Scoring systems for the critically ill: *use, misuse and abuse*

Yoanna Skrobik MD FRCP(C),* Brian P. Kavanagh MB FRCP(C)†

THE Cartesian premise that 'if something can't be quantified, its existence should be questioned', is implicit in much of modern medicine. Many of us are attracted to quantification, and in a world dominated by data, the idea of numeric certainty holds allure. Scoring systems for the purpose of appraising severity and outcome have thus been developed for the critically ill.[1] While accurate and useful across groups of patients, their presence tempts clinicians to suppose that individual patients with good scores will do well, and individual patients with bad scores won't. Attractive as this may seem, it is wrong.

In the current issue of The Journal, Erhmann and colleagues examine a scoring system's performance in a particularly thorny area of critical care – establishing futility, and withdrawing life support.[2] The authors examined 93 intensive care unit (ICU) patients and determined survival as a function of the change in logistic organ dysfunction (LOD) score between admission and day four.[2] They[2] hoped that the trajectory taken by each individual's score would better map out their fate than a single measurement. In short, those who developed progressively better scores would do well and likely survive; conversely, those developing progressively worse scores would do badly and likely die. It was this latter pattern that was thought to select out those for whom limitation of care might have been appropriate.

This paper[2] revisits the use of a severity of illness score to limit care, an idea initially investigated by Atkinson *et al.* in a very controversial manuscript just over a decade ago.[3] That paper reported that by and large, patients predicted to die died, and in doing so over a prolonged ICU stay, consumed a great deal of health-care resources.[3] However, not all of those predicted to die died – some lived with a good quality of life. Far from attempting to save on the financial costs by numbers-driven limitation of care, the paper focused on the human costs of getting it wrong.[3] Thus, the current publication[2] serves as a timely reminder of three generic issues related to the application of scoring systems: use, misuse, and abuse.

How have scoring systems been put to (good) use? Knaus *et al.*'s benchmark study[4] used APACHE scores to compare actual and predicted death rates using group results as the standard. The authors then constructed a ranking of institutional outcomes to demonstrate the superiority of some centres (and their characteristics) and, naturally enough, the inferiority of others. This grading disconcertingly resembles the ranking of the country's universities by a popular Canadian magazine,[5] or the dubious designation of 'best doctor' by others. The assumption is that comparable severity of illness scores equate to comparable illness. Over large groups of patients with similar disease, this is true. If the groups are small or the populations dissimilar, then all bets are off, and validity far less likely. The flaws that are inherent in scoring systems limit the applicability of such scoring systems to quality assurance (between institutions or over time in a single centre). Similar limitations apply to the use of such ratings within clinical trials. Here, scoring systems serve three purposes: to screen trial entrants, as a 'surrogate' measure of outcome, or when implementing the results of the trial in the care of patients.

Good scoring systems provide reliable data for the probability of survival of a group of patients. However, a banal example illustrates the differences by probabilities for groups and probabilities for individuals. Suppose the weights of 1,000 consecutive patients are entered into a database and 90% of the patients weigh less than 100 kg: thus approximately one patient in ten will weigh 100 kg (or greater). There is no indication - without additional information - which of the individual patients would weigh this much. No matter how

From the Intensive Care Unit and Department of Medicine,* Hôpital Maisonneuve-Rosemont, University of Montréal, Montréal, Québec; and the Departments of Anesthesia and Critical Care Medicine,† Hospital for Sick Children and the University of Toronto, Ontario, Canada.

Address correspondence to: Dr. Brian P. Kavanagh, Department of Anesthesia, Hospital for Sick Children, 555 University Avenue, Ontario M5G 1X8, Canada. Phone: 416-813-6860; Fax: 416-813-5313; E-mail: brian.kavanagh@sickkids.ca

such data were processed, there is no predictor for the actual weight of a patient; only the *probability* of their weight is available. The clinician may be bemused or perplexed with this analogy - he or she would simply weigh the patient, and have no need for a population-based scoring system. This simplicity is instructive: when the answer is available directly from the patient, the clinician should not look to a complex statistic. The analogy with ICU scoring systems is carried further by studies documenting that providing clinicians with objective, model-based estimates of probability of survival does not influence mortality.[6,7] Finally, there are pressing biological reasons as to why we should be skeptical towards reliance on scoring systems. For example, the responses - and outcomes - of individual patients with sepsis may well be far more dependent on their particular genetic makeup (e.g., cytokine polymorphisms)[8] than on derived prognostic parameters.

