Observational studies: a perpetual quest for a signal among the noise

Nelson Gonzalez Valencia, MD, MSc

Received: 2 December 2021 / Revised: 2 December 2021 / Accepted: 2 December 2021 / Published online: 10 February 2022
© Canadian Anesthesiologists' Society 2022

During the perioperative care of patients undergoing major surgeries—including cardiac and hepatic procedures—there has been widespread interest in interventions to decrease the morbidity and mortality associated with intraoperative bleeding and allogenic blood transfusions. In the context of liver transplant surgery, such interventions include acute normovolemic hemodilution, intraoperative cell salvage, point-of-care coagulation monitoring through viscoelastic tests, recombinant factor VIIa administration, and the use of antifibrinolytics.¹

Considering the physiologic relationship between the splanchnic and systemic circulation, particularly in liver disease, two approaches to manage intravascular volume therapy aimed at minimizing bleeding during liver transplantation have been conceptualized.² The first seeks a central and splanchnic volume reduction by volume restriction, with or without phlebotomy, to maintain low central venous pressure (CVP) and then “refill” the circulation after the reperfusion phase; the second aims for splanchnic decongestion using vasoconstrictors. In clinical practice, both are often used in combination.

In this issue of the *Journal*, Carrier *et al.*³ report findings from an observational study on a possible beneficial effect of phlebotomy among the *whole* population of liver transplant recipients in terms of reduced bleeding and transfusion (bleeding marginal multiplicative factor, 0.85; 95% confidence interval [CI], 0.72 to 0.99; perioperative risk difference [RD] for transfusion, −15.2% [95% CI, −26.1 to −0.8]; and intraoperative RD, −14.7%; 95% CI,

−23.3 to −2.8) while showing no effect on the treated population (more on the distinction between whole and treated later). The primary mode of analysis was inverse probability of treatment weight (IPTW) analysis—a form of propensity score (PS) method. This statistical technique was used to maximize and separate the “signal” in their observational data set from the “noise” inherent in data sets derived from routinely collected clinical information.

The signal-to-noise ratio is a measure that compares the level of a desired signal to the level of background noise. Using this paradigm, let us consider the incidence of a particular study’s primary outcome as the signal. For the noise quantification, we could infer that there would be a spectrum of values which vary as determined by numerous factors, including the study design. Hence, the magnitude of noise would be the lowest when data came from a methodologically robust randomized controlled trial (RCT) with appropriate sample sizes as well as adequate blinding and follow up, permitting the signal to shine brightly. In turn, the level of noise would increase significantly when moving toward observational studies, and would escalate dramatically in the context of low-quality research with a large degree of uncertainty and highly biased clinical (anecdotal) impressions.

Scientific information to inform clinical decisions comes in different forms; nevertheless, RCTs are considered the “gold standard” for determining causality and provide confidence in a study’s results. If the sample is large enough and there is careful control in the design and the process of randomization, this will, on average, ensure roughly equal and unbiased distributions of known and unknown prognostic factors. As a consequence, RCTs permit causal inferences on treatment effects such as the average treatment effect (ATE). Nevertheless, despite our

N. Gonzalez Valencia, MD, MSc (✉)
Department of Anesthesia & Perioperative Medicine, Schulich
School of Medicine & Dentistry, Western University, 339
Windermere Rd, Office C3-104, London, ON N6A 5A5, Canada
e-mail: Nelson.gonzalezvalencia@lhsc.on.ca

confidence in their results, RCTs are not always possible because of ethical or logistic constraints.

As for noise, non-randomized studies (i.e., observational or quasi-experimental) are ubiquitous in the medical literature and can provide “real world” data. In fact, high-quality observational studies can generate credible evidence of interventions’ effects. Nevertheless, they inherently carry the major limitation that the intervention and control groups could differ systematically with respect to known and/or unknown prognostic factors, and they often have selection bias (or “confounding by indication”) when treatments (e.g., phlebotomy) are chosen by a physician rather than randomly. Attempts to draw conclusions under these conditions will inevitably result in many biased decisions. In this context, adjustment for prognostically important differences becomes necessary to obtain less biased estimates, a process that has typically been achieved through multivariable regression models.

