EXERCICE N° 25 : **Bien prendre en compte les risques d'erreur**

Le *tolvaptan* (Samsca°) est un antagoniste des récepteurs de la vasopressine, autorisé dans l'Union européenne pour le traitement des hyponatrémies secondaires à un syndrome de sécrétion inappropriée d'hormone antidiurétique. Son dossier d'évaluation, que nous analyserons dans un futur numéro, comporte un vaste essai, remarquablement bien conçu, chez des insuffisants cardiaques. Dans cet essai, le *tolvaptan* n'a pas été plus efficace que le placebo sur les principaux critères cliniques d'évaluation. Pour vous exercer à la lecture critique de ce type de publication, l'équipe *Prescrire* vous propose de lire des extraits du protocole de cet essai, puis de répondre à quelques questions. Suivent les réponses et les commentaires de la Rédaction.

EXTRAITS
DE LA VERSION ORIGINALE DE LA PUBLICATION (1)

Rationale and design of the multicenter, randomized, double-blind, placebo-controlled study to evaluate the efficacy of vasopressin antagonism in heart failure: outcome study with tolvaptan (Everest)

(…) Tolvaptan (OPC-41061) is a novel oral, once-daily, nonpeptide vasopressin V2 receptor antagonist without intrinsic agonist properties. Several studies, including the Acute and Chronic Therapeutic Impact of a Vasopressin antagonist in Congestive Heart Failure (ie, ACTIV in CHF) trial, demonstrated that tolvaptan, in addition to standard therapy, resulted in a significant increase in urine output and a decrease in body weight without hypokalemia or worsening renal function. This study generated the hypothesis that tolvaptan might improve survival in patients with chronic heart failure hospitalized from congestion. The EVEREST trial is designed to test this hypothesis.

Methods
EVEREST is an international, multicenter, randomized, double-blind, placebo-controlled study designed to evaluate the long-term efficacy and safety of oral once-daily tolvaptan 30 mg in patients hospitalized with worsening heart failure. Patients must have reduced left ventricular ejection fraction (≤40%), signs of volume expansion, and New York Heart Association class III/IV symptoms. (…)
EVEREST consists of 2 components: the primary outcomes study, and 2 identical embedded studies within the main trial to access global clinical status according to regulatory requirements (Fig. 1A, 1B). The primary outcomes study has 2 coprimary end points: (1) time to all-cause mortality and (2) time to first occurrence of cardiovascular mortality or heart failure hospitalization. The primary efficacy end point of the

TRADUCTION
EN FRANÇAIS DES EXTRAITS CI-CONTRE

Justification et conception de l'essai multicentrique randomisé en double aveugle versus placebo évaluant l'efficacité de s'opposer à l'action de la vasopressine dans l'insuffisance cardiaque : essai sur des critères cliniques avec le tolvaptan (Everest)

(…) Le tolvaptan (OPC-41061) est un nouvel antagoniste non peptidique des récepteurs V2 de la vasopressine sans activité agoniste intrinsèque, par voie orale en une prise quotidienne. Plusieurs essais, dont l'essai Impact thérapeutique aigu et chronique d'un antagoniste de la vasopressine dans l'insuffisance cardiaque congestive (c'est-à-dire l'essai Activ in CHF), ont démontré que le tolvaptan, ajouté au traitement standard, augmente significativement la diurèse et diminue le poids corporel sans hypokaliémie ni aggravation de la fonction rénale. Cet essai a généré l'hypothèse que le tolvaptan pourrait améliorer la survie des patients insuffisants cardiaques chroniques hospitalisés pour poussée congestive. L'essai Everest est conçu pour tester cette hypothèse.