Beyond the issues of use and misuse, there lies the specter of abuse. Abuse, corresponding to bad treatment or exploitation, is especially worrisome in the context of the vulnerable critically ill patient. A most disturbing possibility is that a scoring system might be used alone to justify the withdrawal of active care, or institution of palliation.

We believe that four lines of reasoning explain why we should reject the notion of limiting care based on a scoring system. *First*, as discussed above, predictors derived from populations (as in this study) will never be robust indicators of the fate of individuals. Although some systems perform well[9,10] and despite well-demonstrated weaknesses in physicians' ability predict outcome in critical illness,[11,12] Canadian intensivists predict individual patient mortality better (more than sixfold) than either of two well validated scoring systems.[13] Probabilities, unless *very* close to 100% or to 0% (and then, only if highly accurate), simply have quite limited applicability to individual patients.

Second, straddling the gap between the statistical and the actual is the uneasy notion of futility. Clinicians often invoke futility, but their numerical definition of it varies[14] as does their ability to predict it. In a much discussed study by the Canadian Critical Care Trials Group, actual ICU survival rates were higher than clinician predicted survival across all ranges of survival predictions;[13] of patients whom physicians attributed a less than 10% chance of survival, 29% left the ICU and 22% returned home.[13] Indeed 3.6% of patients in whom mechanical ventilation was withdrawn - not weaned - survived to discharge from hospital.[15] So much for futility.

Third, even if a scoring system was developed that could *accurately* predict the outcome for a specific patient in a 'will definitely live' or 'will definitely die' (soon) format, this still fails to incorporate the patient's wishes, and their beliefs. Autonomy, the right to self determination, is not dependent upon statistics, and certainly not on an uncertain declaration of 'futility'. Patient autonomy is context-specific: Canadian patients vastly prefer shared decision making to relinquishing end-of-life decision responsibility to the physician.[16] In this regard, the prognostic tool may be helpful (assuming it is valid), but only to inform the patient or their surrogate; not to decide for them. As things stand, Canada does not (yet) have statutory limitations to health care provision.

Fourth among these attempts to quantify the unquantifiable, is the issue of quality. This twofold issue speaks to the quality of the life for those who survive, and the quality of death for those who do not. Many attempts have been made to construct metrics of the quality of life among survivors (e.g., the Quality-Adjusted Life Year),[17] but all of us who appreciate the ups and downs of the many dimensions of our own lives will question the notion of a uni-dimensional metric that purports to report on whether - and by how much - a life is worth living or if it should end. Indeed the notion of incorporating quality of life into outcome predictors in the critically ill has been discussed in a previous commentary in The Journal.[18] We certainly wouldn't rely on an isolated 'mark out of ten' to decide on whether to forgo additional therapy.

Intensive care is expensive, and appears all the more so when the costs include those who didn't survive. Expensive as it is, it is getting more so and as long as money is a factor, the attempts to assist (or drive) decision making with numeric scales will be firmly in the sights of those controlling health-care budgets. Society needs to set the agenda, the ground rules and the goals for debating this issue.

In summary, we believe that the overall dynamic between population statistics *vs* individual choices is an important one. Anesthesiologists are fully aware that although mortality directly attributable to anesthesia is in the exceedingly low percentages, each one of those who die is 100% dead. We believe that such clarity can be brought to bear on the use, misuse and abuse of scoring systems for providing care in the critically ill. Patients who have a high *probability* of dying according to a population-derived scoring system should not have their care limited or withdrawn based solely on such a probability.

Acknowledgement
The authors are grateful to Dr. Stéphane Ahern for his thoughtful comments on the manuscript.

Les systèmes de notation pour les grands malades : *usage, mésusage et abus*

La prémisse cartésienne voulant que «si une chose ne peut être quantifiée, son existence doit être remise en question», est implicite dans beaucoup d'aspects de la médecine moderne. Pour nombre d'entre nous, attirés par la quantification dans un monde dominé par les données, l'idée d'une certitude numérique est prenante. Des systèmes de notation ont donc été élaborés pour évaluer la sévérité et l'évolution de l'état des malades en état critique.[1] Exacts et utiles pour des groupes de patients, ces systèmes amènent à penser que les individus qui ont de bons scores vont bien s'en tirer et ceux qui ont de mauvais scores, non. Aussi attrayante soit-elle, cette idée est fausse.