Another approach is the use of PS methods; introduced almost four decades ago,⁴ they are gaining popularity as a good alternative for the analysis of non-randomized intervention trials. They have epistemological advantages over conventional multivariable regression modelling to explore potential causal relationships, possibly producing less noise. Interestingly, observational studies analyzed with PS methods can change the target population (the group of patients for whom the conclusions are drawn) by shifting the distribution of patient characteristics that contribute to analysis. This strategy allows determination of the ATE and the average treatment effect on the treated (ATT). The ATE reflects the effect of the treatment in the scenario that every patient within the population was offered the treatment. In essence, this shifts the entire population from untreated to treated. This contrasts with the ATT, which reflects the effect of the treatment *only among those who were ultimately treated*. In a randomized trial, these two are equal in expectation; in a nonexperimental study, they may differ.

In the case of the study by Carrier *et al.*,³ the effect estimates are measures of association between the hypothetical use of phlebotomy and blood loss with respect to two target populations of patients (i.e., just the treated population “ATT” or the whole population “ATE” created for the purpose of the analysis) using IPTW. This method uses weighting to create a pseudosample in which the distribution of potential confounding factors is conditionally independent of exposure, allowing a more accurate, less biased estimation of the relationship between exposure to phlebotomy and outcome.⁵ Their analysis is elegant and appealing, though it is important for readers to recognize some nuances when interpreting and potentially applying their results.

First, the PS is a function of the quality and the size of the available data and the adequacy of the specification of the PS model (which always involves some arbitrariness). Naturally, the PS model cannot consider factors that are unknown or were not measured, which potentially could influence the treatment selection. Therefore, in the Carrier *et al.*³ study, bias due to unmeasured or unknown differences between patients who received phlebotomy treatment *vs* those who did not cannot be ruled out because, despite its strengths, the PS method is clearly not a substitute for randomization.

Second, the quality of the PS model should in part be judged based on how well patient characteristics are balanced between the pseudosample groups.⁶ Nevertheless, showing balance on the mean or median may not result in complete adjustment for confounding on a specific variable. The reason is because simply including a variable in a regression model may not be sufficient to capture and balance its influence on the outcome model (this often occurs when a continuous variable such as body mass index is dichotomized). In this regard, Reiffel⁷ recently expanded the role of appropriate discrimination of the duration and severity of some diseases and the specification of drugs and doses to prevent misleading information and conclusions when PS methods are used in cardiology studies.

Third, the research question will lead the decision as to whether an estimation of the ATE or ATT is more relevant for analysis and applicability.⁸ Estimation of the ATE (i.e., whole population) may be less appropriate or even unrealistic when a large proportion of patients are not good candidates for the treatment option. Let us consider a hypothetical PS analysis that showed the effectiveness of veno-venous bypass (VVB) to decrease hypotension during liver transplantation. The result cannot be easily translated to the whole population because of patient conditions, logistic barriers, and surgical considerations such as the variability of the VVB technique (or using the piggyback technique without caval occlusion in some centres). On the contrary, a more generic intervention like hemodynamic goal-directed therapy through vasopressors and fluids might be more applicable to the whole population if positive results come from a PS analysis.

Finally, different PS methods could have different impacts on precision, as PS is the probability that each individual will be assigned to receive the treatment of interest given their measured covariates. The conventional IPTW method assigns a weight of $1/PS$ for treated and $1/(1 - PS)$ for untreated patients, allowing individuals with underrepresented characteristics to have increased weighting in the analysis, potentially dominating the results, increasing the variance and compromising the precision.⁹ Trimming the extreme PS values has been

described as a potential option to mitigate this problem but it would be sensitive to the cut-off chosen for the purpose, at the cost of decreasing the representative sample.¹⁰ Propensity score overlap weighting techniques could overcome these limitations by assigning weights to each patient that are proportional to the probability of that patient belonging to the opposite treatment group. The resulting target population simulates the characteristics of a pragmatic randomized trial without excluding study participants from the available sample but emphasizing the comparison of patients at same probability to be treated or not. The advantages of overlap weighting are greatest when groups are initially very different. For applications in which the initial imbalances in patient characteristics between treatment groups are modest, overlap weighting yields similar results to IPTW.¹¹ Of note, like all PS methods, overlap weighting cannot adjust for patient characteristics that are not measured or unknown.

