacionadas ao resultado, se a amostra for pequena.[1]

Cálculo de pontuação de propensão

A maioria das aplicações do EP utiliza a regressão logística para estimar a pontuação[3,4] do tratamento.[7-9] A capacidade preditiva do modelo incluído não deve representar limitação na construção do modelo EP. Na verdade, o modelo EP não é empregado para fins inferenciais, mas simplesmente para criar um equilíbrio, e a prática comum de relatar a estatística-C como medida da adequação de um EP é questionável. Estatística-C muito alta pode indicar não sobreposição na distribuição do EP entre indivíduos tratados e não tratados, e sugere incapacidade de fazer comparações entre sujeitos tratados e não tratados. Além disso, alta estatística-C não pode ser tomada como prova de que o EP incluía todos os fatores importantes de confusão.[1]

Coincidindo

Dois métodos de correspondência comumente selecionados são a mais próxima correspondência vizinha e a correspondência ideal.[3,4] Vizinho mais próximo depende de um algoritmo guloso, que seleciona um participante tratado aleatoriamente e sequencialmente se move por meio da lista de participantes, coincidindo com a unidade tratada com a correspondência mais próximado grupo de comparação. O algoritmo de correspondência ideal minimiza a distância total entre os grupos correspondentes.[1]

Existem várias opções para aumentar a qualidade das correspondências: correspondência com substituição e correspondência com ajuste de pinça. Dentro da correspondência com a substituição, um participante de controle pode ser emparelhado várias vezes, se o EP dessa pessoa fornecer a correspondência mais próxima de participantes de múltiplas intervenções. Combinar com substituição requer que os erros padrão sejam estimados usando métodos, e estimadores sanduíches como dados não são mais independentes, resultando em perda de precisão.[7,8]

A combinação de pinças usa distância pré-especificada, dentro da qual as correspondências são consideradas aceitáveis. Se a melhor correspondência estiver fora da distância do cursor, as partidas não são incluídas no conjunto final. A distância designada é geralmente uma fração de um desvio padrão do logit do EP (por exemplo, 0,20 SD).[9]

Simulações de Monte Carlo mostraram que, quando comparados com outros métodos, o ajuste de calibre resulta em estimativas com menos vieses, quando comparado com a correspondência ótima e de vizinhos mais próximos, e o melhor desempenho foi avaliado usando erro quadrático médio. A correspondência com a substituição não tem desempenho superior quando comparada com comparação de calibradores sem substituição.[10] Comparada com outros métodos, a correspondência de calibradores resulta em estimativas com menor viés, quando comparada com a correspondência ótima e mais próxima, e tem o melhor desempenho quando avaliada usando média quadrática erro. A correspondência com a substituição não tem desempenho superior quando comparada com a correspondência de calibradores sem substituição.[11]

Finalmente, o número de unidades de comparação selecionadas para cada unidade tratada deve ser >1 quando houver poucas boas combinações nos grupos de controle para cada unidade tratada, pois maior proporção aumenta a precisão. No entanto, se houver um número limitado de unidades de comparação, uma proporção >1 pode ser selecionada como mal correspondida, levando a uma tendência.[11]

Estratificação

A estratificação subclassifica os indivíduos com base nos quantis dos EPs[12]. Os resultados dos indivíduos são então comparados dentro de cada um dos estratos, e um estimador comum do efeito de tratamento é derivado da combinação dos resultados ao longo dos cinco estratos.[1]

Uma prática comum é dividir o EP em cinco estratos. Este tem mostrado eliminar 90% do viés de confundidores medidos. A estratificação aproxima correspondência sem executar o risco de perder pacientes incomparáveis. Outra vantagem da técnica de estratificação é que permite o cálculo de ambos o ATE e o ATT. As estimativas de efeito específicas do estrato são ponderadas pela proporção de indivíduos que se encontram nesse estrato. Assim, quando a amostra é estratificada em n estratos de tamanho igual, os pesos específicos do estrato de 1/n são comumente usados quando se agrupam os efeitos do tratamento específico do estrato, permitindo estimar o ATE. O uso de pesos específicos do estrato, que são iguais à proporção de sujeitos tratados que se encontram dentro de cada estrato, permite estimar o TCA.[12]

Uma desvantagem da estratificação é que ela reduz os vieses menos do que outros métodos em particular à análise de sobrevivência.[8] Outra desvantagem é a complexidade de agrupar os efeitos dos estratos (por exemplo, o uso do método de Cochran-Mantel-Haenszel).[12]

Probabilidade inversa de ponderação de tratamento

O EP também pode ser usado como pesos inversos nas estimativas do ATE, conhecido como PIPT.[13] O peso de cada participante é calculado usando duas variáveis: T (indicador do *status* de tratamento do participante, sendo zero no braço de controle e 1 no braço de tratamento) e EP de cada participante. O peso (w) do participante [wATE = T / EP + (1 - T) / (1 - EP)] é igual ao inverso da probabilidade de receber o tratamento recebido pelo participante.[1]