Dans le présent numéro du Journal, Erhmann et ses collègues étudient la performance de systèmes de notation dans un domaine particulièrement épineux des soins intensifs - l'établissement de la futilité, et le retrait de la réanimation.[2] Quatre-vingt-treize patients admis aux soins intensifs (USI) ont été suivis, et leur survie déterminée en fonction de la modification du score logistique de dysfonction organique (LDO) entre le jour de l'admission et le quatrième jour.[2] Ils[2] espéraient que la trajectoire de chaque score individuel permettrait mieux qu'une seule mesure de représenter l'évolution du patient. Bref, ceux dont les scores s'améliorent progressivement vont aller bien et probablement survivre et ceux dont les scores se dégradent progressivement vont aller mal et probablement mourir. C'est à ce dernier modèle qu'on a pensé pour établir les patients chez qui un arrêt de réanimation serait approprié.

Cet article[2] revoit l'utilisation du score de sévérité de la maladie pour limiter les soins, un sujet initialement exploré par Atkinson *et coll.* dans un article très controversé paru il y a un peu plus de dix ans.[3] L'article rapportait qu'en général les patients de qui on avait prédit le décès étaient morts et ce, après un séjour prolongé à l'USI et une grande consommation de ressources en soins de santé.[3] Pourtant, tous les décès prédits ne sont pas survenus, certains patients ont survécu et ont eu une bonne qualité de vie. Loin de vouloir épargner de l'argent en se basant sur des chiffres pour limiter les soins, l'article mettait l'accent sur les coûts humains de s'être trompé.[3] Ainsi, la présente publication[2] sert à rappeler la pertinence de trois aspects généraux reliés à l'application des systèmes de notation : usage, mésusage et abus.

Comment les systèmes de notation ont-ils été mis en (bon) usage ? Dans l'étude de référence de Knaus *et coll.*[4], ils ont utilisé les scores APACHE pour comparer les taux de décès prédits et réels et se sont servis des résultats de groupes comme norme. Ils ont ensuite construit un palmarès des résultats institutionnels pour démontrer la supériorité de certains centres, et leurs caractéristiques, et, bien sûr, l'infériorité des autres. Ce classement ressemble étrangement au palmarès des universités du pays, publié par un magazine canadien populaire,[5] ou à la désignation discutable du «meilleur médecin» par d'autres. L'hypothèse est que des scores de sévérité comparables égalent une maladie comparable. C'est vrai pour de grands groupes de patients qui ont des maladies similaires. Mais pour des groupes restreints ou une population diversifiée, il n'y a plus de pari qui tienne et encore moins de validité. Les défauts inhérents aux systèmes de notation limitent leur applicabilité à l'assurance de la qualité entre institutions ou, avec le temps, à l'intérieur d'un même centre. Des limites similaires s'appliquent à l'usage de ces notations dans les études cliniques. Ici, les systèmes de notation servent trois objectifs : la sélection des participants à l'étude, la mesure «porteuse» des résultats ou l'application des résultats expérimentaux aux soins des patients.

Les bons systèmes de notation fournissent des données fiables sur la probabilité de survie d'un groupe de patients. Cependant, un exemple banal illustre les différences entre les probabilités pour des groupes et pour des individus. Supposons que les poids de 1 000 patients consécutifs sont entrés dans une base de données et que 90 % des patients pèsent moins de 100 kg : il y aura donc environ un patient sur dix pesant 100 kg ou plus. Rien n'indique, sans autre information, quels patients pourraient peser autant. Peu importe comment ces données ont été traitées, il n'y a pas de prédicteur du poids exact d'un patient ; il n'y a que la *probabilité* de leur poids qui soit disponible. Le clinicien pourrait être confondu ou perplexe face à cette analogie, et pèserait tout simplement le patient sans avoir besoin d'un système de notation fondé sur une population. Cette simplicité est instructive : quand la réponse peut être obtenue directement du patient, le clinicien ne cherche pas de statistique complexe. L'analogie avec les systèmes de notation de l'USI est poussée encore plus dans des études qui montrent que de fournir aux cliniciens des estimations objectives et fondées sur des modèles de probabilité

de survie ne modifie pas la mortalité.[6,7] Finalement, d'importantes raisons biologiques expliquent pourquoi nous devrions être sceptiques face aux systèmes de notation. Par exemple, les réactions, et l'évolution, de patients atteints de septicémie pourraient dépendre beaucoup plus de leur bagage génétique (comme les polymorphismes de la cytokine)[8] que de paramètres pronostiques dérivés.