Succinctly, PS methods certainly are a powerful strategy that can potentially produce less “noise” in analyzing non-randomized trials compared with multivariable regression analysis, but they are not completely reassuring. Neither should be considered as reliable in balancing intergroup differences and prognostic factors as randomization. As an aside, to mitigate the susceptibility of observational trials to biased analyses, it is worth mentioning the growing appreciation for observational study registration, ideally detailing the variables intended to be used in the models for regression or PS, in addition to clearly prespecified statistical analyses.¹²

Carrier *et al.* have assembled a comprehensive database of patients who underwent liver transplantation at their institution, allowing us to understand their accumulated clinical experience through several scientific contributions.^{13–15} Their effort highlights the importance of having a culture of data capturing for research, data-driven decision-making, and quality improvement initiatives. In their current analysis, the researchers not only have interrogated their local database for a fresh clinical question regarding the role of intraoperative phlebotomy, integrating a causal framework, but also have extended their period of observations to January 2021, consequently including 30% more patients compared with their last trackable report.¹³

Accordingly, the clinical relevance of Carrier *et al.*'s study³ depends not only on the statistical issues discussed, but also on whether the study sample is representative of patients to whom the results could be generalized. This is particularly important given it was a single centre study. The whole target population created in their IPTW analysis seems to represent many “usual patients” undergoing liver transplantations in transplant centres across the world—excluding those patients on preoperative renal replacement

therapy or with glomerular filtration rates below 30 mL·min⁻¹·1.73 m⁻². Based on their conclusions, phlebotomy was effective in decreasing the amount of perioperative blood loss and the probability of transfusion for the ATE. Nevertheless, phlebotomy was not effective for the patients among the treated (i.e., ATT). Because this differentiation is a side effect of the statistical nuances of PS analysis, we are left with substantial uncertainty as to whether we should consider phlebotomy for almost everybody in routine liver transplantation practice, especially facing the common context of significant comorbidity, hemodynamic instability, or anemia that many liver transplant recipients endure.

After all, in the pursuit of reducing bleeding and transfusion during liver transplantation, practitioners must also consider the undeniable variability in practices and outcomes of interest in this research field.¹⁶ Despite efforts to standardize the decision to transfuse, inconsistency remains among institutions. In addition, accurate quantification of blood loss during major surgery has been elusive and represents a surrogate outcome. Finally, using CVP as a marker of volume status has been widely questioned,^{17–19} although it remains popular in liver surgery and liver transplantation for reasons regarding splanchnic and graft congestion.^{19, 20} Methods to achieve CVP reduction also vary between institutions and include the use of not only phlebotomy but also venodilator administration, inferior vena cava clamping, neuromuscular blockade, reverse Trendelenburg positioning, fluid restriction, or a combination of these techniques.

To conclude, in the perpetual quest for “signal” in observational studies, efforts to attenuate the “noise” often include the use of PS methods. Although PS methods offer some methodological advantages, they are not a substitute for randomization, and their results can still be biased. The decision to use of phlebotomy during liver transplantations must consider the ongoing uncertainty of the evidence base, the relevance of patient-centred outcomes for clinical decision-making in perioperative medicine,²¹ and the call for a clear “signal” obtained ideally through multicentred RCTs.