Méthodes
Everest est un essai international multicentrique, randomisé en double aveugle versus placebo conçu pour évaluer l'efficacité et les effets indésirables à long terme du tolvaptan, 30 mg une fois par jour par voie orale chez des patients hospitalisés pour aggravation d'une insuffisance cardiaque. Les patients doivent avoir une diminution de la fraction d'éjection ventriculaire gauche (≤ 40 %), des signes congestifs, et des symptômes des classes III ou IV de la classification de la New York Heart Association. (…)
Everest est constitué de 2 composants : l'essai principal, et 2 essais identiques intégrés dans l'essai principal pour évaluer le statut clinique global selon les exigences des agences du médicament (fig. 1A, 1B). L'essai principal a 2 critères d'évaluation principaux : 1°) le délai de survenue du décès, quelle qu'en soit la cause et 2°)

embedded studies is the change from baseline in patient-assessed global clinical status using a 100-point visual analog scale on the 7th inpatient day or discharge, whichever occurs earlier. (…)

Statistical Analysis

The EVEREST statistical design allows for the analysis of the acute and chronic effects of tolvaptan using 3 studies in a integrated, but independent manner. The primary outcome study is a single study designed to assess the effect of tolvaptan on clinical outcomes. Embedded within the primary outcome study are 2 distinct but identical studies, study A and study B. These studies are designed and powered to assess the short-term effect of tolvaptan on patient-assessed global clinical status. Studies A and B are designed to be analyzed separately for the global clinical status end points and to be pooled for the analysis of the outcome study. The intent-to-treat dataset including all randomized subjects will be used for the efficacy analysis of the main study and for the global clinical status outcome of the embedded studies. The patient population for the embedded studies will be obtained by randomizing centers to study A or study B at the conclusion of enrollment (Fig. 1B). (…)

All-cause mortality, total cardiovascular mortality, and total heart failure hospitalizations occurring from randomization until study termination will be included in the analysis. The study will be terminated after the 1065th death and a minimum of 60 days of therapy for all current subjects. The assessment of global clinical status will be evaluated separately in studies A and B. (…)

Efficacy monitoring will be based on the primary end point of all-cause mortality with the treatment comparison performed using the Peto-Peto-Wilcoxon log-rank test. A boundary based on the O'Brien-Fleming α spending function will be used to account for interim analyses of unblinded mortality data. After adjusting for interim analysis, the significance level for the all-cause mortality coprimary end point will be approximately 0.0376, for an overall type I error rate of 0.0402. The significance level for the cardiovascular mortality or heart failure hospitalization coprimary end point will be 0.009. (…)

le délai avant la première survenue soit d'un décès d'origine cardiovasculaire soit d'une hospitalisation pour insuffisance cardiaque. Le critère d'évaluation principal des deux essais intégrés est le changement par rapport à l'inclusion de l'état clinique global mesuré par le patient à l'aide d'une échelle visuelle analogique, graduée de 0 à 100, remplie au 7e jour d'hospitalisation ou lors de la sortie d'hôpital, selon l'événement survenant en premier. (…)

Analyses statistiques

Le plan d'analyse statistique d'Everest permet l'analyse des effets aigus et chroniques du tolvaptan en utilisant 3 essais d'une manière intégrée mais indépendante. L'essai principal est un essai unique conçu pour évaluer l'effet du tolvaptan sur le devenir clinique. Deux essais distincts mais identiques, essai A et essai B, sont intégrés à l'essai principal. Ces essais sont conçus avec une puissance suffisante pour évaluer l'effet à court terme du tolvaptan sur l'état clinique global évalué par le patient lui-même. Les essais A et B ont été conçus pour être analysés séparément pour l'état clinique global et pour être regroupés pour l'analyse de l'essai principal. Les données en intention de traiter provenant de tous les patients randomisés seront utilisées pour l'analyse de l'efficacité dans l'essai principal et pour le critère état clinique global des essais intégrés. La population de patients pour les essais intégrés sera obtenue en randomisant les centres entre essai A et essai B, à la fin du recrutement (fig. 1B). (…)