Au delà de l'usage et du mauvais usage, apparaît le spectre de l'abus. Correspondant au mauvais traitement ou à l'exploitation, l'abus est particulièrement inquiétant pour le patient gravement malade et vulnérable. La plus dérangeante possibilité serait qu'un système de notation soit utilisé seul pour justifier le retrait de soins actifs ou l'institution d'un traitement palliatif.

Quatre raisons peuvent expliquer pourquoi nous devrions rejeter l'idée d'une limite des soins fondée sur un système de notation. *Premièrement*, les prédicteurs dérivés de populations, comme dans la présente étude, ne seront jamais des prédicteurs solides de l'évolution individuelle. Même si certains systèmes affichent de bonnes performances[9,10] et malgré la faiblesse bien démontrée de l'habileté des médecins à prédire l'évolution d'une maladie grave,[11,12] les intensivistes canadiens prédisent mieux, plus que six fois mieux, la mortalité d'un patient particulier que deux systèmes de notation bien validés.[13] Les probabilités, à moins d'être *très* près de 100 % ou de 0 % et, dans ce cas, seulement si elles sont hautement exactes, n'ont qu'une applicabilité individuelle plutôt limitée.

Deuxièmement, occupant l'espace entre les statistiques et le réel se trouve l'embarrassante notion de futilité. Les cliniciens l'invoquent souvent, mais la définition numérique qu'ils en font varie comme leur habileté à la prédire.[14] Dans une étude, objet de discussions animées par le Groupe canadien sur les études en soins intensifs, les taux réels de survie à l'USI ont été plus élevés que les taux prédits par les cliniciens au moyen de tous les systèmes de prédictions de survie.[13] Parmi les patients à qui les médecins ont attribué moins de 10 % de chance de survie, 29 % ont quitté l'USI et 22 % sont retournés à la maison.[13] En effet, 3,6 % des patients chez qui la ventilation mécanique fut cessée et non sevrée, ont survécu au départ de l'hôpital.[15] Voilà pour la futilité.

Troisièmement, même si un système de notation était élaboré, qui pouvait *exactement* prédire l'évolution d'un patient en particulier sous forme de «va certainement vivre» ou «va certainement mourir» (bientôt), il ne réussirait toujours pas à incorporer les souhaits et les croyances du patient. L'autonomie, le droit à l'autodétermination, ne dépend pas des statistiques et certainement pas d'une déclaration incertaine de «futilité». L'autonomie du patient est spécifique au contexte : les patients canadiens préfèrent de beaucoup la prise de décision partagée que l'abandon au médecin de la responsabilité de la décision de fin de vie.[16] À cet égard, l'outil de pronostic peut être utile, en supposant qu'il soit valide, mais seulement pour informer le patient ou son substitut, non pour décider pour lui. Actuellement, le Canada n'a pas, encore, de restriction statutaire pour la prestation de soins de santé.

Quatrièmement, parmi ces tentatives pour quantifier l'inquantifiable se trouve la question de la qualité. Cette question double traite de la qualité de la vie de ceux qui survivent et de la qualité de la mort de ceux qui meurent. Beaucoup d'essais ont été faits pour construire des mesures de la qualité de vie des survivants, comme la survie pondérée par la qualité de vie,[17] mais nous tous qui apprécions les grandeurs et les misères de notre propre existence mettrions en doute une mesure unidimensionnelle prétendant montrer qu'une vie vaut ou non la peine d'être vécue. L'inclusion de la qualité de vie aux prédicteurs de l'évolution des grands malades a déjà été discutée dans un commentaire antérieur dans le Journal.[18] Nous ne pourrions certainement pas compter sur une seule «note sur dix» pour décider de l'abandon d'une réanimation.