Les études observationnelles, ou la quête perpétuelle d'un signal dans le bruit

Au cours des soins périopératoires des patients bénéficiant de chirurgies majeures, notamment d'interventions

cardiaques et hépatiques, un intérêt généralisé s'est manifesté pour les interventions visant à réduire la morbidité et la mortalité associées aux saignements peropératoires et aux allotransfusions. Dans le contexte de la chirurgie de transplantation hépatique, ces interventions comprennent l'hémodilution aiguë normovolémique, l'autotransfusion peropératoire, le monitoring de la coagulation au point de service par des tests viscoélastiques, l'administration de facteur recombinant VIIa, et l'utilisation d'antifibrinolytiques.¹

Étant donné la relation physiologique entre la circulation splanchnique et systémique, en particulier dans les maladies hépatiques, deux approches pour prendre en charge le traitement du volume intravasculaire visant à minimiser les saignements lors d'une transplantation hépatique ont été conceptualisées.² La première vise une réduction du volume central et splanchnique par restriction volumique, avec ou sans phlébotomie, afin de maintenir une pression veineuse centrale (PVC) basse puis de « re-remplir » la circulation après la phase de reperfusion; la seconde a pour objectif la décongestion splanchnique à l'aide de vasoconstricteurs. Dans la pratique clinique, les deux sont souvent combinées.

Dans ce numéro du *Journal*, Carrier *et coll.*³ rapportent les résultats d'une étude observationnelle concluant à un effet bénéfique potentiel de la phlébotomie parmi la population *entière* de receveurs de greffes de foie en termes de réduction des saignements et des transfusions (facteur multiplicatif marginal de saignement, 0,85; intervalle de confiance [IC] à 95 %, 0,72 à 0,99; différence de risque périopératoire [DR] pour la transfusion, -15,2 % [IC 95 %, -26,1 à -0,8]; et RD peropératoire, -14,7 %; IC 95 %, -23,3 à -2,8), tout en ne montrant aucun effet sur la population traitée (nous décrivons la distinction entre population entière et traitée plus loin). Le principal mode d'analyse a consisté en l'analyse de la probabilité inverse de pondération du traitement (ou IPTW - *inverse probability of treatment weight*), une forme de méthode du score de propension (SP). Cette technique statistique a été utilisée pour maximiser et distinguer le « signal », dans leur ensemble de données observationnelles, du « bruit » inhérent aux ensembles de données dérivées des informations cliniques recueillies systématiquement.

Le ratio signal/bruit est une mesure qui compare le niveau d'un signal souhaité au niveau de bruit de fond. En nous fondant sur ce paradigme, considérons l'incidence du critère d'évaluation principal d'une étude en particulier comme étant notre signal. En matière de quantification du bruit, nous pourrions déduire qu'il y aurait un spectre de valeurs qui varieraient en fonction de nombreux facteurs, y compris de la méthodologie de l'étude. Par conséquent, l'ampleur du bruit serait la plus faible lorsque les données

sont issues d'une étude randomisée contrôlée (ERC) robuste d'un point de vue méthodologique, avec des tailles d'échantillon appropriées ainsi qu'une procédure en aveugle et un suivi adéquats, permettant ainsi au signal de briller de mille feux. En revanche, le niveau de bruit augmenterait considérablement dans les études observationnelles et croîtrait de façon spectaculaire dans le contexte d'une recherche de faible qualité, comportant un degré élevé d'incertitude et des impressions cliniques (anecdотiques) très biaisées.

L'information scientifique utilisée pour éclairer les décisions cliniques se présente sous différentes formes; les ERC sont toutefois considérées comme « l'étalon or » si l'on souhaite déterminer la causalité et augmenter la confiance envers les résultats d'une étude. Si l'échantillon est suffisamment grand et qu'il y a un contrôle minutieux dans la conception et le processus de randomisation, cela garantirait, en moyenne, des distributions à peu près égales et non biaisées des facteurs pronostiques connus et inconnus. En conséquence, les ERC permettent des inférences causales sur les effets d'un traitement, tels que l'effet moyen du traitement (EMT). Néanmoins, malgré notre confiance en leurs résultats, les ERC ne sont pas toujours réalisables, car elles sont tenues à des contraintes éthiques ou logistiques.