La mortalité quelle qu'en soit la cause, la mortalité cardiovasculaire totale et le nombre total d'hospitalisations pour insuffisance cardiaque survenues entre le tirage au sort et la fin de l'essai seront inclus dans l'analyse. L'essai sera terminé après la survenue du 1065e décès et un minimum de 60 jours de traitement pour tous les patients. L'évaluation de l'état clinique global sera effectuée séparément dans les essais A et B. (…)

Le suivi de l'efficacité sera fondé sur le critère principal de la mortalité de toutes causes en comparant les traitements par le test log rank de Peto-Peto Wilcoxon. Une limite fondée sur la fonction α de O'Brien-Fleming sera utilisée pour tenir compte de l'analyse intermédiaire des données non aveugles de mortalité. Après ajustement tenant compte des analyses intermédiaires, le niveau de significativité pour le critère mortalité de toutes causes sera approximativement de 0,0376, correspondant à un risque global d'erreur de type I de 0,0402. Le niveau de significativité sera 0,009 pour le critère mortalité cardiovasculaire ou hospitalisation pour insuffisance cardiaque.

(…)

EXERCICE N° 25 : **Bien prendre en compte les risques d'erreur**

Statistical Power and Sample Size

The sample size for the study was estimated based on the number of death events. An overall 20% reduction in all-cause mortality was originally assumed. In addition, a 15% dropout rate was assumed during the course of the study. These dropouts would have no effect on the mortality rate in placebo group, but were assumed to reduce the magnitude of the effect of tolvaptan in the tolvaptan treatment group because these dropouts would no longer be on active treatment. After all these considerations, 1065 deaths are required to provide 90% power to compare all-cause mortality at a 2-sided alpha of 0.009 (a level that has regulatory significance), assuming a hazard ratio of 0.7865 in all-cause mortality. (…)

For a less stringent 2-sided alpha of 0.0402 (…), 1065 deaths will provide 90% power to detect a hazard ratio of 0.8132 for all-cause mortality. Based on a projected annual mortality rate of 35% in the placebo group, this hazard ratio would translate to a 15.6% reduction in the annual mortality rate. An estimated 3600 patients will be enrolled to obtain 1065 deaths. The following assumptions were used to generate this estimate:

• The placebo rates of all-cause mortality are assumed to be 10%, 25% and 35% at 2 months, 6 months and 1 year, respectively. An annual rate of 30.5% is assumed beyond 1 year.

• A proportional hazards model is assumed such that tolvaptan reduces the 6-month mortality rate by 20% (hazard ratio = 0.776), which result in mortality rates in the tolvaptan group of 7.8%, 20%, and 28.4% at 2 months, 6 months and 1 year, respectively.

• It is also assumed that the discontinuation rate at 18 months will be 15% with treatment discontinuations uniformly distributed over the follow-up time.

(…) »

1- Gheorghiade M et coll. "Rationale and design of the multicenter, randomized, double-blind, placebo-controlled study to evaluate the efficacy of vasopressin antagonism in heart failure: outcome study with tolvaptan (EVEREST)" *J Card Fail* 2005 ; **11** (4) : 260-269.

Puissance statistique et taille de l'échantillon

La taille de l'échantillon pour cet essai a été estimée sur la base du nombre de décès. L'hypothèse initiale a été une diminution globale de 20 % de la mortalité toutes causes confondues. De plus, il a été fait l'hypothèse d'un taux d'abandon de 15 % durant la durée de l'essai. Ces abandons n'auraient pas d'effet sur le taux de mortalité dans le groupe placebo, mais ont été supposés réduire l'amplitude de l'effet du tolvaptan dans le groupe tovalptan, parce que ces patients ne seraient plus sous traitement actif. D'après toutes ces considérations, 1 065 décès sont requis pour avoir une puissance de 90 % pour comparer la mortalité de toutes causes avec un risque alpha bilatéral de 0,009 (un seuil considéré comme significatif par les agences du médicament), en supposant un risque relatif de 0,7865 pour la mortalité quelle qu'en soit la cause. (…)

Pour un risque alpha bilatéral moins strict de 0,0402 (…) 1 065 décès fourniront une puissance de 90 % de détecter un risque relatif de 0,8132 pour la mortalité quelle qu'en soit la cause. En se basant sur une mortalité annuelle prévisible de 35 % dans le groupe placebo, ce risque relatif se traduirait par une réduction de 15,6 % du taux de mortalité annuelle. Environ 3 600 patients seront inclus pour obtenir 1 065 décès. Les hypothèses suivantes ont été utilisées pour générer cette estimation.