Nesta abordagem, as contribuições dos sujeitos do estudo são ponderadas por 1/EP para pacientes experimentais e por 1/(1 - EP) para pacientes de controle. No entanto, um conjunto diferente de pesos permite estimar o ATE no tratado (ATT): wATT: T + EP (1-T)/(1-EP). Os indivíduos tratados recebem peso 1. Assim, a amostra tratada está sendo usada como a população de referência, para a qual as amostras tratadas e de controle estão sendo padronizadas.[1]

Além disso, para cenários com mais de dois tratamentos, a ponderação de EP inversa (PEPI) com o EP estimado por meio de modelos generalizados impulsionados pode ser implementada usando esses escores para estimar pesos e efeitos causais. As vantagens de usar a PEPI é que ela retém todos os dados do paciente e reduz o viés mais do que a estratificação e o ajuste de covariável.[8]

Ajuste de covariável

O EP pode ser usado como covariável no ajuste do efeito do tratamento para diferenças de linha de base. Sua vantagem é que o próprio EP pode incluir muitas covariáveis junto de interações. Isso permite que o modelo de regressão de covariáveis subsequente seja mais parcimonioso, incluindo apenas as covariáveis relevantes, juntamente da propensão variável de pontuação. No entanto, uma avaliação formal do equilíbrio entre os grupos de tratamento não é possível. Além disso, produzem estimativas mais tendenciosas, e suposições errôneas sobre a relação funcional entre EP e resultado (linearidade e risco proporcional) podem levar diretamente a estimativas enviesadas.[8]

Verificando o equilíbrio

A qualidade dos jogos é baseada na comparação da distribuição de fatores de confusão na amostra casada.[14]

O uso de testes de hipóteses e valores de p para comparar o equilíbrio não é apropriado, porque não há inferências em relação a uma população.

Eles também combinam mudanças no equilíbrio com mudanças no poder estatístico. Os vieses padronizados (também conhecidos como diferença média padronizada) são recomendados para avaliar o equilíbrio das covariáveis entre os dois grupos.[14]

A diferença média padronizada compara a diferença de médias em unidades do desvio padrão agrupadas.[14] Um valor maior que 0,10 (10%, caso seja reportado como porcentagem) é comumente considerado índice de desbalanceamento residual. A falta de equilíbrio pode indicar a necessidade de adicionar termos de ordem superior ou não lineares. Os termos de interação também devem ser considerados, em particular entre as covariáveis mais desequilibradas. A análise também pode ser restrita apenas àqueles sujeitos com EP que se sobrepõem a outro grupo (suporte comum).[15]

O diagnóstico gráfico pode ser útil para obter uma avaliação rápida do equilíbrio da covariável na presença de muitas covariáveis. O primeiro passo é examinar a distribuição do EP nos grupos usando o histograma, e também o gráfico das diferenças padronizadas de médias (nos dá uma visão geral de se o equilíbrio é adequado.[1].

Estimativa do efeito do tratamento

Os dados correspondentes devem ser analisados usando procedimentos para análises casadas, como testes *t* pareados para variáveis contínuas, enquanto a regressão logística de teste de McNemar, logit condicional ou efeito misto (pares combinados como efeito aleatório) podem ser usados para desfechos binários.[16]

Para os resultados de tempo até o evento (sobrevivência), o teste *log-rank* estratificado, o modelo de Cox estratificado ou o modelo de efeito misto de Cox são necessários.[3] Os dados também podem ser analisados usando uma regressão padrão na amostra casada, que inclui um indicador de tratamento e as variáveis usadas no modelo EP (duplo robusto), no qual o ajuste de regressão é usado para "limpar" o pequeno desequilíbrio residual da covariância entre os grupos. Métodos para amostras pareadas fornecem melhores estimativas, e métodos não pareados podem ser apresentados como análise de sensibilidade.[16]

Pontuação de propensão e dados ausentes

Na presença de dados ausentes, a imputação múltipla pode ser usada para criar conjuntos de dados completos, dos quais o EP pode ser estimado. Existem dois métodos propostos: (i) a média dos EPs após múltipla imputação, seguida por inferência causal, ou (ii) inferência causal usando cada conjunto dos EPs das múltiplas imputações seguidas pela média das estimativas causais. É aconselhável incluir o resultado no modelo de imputação.[17]