En ce qui touche au bruit, les études non randomisées (soit les études observationnelles ou quasi expérimentales) sont omniprésentes dans la littérature médicale et peuvent fournir des données du « monde réel ». En fait, les études observationnelles de haute qualité peuvent générer des données probantes crédibles sur les effets des interventions. Néanmoins, leur limite principale et intrinsèque réside dans le fait que les groupes d'intervention et témoin pourraient différer de manière systématique en matière de facteurs pronostiques connus et/ou inconnus; qui plus est, elles comportent souvent un biais de sélection (ou de « confusion par indication ») lorsque les traitements (par exemple, la phlébotomie) sont choisis par un médecin plutôt que de manière aléatoire. Dans de telles conditions, les tentatives de tirer des conclusions aboutiront inévitablement à de nombreuses décisions biaisées. Dès lors, un ajustement pour tenir compte des différences importantes d'un point de vue pronostique devient nécessaire pour obtenir des estimations moins biaisées – un processus généralement réalisé à l'aide de modèles de régression multivariée.

Une autre approche consiste à utiliser des méthodes de score de propension; introduites il y a près de quarante ans,⁴ ces méthodes gagnent en popularité en tant qu'alternative intéressante pour l'analyse d'études interventionnelles non randomisées. Ces méthodes ont des avantages épistémologiques par rapport aux modèles de régression multivariée conventionnels pour explorer les

relations causales potentielles, produisant peut-être moins de bruit. Il est intéressant de noter que les études observationnelles analysées à l'aide de méthodes de score de propension peuvent modifier la population cible (le groupe de patients pour lequel les conclusions sont tirées) en modifiant la distribution des caractéristiques des patients qui contribuent à l'analyse. Cette stratégie permet de déterminer l'EMT et l'effet moyen du traitement sur les personnes traitées (EMTT). L'EMT reflète l'effet du traitement dans la situation dans laquelle chaque patient de la population s'est vu offrir le traitement. Essentiellement, cela fait passer l'ensemble de la population de non traitée à traitée. Cela contraste avec l'EMTT, qui reflète l'effet du traitement *seulement parmi les individus qui ont finalement été traités*. Dans une étude randomisée, les deux sont égaux quant aux attentes; dans une étude non expérimentale, ils peuvent différer.

Dans le cas de l'étude de Carrier *et coll.*,³ les estimations de l'effet sont des mesures de l'association entre l'utilisation hypothétique de la phlébotomie et les pertes de sang par rapport à deux populations cibles de patients (c.-à-d. seulement la population traitée « EMTT » ou l'ensemble de la population « EMT » créée aux fins de l'analyse) à l'aide de l'IPTW. Cette méthode utilise la pondération pour créer un pseudo-échantillon dans lequel la distribution des facteurs de confusion potentiels est conditionnellement indépendante de l'exposition, ce qui permet une estimation plus précise et moins biaisée de la relation entre l'exposition à la phlébotomie et le devenir.⁵ Leur analyse est élégante et séduisante, mais il est important que le lecteur demeure conscient de certaines nuances lors de l'interprétation et de l'application potentielle de leurs résultats.

Premièrement, le score de propension est fonction de la qualité et de la taille des données disponibles et de l'adéquation de la spécification du modèle de score de propension (ce qui implique toujours un certain arbitraire). Naturellement, le modèle de score de propension ne peut pas tenir compte des facteurs inconnus ou qui n'ont pas été mesurés, ce qui pourrait influencer le choix du traitement. Ainsi, dans l'étude de Carrier *et coll.*,³ le biais dû à des différences non mesurées ou inconnues entre les patients ayant reçu un traitement par phlébotomie vs ceux qui n'en ont pas reçu ne peut pas être exclu car, malgré ses points forts, la méthode de score de propension n'est clairement pas un substitut à la randomisation.

Deuxièmement, la qualité du modèle du score de propension devrait en partie être jugée en fonction de l'équilibre des caractéristiques des patients entre les groupes de pseudo-échantillons.⁶ Néanmoins, même avec une moyenne ou une médiane équilibrée, il pourrait ne pas y avoir un ajustement complet qui tienne compte des facteurs de confusion sur une variable spécifique. La raison

en est que le simple fait d'inclure une variable dans un modèle de régression pourrait ne pas suffire à saisir et équilibrer son influence sur le modèle de résultat (cela se produit souvent lorsqu'une variable continue, telle que l'indice de masse corporelle, est dichotomisée). À cet égard, Reiffel⁷ a récemment élargi le rôle de discrimination appropriée de la durée et de la gravité de certaines maladies et la spécification des médicaments et des doses afin d'éviter d'obtenir des informations et des conclusions trompeuses lorsque des méthodes de score de propension sont utilisées dans des études de cardiologie.