• Le taux de mortalité de toutes causes sous placebo a été supposé être de 10 %, 25 % et 35 % à 2 mois, 6 mois et 1 an respectivement. Le taux annuel est supposé être de 30,5 % au-delà de la première année.

• Le risque est supposé suivre un modèle proportionnel, selon lequel le tolvaptan réduit de 20 % le taux de mortalité à 6 mois (risque relatif = 0.776) et conduit à un taux de mortalité dans le groupe tolvaptan de 7,8 %, 20 % et 28,4 % à 2 mois, 6 mois et 1 an respectivement.

• Il a aussi été supposé que le taux d'arrêt à 18 mois sera de 15 % avec des arrêts de traitement uniformément répartis durant tout le suivi.

(…) »

Traduction©Prescrire

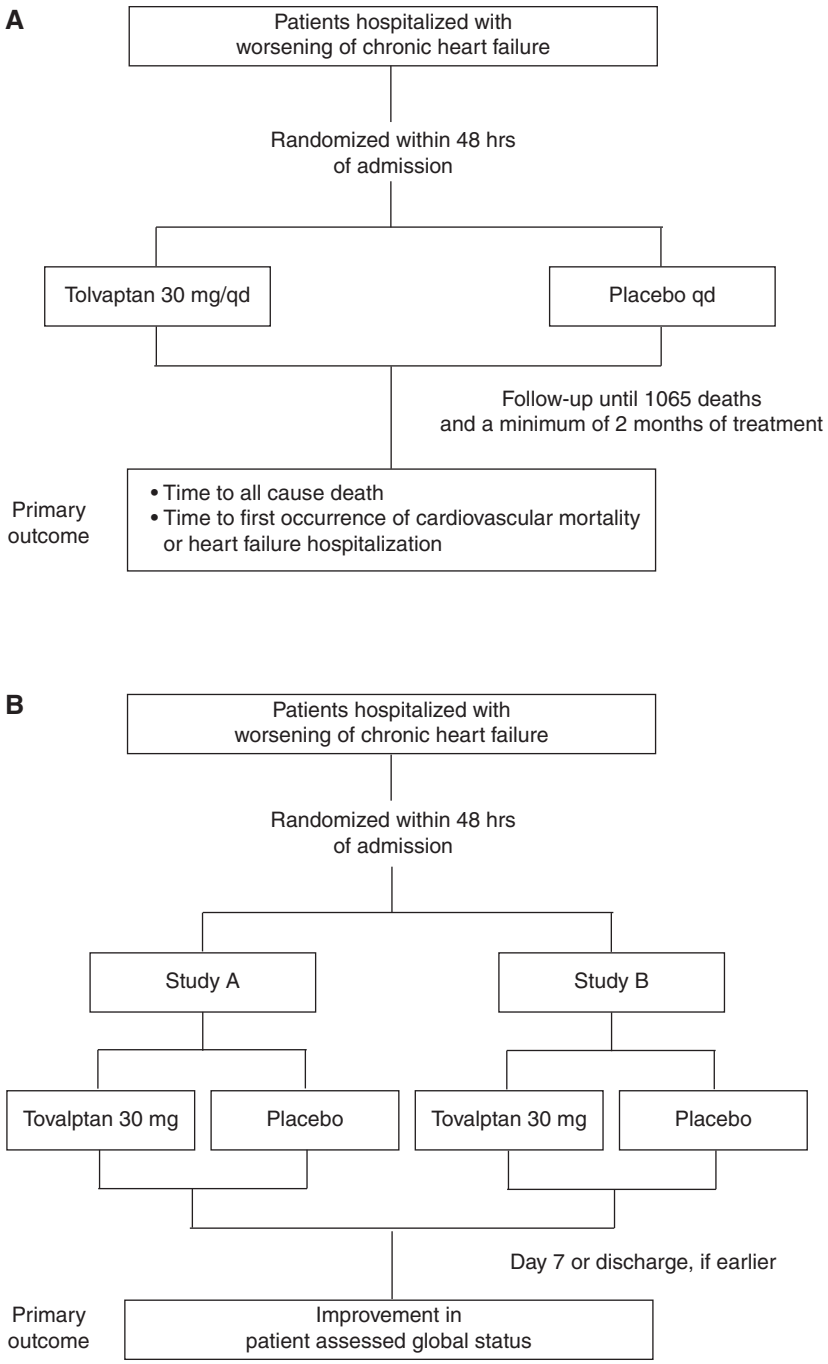

Fig.1 (A) Schematic of the main study design. (B) Schematic of embedded study design.