Les soins intensifs sont chers et le sont plus encore quand les coûts comprennent ceux des non-survivants. Aussi chers soient-ils, ils le seront encore plus et aussi longtemps que l'argent sera un facteur ; les tentatives pour assister, ou diriger, la prise de décision avec des échelles numériques resteront dans la mire des contrôleurs de budgets des soins de santé. La société doit fixer le programme, les règles fondamentales et les objectifs pour débattre de cette question.

En résumé, il y a une importante dynamique globale entre les statistiques des populations et les choix individuels. Les anesthésiologistes savent pertinemment que même si la mortalité directement attribuable à l'anesthésie représente des pourcentages extrêmement faibles, chaque patient qui meurt est mort à 100 %. Il faut adopter la même clarté en ce qui a trait à l'usage, le mésusage et l'abus des systèmes de notation dans les soins aux grands malades. Les patients pour qui il y a une forte *probabilité* de mourir, selon un système de notation fondé sur une étude de population, ne devraient pas voir leurs soins limités ou retirés d'après cette seule probabilité.

References

1 *Le Gall JR*. The use of severity scores in the intensive care unit. Intensive Care Med 2005; 31: 1618–23.

2 *Ehrmann S, Mercier E, Bertrand P, Dequin PF*. The

logistic organ dysfunction score as a tool for making ethical decisions. Organ dysfunction & therapeutic limitations. Can J Anesth 2006; 53: 518–23.

3 *Atkinson S, Bihari D, Smithies M, Daly K, Mason R, McColl I.* Identification of futility in intensive care. Lancet 1994; 344: 1203–6.

4 *Knaus WA, Draper EA, Wagner DP, Zimmerman JE.* An evaluation of outcome from intensive care in major medical centers. Ann Intern Med 1986; 104: 410–8.

5 *Dowsett Johnston A, Dwyer M, Farran S, Kim g, Marley K, Wright D.* Universities. Rankings 2005. Waging a war for talent. URL available from; http://www.macleans.ca/universities/.

6 *Knaus WA, Rauss A, Alperovitch A*, et al. Do objective estimates of chances for survival influence decisions to withhold or withdraw treatment? The French Multicentric Group of ICU Research. Med Decis Making 1990; 10: 163–71.

7 *Murray LS, Teasdale GM, Murray GD, et al.* Does prediction of outcome alter patient management? Lancet 1993; 341: 1487–91.

8 *Mira JP, Cariou A, Grall F, et al.* Association of TNF2, a TNF-alpha promoter polymorphism, with septic shock susceptibility and mortality: a multicenter study. JAMA 1999; 282: 561–8.

9 *Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE.* APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981; 9: 591–7.

10 *Le Gall JR, Klar J, Lemeshow S, et al.* The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA 1996; 276: 802–10.

11 *Perkins HS, Jonsen AR and Epstein WV.* Providers as predictors: using outcome predictions in intensive care. Crit Care Med 1986; 14: 105–10.

12 *Poses RM, Bekes C, Copare FJ, Scott WE.* The answer to «What are my chances, doctor?» depends on whom is asked: prognostic disagreement and inaccuracy for critically ill patients. Crit Care Med 1989; 17: 827–33.

13 *Rocker G, Cook D, Sjokvist P, et al.* Clinician predictions of intensive care unit mortality. Crit Care Med 2004; 32: 1149–54.

14 *Wilson IB, Green ML, Goldman L, Tsevat J, Cook EF, Phillips RS.* Is experience a good teacher? How interns and attending physicians understand patients' choices for end-of-life care. SUPPORT Investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. Med Decis Making 1997; 17: 217–27.

15 *Cook D, Rocker G, Marshall J, et al.* Withdrawal of mechanical ventilation in anticipation of death in the intensive care unit. N Engl J Med 2003; 349: 1123–32

16 *Heyland DK, Tranmer J, O'Callaghan CJ, Gafni A.* The seriously ill hospitalized patient: preferred role in end-of-life decision making? J Crit Care 2003; 18: 3–10.

17 *Angus DC, Musthafa AA, Clermont G, et al.* Quality-adjusted survival in the first year after the acute respiratory distress syndrome. Am J Respir Crit Care Med 2001; 163: 1389–94.

18 *Marshall JC.* Measuring organ dysfunction in the intensive care unit: why and how? (Editorial). Can J Anesth 2005; 52: 224–30.