Troisièmement, la question de recherche conduira à décider si une estimation de l'EMT ou de l'EMTT est plus pertinente pour l'analyse et l'applicabilité.⁸ L'estimation de l'EMT (c.-à-d. l'ensemble de la population) peut être moins appropriée, voire irréaliste, lorsqu'une grande proportion de patients ne sont pas de bons candidats pour l'option de traitement. Considérons une analyse de score de propension hypothétique qui a montré l'efficacité d'un pontage veino-veineux (PVV) pour diminuer l'hypotension pendant la transplantation hépatique. Le résultat ne peut pas être facilement appliqué à l'ensemble de la population en raison des conditions des patients, des obstacles logistiques et de considérations chirurgicales telles que la variabilité de la technique de PVV (ou l'utilisation d'une technique de perfusion double sans occlusion cavale dans certains centres). En revanche une intervention plus générique, comme la thérapie dirigée vers un objectif hémodynamique par des vasopresseurs et des fluides, pourrait être plus applicable à l'ensemble de la population si des résultats positifs proviennent d'une analyse par score de propension.

Enfin, différentes méthodes de score de propension pourraient avoir des impacts différents sur la précision, car le score de propension est la probabilité que chaque individu soit affecté au traitement d'intérêt en fonction de ses covariables mesurées. La méthode IPTW conventionnelle attribue un poids de $1/SP$ pour les patients traités et de $1/(1 - SP)$ pour les patients non traités, ce qui permet aux personnes présentant des caractéristiques sous-représentées de bénéficier d'une pondération accrue dans l'analyse, dominant potentiellement les résultats, augmentant la variance et compromettant la précision.⁹ La réduction des valeurs du SP extrêmes a été décrite comme une option potentielle pour atténuer ce problème, mais il serait sensible au seuil choisi à cet effet, au prix d'une diminution de l'échantillon représentatif.¹⁰ Les techniques de pondération du chevauchement des scores de propension pourraient surmonter ces limites en attribuant à chaque patient des poids proportionnels à la probabilité que ce patient appartienne au groupe de traitement opposé. La population cible résultante simule les caractéristiques

d'une étude randomisée pragmatique sans exclure les participants à l'étude de l'échantillon disponible, mais en mettant l'accent sur la comparaison des patients ayant la même probabilité d'être traités ou non. Les avantages de la pondération de chevauchement sont les plus importants lorsque les groupes sont initialement très différents. Pour les applications dans lesquelles les déséquilibres initiaux dans les caractéristiques des patients entre les groupes de traitement sont modestes, la pondération par chevauchement donne des résultats similaires à ceux de l'IPTW.¹¹ Il convient de noter que, comme toutes les méthodes de score de propension, la pondération par chevauchement ne peut pas s'ajuster aux caractéristiques du patient qui ne sont pas mesurées ou qui sont inconnues.

Pour résumer, les méthodes de score de propension constituent indubitablement de puissantes stratégies qui pourraient potentiellement produire moins de « bruit » dans l'analyse des études non randomisées par rapport aux analyses de régression multivariée, mais elles ne sont pas entièrement rassurantes. Ni les unes ni les autres ne doivent être considérées comme aussi fiables que la randomisation pour équilibrer les différences intergroupes et les facteurs pronostiques. Soit dit en passant, pour atténuer la susceptibilité des études observationnelles aux analyses biaisées, il convient de mentionner l'engouement croissant pour l'enregistrement des études observationnelles, lequel détaille, dans l'idéal, les variables destinées à être utilisées dans les modèles de régression ou de score de propension, en plus d'analyses statistiques clairement prédéfinies.¹²