EXERCICE N° 25 : **Bien prendre en compte les risques d'erreur**

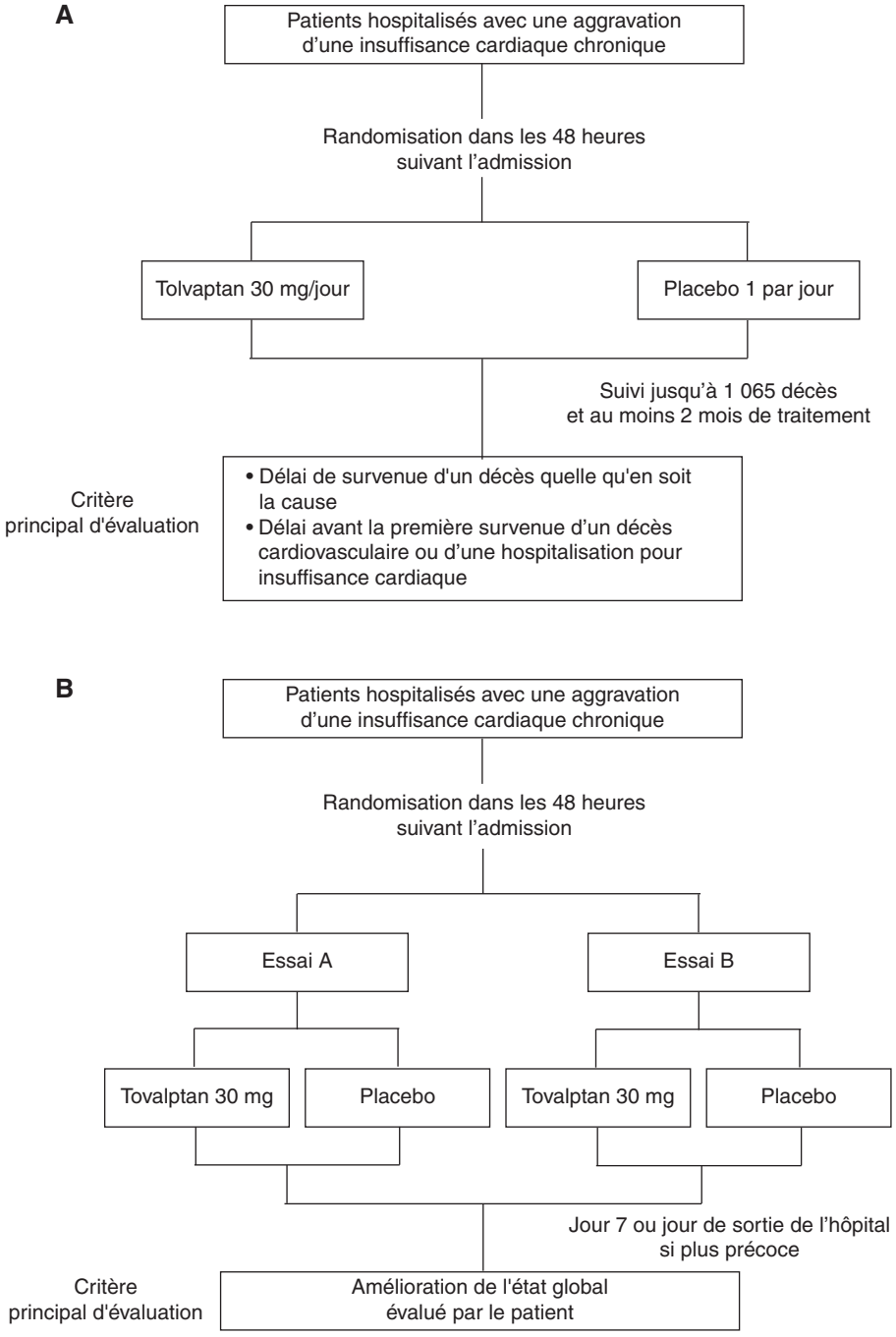

Fig.1 (A) Schéma du protocole de l'essai principal. (B) Schéma du protocole des essais intégrés.

EXERCICE N° 25 : **Bien prendre en compte les risques d'erreur**

Questions

Question n° 1

Quels avantages y a-t-il à étudier l'efficacité à court terme du *tolvaptan* sur le statut clinique global dans deux essais de protocoles identiques plutôt qu'un seul essai ?

Question n° 2

Pour l'essai principal, évaluant l'efficacité à long terme du *tolvaptan*, il y a deux critères principaux d'évaluation. Lequel vous paraît le plus pertinent ?

Question n° 3

Le *tolvaptan* étant censé corriger l'hyponatrémie des insuffisants cardiaques, n'aurait-il pas été préférable d'utiliser la natrémie comme critère d'évaluation principal dans au moins une des évaluations (à court terme, à long terme) ? Si oui, dans laquelle ? Si non, pourquoi ?

Question n° 4

Quel est le seuil de significativité statistique retenu pour le second critère d'évaluation principal ?

Question n° 5

À partir de quel critère d'évaluation a été déterminé le nombre de patients à inclure ? Pour ce critère, quelle est la puissance statistique de cet essai ?

Question n° 6

L'essai étant multicentrique, quelle précaution doit être observée lors de l'analyse statistique ?

Propositions de réponses de la Rédaction

Question n° 1