Relatando

Embora o uso dessa abordagem analítica tenha aumentado significativamente na pesquisa clínica, o relato atual é, muitas vezes, inadequado e ambíguo, e isso resulta em problemas de reprodutibilidade e interpretação do estudo. Para melhorar a consistência e a reprodutibilidade, um conjunto de itens a serem relatados tem sido recomendado.[18] Esses itens devem ser integrados com as categorias Fortalecendo o Relato sobre Estudos Observacionais em Epidemiologia (STROBE).[19]

Exemplo

Um estudo transversal multicêntrico foi realizado em 2015 em 30 províncias do Irã, com amostra de 4.200 estudantes da escola, com idades entre 7 e 18 anos. Foram realizados exames físicos e testes laboratoriais utilizando protocolos padrão. A análise foi conduzida com base no EP, e a regressão logística condicional foi utilizada para avaliar a associação entre sono curto (menos de 8 horas por dia) e o início do sono com a síndrome metabólica (SM) e seus componentes. Os resultados da regressão logística condicional foram relatados como OR e intervalos de confiança de 95% (IC95%). O EP foi calculado com base no modelo logístico condicional, com potenciais variáveis de confusão (idade, sexo, área de vida, comportamentos alimentares saudáveis e não saudáveis, história familiar de doenças crônicas e índice de massa corporal – IMC – parental). Dois grupos (com SM e sem SM) foram pareados com base no método de correspondência 1:1, sem substituição pelo escore. Em seguida, os percentuais de assimetria padronizados (AP) antes e depois da sincronização foram calculados e, em seguida, a média AP foi calculada para todas as variáveis. Devido à correspondência do EP, a regressão logística condicional foi utilizada para avaliar a associação entre a duração curta do sono (como variável categórica) e o início do sono (como variável contínua) com SM e seus componentes.[20]

No geral, 3.843 dos participantes completaram a pesquisa, no modelo multivariado. Os indivíduos que dormiam menos de 8 horas por dia tinham probabilidade significativamente maior de SM (OR=2,05; IC95% 1,19-3,63) e pressão arterial elevada (OR=1,46; IC95% 1,04-2,06). A associação entre a duração curta do sono com outros componentes da SM (incluindo obesidade abdominal, hipertrigliceridemia, hiperglicemia e baixos níveis de lipoproteína de alta densidade) não foi esta-

tisticamente significativa (p>0,05). Além disso, a associação entre o início do sono com SM e seus componentes não foi estatisticamente significante (p>0,05). A curta duração do sono foi associada ao aumento do risco de SM e PA elevada em crianças e adolescentes.[20]

Limitações

A principal limitação dos métodos EP é sua incapacidade de controlar confundimentos não mensurados. Uma desvantagem da correspondência é um tamanho de amostra muitas vezes substancialmente reduzido porque, para alguns pacientes, as correspondências podem não ser encontradas. Isso pode afetar significativamente as conclusões finais do estudo, que então se aplicam apenas ao subconjunto selecionado de pacientes que poderiam ser correspondidos.[1]

O EP tende a funcionar melhor em amostras maiores, e os desequilíbrios significativos de certas covariáveis podem ser inevitáveis, apesar de EP bem construído, secundário a um pequeno número de observações. Como estudos randomizados, os métodos EP geram efeito médio e, portanto, não abordam qual tratamento pode ser adequado para um determinado paciente.[1]

CONCLUSÃO

Os métodos EP reduzem um conjunto de fatores de confusão em uma única e intuitiva variável que aperfeiçoa a correspondência e possibilita o ajuste estatístico, quando a proporção de eventos para confundidores é baixa. Eles também podem revelar casos em que as populações de pacientes são muito divergentes para fazer comparações significativas.

REFERÊNCIAS

1. Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg. 2018;53(6):1112-7.
2. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics. 1996;52(1):249-64.
3. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399-424.
4. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25(1):1-21.
5. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26(1):20-36.
6. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008; 27(12):2037-49.
7. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078-94.
8. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013;32(16):2837-49.
9. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stumer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56.
10. Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf. 2011;20(3):317-20.
11. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014;33(6):1057-69.
12. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24(2): 295-313.
13. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013;32(19):3388-414.
14. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity score matched samples. Stat Med. 2009;28(25):3083-107.
15. Blackstone EH. Comparing apples and oranges. J Thorac Cardiovasc Surg. 2002;123(1):8-15.
16. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med. 2011;30(11):1292-301.
17. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med. 2009;28(9):1402-14.
18. Yao X, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. J Natl Cancer Inst. 2017;109(8).doi: 10.1093/jnci/djw323.
19. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335 (7624):806-8.
20. Hemati Z, Mozafarian N, Heshmat R, Ahadi Z, Motlagh ME, Ziaodini H, et al. Association of sleep duration with metabolic syndrome and its components in children and adolescents; a propensity score-matched analysis: the CASPIAN-V study.Diabetol Metab Syndr. 2018;10:78. doi: 10.1186/s13098-018-0381-y. eCollection 2018.