Carrier *et coll.* ont rassemblé une base de données exhaustive de patients qui ont subi une transplantation hépatique dans leur établissement, et nous permettent de comprendre leur expérience clinique accumulée grâce à plusieurs contributions scientifiques.^{13–15} Leurs efforts soulignent l'importance de disposer d'une culture de collecte de données pour la recherche, de prises de décision axées sur les données et d'initiatives d'amélioration de la qualité. Dans leur analyse actuelle, les chercheurs ont non seulement interrogé leur base de données locale avec une nouvelle question clinique concernant le rôle de la phlébotomie peropératoire, intégrant un cadre causal, mais ils ont également prolongé leur période d'observations jusqu'en janvier 2021, incluant ainsi 30 % de patients en plus par rapport à leur dernier compte rendu traçable.¹³

Par conséquent, la pertinence clinique de l'étude de Carrier *et coll.*³ dépend non seulement des questions statistiques discutées, mais aussi de la question de savoir si l'échantillon de l'étude est représentatif des patients auxquels les résultats pourraient être généralisés. Ceci est particulièrement important étant donné qu'il s'agissait d'une étude monocentrique. L'ensemble de la population cible créée dans leur analyse d'IPTW semble représenter

de nombreux « patients habituels » bénéficiant de transplantations hépatiques dans des centres de transplantation à travers le monde, à l'exclusion des patients sous traitement substitutif de l'insuffisance rénale préopératoire ou dont le taux de filtration glomérulaire était inférieur à $30 \text{ mL} \cdot \text{min}^{-1} \cdot 1,73 \text{ m}^{-2}$. Sur la base de leurs conclusions, la phlébotomie a été efficace pour réduire la quantité de perte de sang périopératoire et la probabilité de transfusion pour l'EMT. Néanmoins, la phlébotomie n'a pas été efficace pour les patients parmi les patients traités (c.-à-d. EMTT). Parce que cette différenciation est un effet secondaire des nuances statistiques de l'analyse par score de propension, nous nous retrouvons avec une incertitude considérable quant à savoir si nous devrions envisager la phlébotomie pour la majorité de nos patients bénéficiant d'une transplantation hépatique de routine, en particulier face au contexte commun de comorbidité significative, d'instabilité hémodynamique ou d'anémie dont souffrent de nombreux receveurs de greffe de foie.

Après tout, dans la poursuite de la réduction des saignements et des transfusions lors d'une transplantation hépatique, les praticiens doivent également tenir compte de la variabilité indéniable des pratiques et des résultats d'intérêt dans ce domaine de recherche.¹⁶ Malgré les efforts déployés pour standardiser la décision de transfuser, des disparités subsistent entre les institutions. En outre, la quantification précise de la perte de sang lors d'une chirurgie majeure demeure difficile à réaliser et représente un résultat de substitution. Enfin, l'utilisation de la PVC comme marqueur de l'état volémique a été abondamment remise en question^{17–19} bien que, pour des raisons touchant à la congestion splachnique et du greffon, elle demeure populaire en chirurgie hépatique et en transplantation hépatique.^{19, 20} Les méthodes utilisées pour obtenir une réduction de la PVC varient également d'une institution à l'autre et comprennent non seulement la phlébotomie, mais également l'administration de vénodilatateurs, le clampage inférieur de la veine cave, les blocs neuromusculaires, un positionnement de Trendelenburg inversé, la restriction hydrique ou une combinaison de ces techniques.

Pour conclure, dans la quête perpétuelle du « signal » dans les études observationnelles, les efforts pour atténuer le « bruit » incluent souvent l'utilisation de méthodes de score de propension. Bien que ces dernières offrent certains avantages méthodologiques, elles ne remplacent pas la randomisation et leurs résultats peuvent tout de même être biaisés. La décision de recourir à la phlébotomie lors de transplantations hépatiques doit tenir compte de l'incertitude persistante de la preuve, de la pertinence des résultats centrés sur le patient pour la prise de décision clinique en médecine périopératoire,²¹ et du besoin de « signal » clair, idéalement obtenu par le biais d'ERC multicentriques.

Disclosures None.

Funding statement None.

Editorial responsibility This submission was handled by Dr. Philip M. Jones, Deputy Editor-in-Chief, *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*.