Des résultats concordants avec deux essais de taille suffisante sont plus convaincants que les résultats d'un seul essai. En effet, il y a moins de chance que le seul hasard explique une différence observée dans deux essais.

Question n° 2

Pour l'étude de l'efficacité à long terme du *tolvaptan* la mortalité toutes causes confondues est le critère d'évaluation le plus pertinent. En effet, l'insuffisance cardiaque chronique est associée à une mortalité élevée et une baisse de cette mortalité est habituellement l'attente principale des malades. Par ailleurs, la baisse de la mortalité toutes causes confondues est plus satisfaisante que le second critère, mortalité cardiovasculaire ou hospitalisation pour insuffisance cardiaque, car on ne peut exclure qu'une baisse de la mortalité cardiovasculaire due au médicament ne soit compensée par une augmentation de la mortalité d'autres causes en raison d'une toxicité particulière.

Question n° 3

La natrémie n'est qu'un critère biologique intermédiaire. Mieux vaut privilégier les critères cliniques réellement pertinents pour les malades. Les investigateurs ont eu raison de choisir un critère d'évaluation global de l'état de santé des patients, même pour les essais à court terme.

Question n° 4

Pour le second critère d'évaluation principal (mortalité cardiovasculaire ou hospitalisation pour insuffisance cardiaque), le seuil de significativité statistique retenu est de 0,009.

Question n° 5

Le nombre de patients à inclure a été déterminé de manière à avoir une probabilité élevée de mettre en évidence, si elle existe, une différence pour le critère "mortalité toutes causes confondues". D'après le paragraphe "puissance statistique", pour ce critère, avec 1 065 décès, il y a 90 % de chances (autrement dit une puissance statistique de 90 %) de détecter une réduction relative d'au moins 21,35 % (= 1-0,7865) de la mortalité toutes causes confondues et de conclure qu'elle n'est pas due au hasard, si l'on accepte un risque alpha de 0,9 % de se tromper en faisant cette conclusion. Si l'on choisit un risque alpha de 4,02 % de se tromper, il y a 90 % de chances de détecter une réduction relative d'au moins 18,7 % (= 1-0,8132) de la mortalité toutes causes confondues et de conclure qu'elle n'est pas due au hasard.

EXERCICE N° 25 : **Bien prendre en compte les risques d'erreur**

Question n° 6

Dans un essai multicentrique, il est nécessaire d'évaluer si les résultats sont homogènes d'un centre à l'autre et de vérifier qu'un seul centre aux résultats divergents des autres ne soit pas la cause des résultats finaux. C'est ce qu'on appelle l'effet centre.

Commentaires de la Rédaction

Commentaires de la Rédaction sur la question 1. La reproductibilité des données expérimentales est une exigence commune à toute démarche scientifique fondée sur l'observation. Elle a pour but de diminuer la probabilité qu'un résultat soit observé par hasard ou à cause d'un biais méthodologique (lire aussi dans ce numéro, page 4).
La statistique ne permet jamais d'exclure totalement qu'une différence observée entre deux traitements soit en fait due au hasard, même lorsqu'elle est "statistiquement significative". Mais une même différence constatée dans deux essais ayant chacun inclus 100 patients a moins de chances d'être liée au hasard que la même différence constatée dans un seul essai ayant inclus 200 patients.
Si les deux essais concordants sont réalisés par des équipes différentes, ce qui n'est pas le cas ici, leur niveau de preuves est encore plus important. Lorsque, comme ici, les essais ont le même protocole et sont réalisés dans des populations similaires, il est possible qu'un même biais expliquant les résultats soit présent dans les deux essais.
Disposer d'au moins 2 essais aux résultats concordants a longtemps été une exigence des agences du médicament, changées notamment de conseiller l'octroi ou non des autorisations de mise sur le marché (AMM). Mais, dans certains domaines, notamment en cancérologie, les nouveaux médicaments sont de plus en plus souvent autorisés sur la base d'un seul essai comparatif, parfois terminé prématurément à la suite d'une analyse intermédiaire. L'excuse invoquée est de ne pas retarder l'accès des patients à des médicaments "innovants". Nous regrettons cet empressement, car les exemples ne manquent pas d'un premier essai favorable démenti par des essais ultérieurs.

Commentaires de la Rédaction sur les questions 2 et 3. Avant de lire le compte rendu dans un essai il faut se poser la question de ce qu'on en attend pour sa pratique. Quelle est l'attente des malades vis-à-vis de ce nouveau médicament ? C'est en fonction de cette attente que les critères d'évaluation doivent être choisis. Ici, mesurer la natrémie comme critère d'évaluation principal reviendrait juste à vérifier un effet pharmacologique.