Déclaration Aucune.

Déclaration de financement Aucune.

Responsabilité éditoriale Ce manuscrit a été traité par Dr Philip M. Jones, rédacteur adjoint, *Journal canadien d'anesthésie*.

References

1. Gurusamy KS, Pissanou T, Pikhart H, Vaughan J, Burroughs AK, Davidson BR. Methods to decrease blood loss and transfusion requirements for liver transplantation. *Cochrane Database Syst Rev* 2011; DOI: <https://doi.org/10.1002/14651858.CD009052.pub2>.
2. Mukhtar A, Lotfy A, Hussein A, Fouad E. Splanchnic and systemic circulation cross talks: implications for hemodynamic management of liver transplant recipients. *Best Pract Res Clin Anaesthesiol* 2020; 34: 109-18.
3. Carrier FM, Ferreira Guerra S, Coulombe J, et al. Intraoperative phlebotomies and bleeding in liver transplantation: a historical cohort study and causal analysis. *Can J Anesth* 2022; <https://doi.org/10.1007/s12630-022-02197-1>.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41-55.
5. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11: 550-60.
6. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol* 2015; 68: 112-21.
7. Reiffel JA. Propensity score matching: the 'devil is in the details' where more may be hidden than you know. *Am J Med* 2020; 133: 178-81.
8. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; 46: 399-424.
9. Thomas L, Li F, Pencina M. Using propensity score methods to create target populations in observational clinical research. *JAMA* 2020; 323: 466-7.
10. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol* 2019; 188: 250-7.
11. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. *JAMA* 2020; 323: 2417-8.
12. Jones PM, Martin J. Increasing the reproducibility of research will reduce the problem of apophenia (and more). *Can J Anesth* 2021; 68: 1120-34.
13. Lariviere J, Giard JM, Zuo RM, Massicotte L, Chasse M, Carrier FM. Association between intraoperative fluid balance, vasopressors and graft complications in liver transplantation: a cohort study. *PLoS One* 2021; DOI: <https://doi.org/10.1371/journal.pone.0254455>.
14. Carrier FM, Chassé M, Sylvestre MP, et al. Effects of intraoperative fluid balance during liver transplantation on postoperative acute kidney injury: an observational cohort study. *Transplantation* 2020; 104: 1419-28.
15. Carrier FM, Sylvestre MP, Massicotte L, Bilodeau M, Chasse M. Effects of intraoperative hemodynamic management on postoperative acute kidney injury in liver transplantation: an observational cohort study. *PLoS One* 2020; DOI: <https://doi.org/10.1371/journal.pone.0237503>.
16. Bartoszko J, Vorobeichik L, Jayarajah M, et al. Defining clinically important perioperative blood loss and transfusion for the Standardised Endpoints for Perioperative Medicine (StEP) collaborative: a protocol for a scoping review. *BMJ Open* 2017; DOI: <https://doi.org/10.1136/bmjopen-2017-016743>.
17. Kumar A, Anel R, Bunnell E, et al. Pulmonary artery occlusion pressure and central venous pressure fail to predict ventricular filling volume, cardiac performance, or the response to volume infusion in normal subjects. *Crit Care Med* 2004; 32: 691-9.
18. De Backer D, Vincent JL. Should we measure the central venous pressure to guide fluid management? Ten answers to 10 questions. *Crit Care* 2018; DOI: <https://doi.org/10.1186/s13054-018-1959-3>.
19. Liu TS, Shen QH, Zhou XY, et al. Application of controlled low central venous pressure during hepatectomy: a systematic review and meta-analysis. *J Clin Anesth* 2021; DOI: <https://doi.org/10.1016/j.jclinane.2021.110467>.
20. Park L, Gilbert R, Baker L, et al. The safety and efficacy of hypovolemic phlebotomy on blood loss and transfusion in liver surgery: a systematic review and meta-analysis. *HPB (Oxford)* 2020; 22: 340-50.
21. Moonesinghe SR, Jackson AI, Boney O, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine initiative: patient-centred outcomes. *Br J Anaesth* 2019; 123: 664-70.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.