Commentaires de la Rédaction sur la question 4. Toute comparaison statistique comporte deux risques d'erreur. L'un est de conclure à une différence alors, qu'en réalité, il n'y en a pas (erreur de type I, alias risque de première espèce alpha). Avant de réaliser un essai, il convient de fixer quel risque alpha on acceptera pour chaque comparaison effectuée. C'est le "seuil de significativité statistique". Habituellement, le risque alpha choisi est de 5 % ou 0,05. Si la valeur p est en dessous de ce seuil, on considère qu'une différence est "statistiquement significative". Autrement dit, si les deux traitements avaient exactement le même effet, il y aurait moins de 5 chances sur 100 de constater par hasard une telle différence.
Dans la plupart des essais, les investigateurs font plusieurs comparaisons. Même si les deux traitements ont exactement le même effet, la multiplication des comparaisons augmente le risque d'observer une différence entre les deux traitements, qui ne serait due qu'au hasard. Pour tenir compte de la multiplication des comparaisons, mieux vaut alors choisir un seuil de significativité statistique plus bas : 0,01, voire moins.
Dans cet essai, pour que le risque alpha de conclure à tort à l'existence d'une différence cliniquement pertinente entre le *tolvaptan* et le placebo n'excède pas 5 % (0,05) pour l'ensemble de l'essai, les auteurs ont choisi un risque alpha = 0,0402 pour le premier critère principal de jugement, un risque alpha = 0,009 pour le second critère principal de jugement et un risque

alpha = 0,0008 pour les deux essais intégrés évaluant le statut clinique global à court terme (ce dernier chiffre n'est pas précisé dans cet article, mais dans la publication finale des résultats). Ainsi 0,0402 + 0,009 + 0,0008 = 0,05. Par ailleurs, il a été prévu que la mortalité dans les deux groupes soit comparée plusieurs fois en cours d'essai. Pour que ces comparaisons intermédiaires n'augmentent pas le risque alpha au-delà de 0,0402, les auteurs ont calculé qu'il faudrait retenir 0,0376 comme seuil de significativité statistique lors de l'analyse finale du premier critère principal de jugement. Une telle rigueur est assez rare.

Commentaires de la Rédaction sur la question 5. Le deuxième risque d'erreur lors d'une comparaison statistique est de conclure à l'absence de différence alors qu'en réalité il y en a une. C'est le risque de deuxième espèce bêta, alias erreur de type II. Ce risque est lié à un manque de puissance de l'essai (la puissance statistique est la probabilité 1-bêta). Ici, les investigateurs ont calculé a priori la puissance de l'essai. Cette puissance est telle qu'ils avaient 90 % de chance de mettre en évidence une différence relative d'au moins 18,7 % entre le *tolvaptan* et le placebo pour le critère "mortalité de toute cause".
À partir de ces éléments définis a priori : risque alpha, puissance statistique, différence minimale considérée comme cliniquement pertinente (et, dans le cas d'une comparaison de deux moyennes, variabilité du paramètre exprimé par sa variance), des abaques ou des logiciels permettent de déterminer le nombre de patients à inclure.

Commentaires de la Rédaction sur la question 6. L'analyse de l'homogénéité de l'efficacité observée d'un centre à l'autre est souvent une donnée absente des comptes rendus d'essai clinique. Quand un centre a obtenu des résultats divergents, il importe d'en rechercher la cause : cela peut résulter d'un biais propre à ce centre. Pour vérifier la robustesse des conclusions, mieux vaut vérifier que l'exclusion du centre ou des centres dont les résultats paraissent discordants ne modifient pas les résultats.

Commentaires de la rédaction sur l'ensemble de l'exercice. Il est malheureusement rare qu'un essai clinique financé par une firme soit aussi bien conçu. Notamment qu'il utilise la mortalité toutes causes confondues comme critère principal d'évaluation, qu'il ait une puissance statistique suffisante pour ce critère, et qu'il prenne aussi rigoureusement en compte la multiplication des comparaisons pour déterminer le seuil de significativité statistique.

Pour aller plus loin

– Prescrire Rédaction "La valeur "p" résume-t-elle la pertinence d'une comparaison ?" *Rev Prescrire* 2008 ; **28** (298) : 621-622.
– Prescrire Rédaction "Puissance d'une étude comparative. À prendre en compte pour interpréter certains résultats" *Rev Prescrire* 2008 ; **28** (298) : 634-636.

Cet exercice aborde certains objectifs pédagogiques proposés en France pour l'épreuve de lecture critique d'un article médical : les objectifs n° 7, 9, 13, 16.

■