CHAPITRE

16

Les études de la fiabilité diagnostique

« Rien n'est constant dans ce monde que l'inconstance. »
Jonathan Swift

L'ESSENTIEL À RETENIR

- Une étude de fiabilité est une expérimentation dans laquelle l'ordre d'application ou d'interprétation de l'outil évalué doit être tiré au sort.
- Toutes les sources de variation considérées doivent être pertinentes.
- La population étudiée doit être aussi hétérogène que possible, mais représentative des conditions d'utilisation de l'outil évalué.
- Les modalités pratiques d'application et d'interprétation des résultats, si possible en insu, doivent garantir l'indépendance des observations.
- Le délai entre deux applications du test doit être documenté et suffisamment court pour que l'état des sujets n'ait pas changé.
- L'analyse d'une étude de fiabilité repose sur l'estimation d'une mesure de concordance, un coefficient kappa pour les mesures qualitatives ou un coefficient de corrélation intraclasse pour les mesures quantitatives.
- Ni les mesures d'association ou de corrélation simple ni les tests statistiques ne donnent d'indications sur la fiabilité.
- La signification clinique des résultats se juge sur la largeur des intervalles de confiance autour des paramètres, sur une mesure de précision relative et sur la nature des discordances et de leurs conséquences.

Guide pratique de méthodes de recherche en santé
© 2026 Elsevier Masson SAS. Tous les droits sont réservés, y compris ceux relatifs à l'exploration de textes et de données, à la formation en intelligence artificielle et aux technologies similaires.

Schéma d'étude expérimental

Sources de variation pertinentes

L'évaluation de la fiabilité d'un outil diagnostique doit se faire dans des **conditions expérimentales** (figure 16.1). Le schéma d'étude idéal consiste à **comparer les résultats de l'outil diagnostique soumis expérimentalement à toutes les sources possibles de variation**. Quand ces dernières sont nombreuses, le schéma d'étude peut se limiter à observer l'effet d'une ou de deux sources de variation sur les résultats. Les conclusions de l'étude, cependant, seront limitées à la fiabilité vis-à-vis de ces sources.

Les sources de variation le plus souvent évaluées sont la **répétition de la mesure chez le même sujet** (fiabilité « test-retest » ou **reproductibilité**), la **variation entre observateurs** et l'**accord entre différents outils imparfaits**. L'évaluation de l'accord entre un outil imparfait et un outil parfait ne relève pas de la fiabilité mais de l'exactitude de l'outil imparfait. La revue des études préalables doit préciser quelles sources de variation ont déjà été étudiées et pourquoi les sources de variation considérées dans l'étude sont pertinentes.

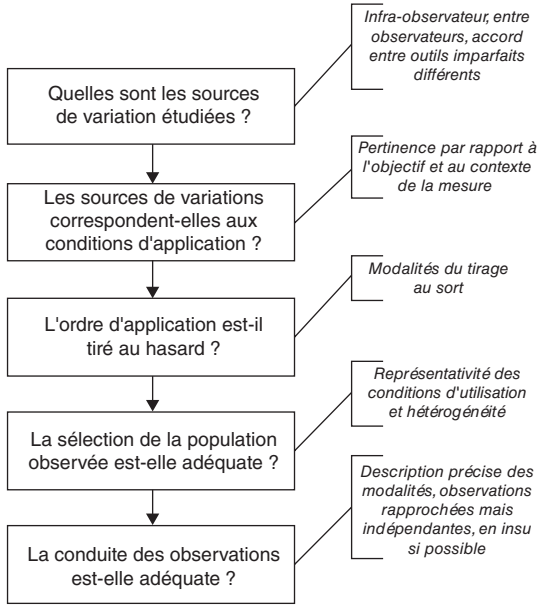


Figure 16.1. Questions à se poser sur la conduite d'une étude de fiabilité.
Source : Salmi LR. Lecture critique et communication médicale scientifique. 3ème édition. Elsevier Masson ; 2012.

Le PICOT d'une étude de fiabilité

Population (P), *Index test* (I, test évalué), *Variability sources* (C ou V, les sources de variabilité), *Outcome* (O, la maladie ou le phénomène à identifier), *Time* (T).

Toutes les sources considérées dans l'étude doivent être pertinentes d'un point de vue clinique ou épidémiologique. Cette pertinence doit être jugée **par rapport aux conditions d'application** qui sont ou seront effectivement celles **dans lesquelles le test ou l'outil de mesure sera utilisé**.

Remarque

Pour que la diffusion d'un nouvel outil diagnostique soit envisagée, il faut que sa fiabilité ait été évaluée vis-à-vis de toutes les sources pertinentes de variation. Par exemple, la comparaison de l'interprétation des résultats d'une radiographie par deux radiologues est pertinente si l'objectif est le diagnostic différentiel en milieu spécialisé, mais manque de pertinence si l'objectif est de dépister une maladie en soins de première ligne. Un test sérologique dont la fiabilité a été démontrée en milieu hospitalier universitaire n'est pas forcément prêt pour une diffusion dans tous les laboratoires d'analyses médicales, où les sources de variation peuvent être différentes. Toutes les sources pertinentes doivent donc avoir été considérées, éventuellement dans plusieurs études, pour que la diffusion de l'outil diagnostique soit justifiée.

Tirage au sort de l'ordre d'application

L'ordre d'application de l'outil ou d'interprétation de ses résultats doit être tiré au hasard. Par exemple, lors de l'évaluation de l'effet des modalités d'administration d'un questionnaire, la moitié des individus peut être tirée au hasard pour remplir le questionnaire et être interrogée ensuite par un médecin qui remplit le même questionnaire ; les autres individus seraient d'abord interrogés, puis rempliraient eux-mêmes le questionnaire. **Le tirage au sort peut porter sur plusieurs sources**, dans des plans statistiques à plusieurs facteurs. Par exemple, une évaluation de la fiabilité du contrôle ultime du groupe sanguin au lit du malade comportait une répartition au hasard de l'ordre dans lequel une série de douze cartes de contrôle était interprétée par des infirmières qui différaient les unes des autres selon l'ancienneté et le type de service dans lequel elles exerçaient (Dujardin PP, et al. Vox Sang. 2000;78:37-43). Ce tirage au sort, **associé à des procédures d'insu** (*cf.* plus loin), est la meilleure garantie de neutraliser les biais d'information.

Échantillon représentatif mais hétérogène

La sélection des individus pour une étude de fiabilité est une étape difficile car **les deux éléments à respecter peuvent paraître contradictoires**. La population étudiée doit en effet être représentative des conditions d'utilisation mais doit rester hétérogène. La

définition de la population et les modalités pratiques de sa sélection doivent apparaître dans la section « Méthodes » du protocole et de l'article et la description des individus effectivement étudiés devra apparaître au début de la section « Résultats » de l'article. L'absence de ces descriptions peut remettre les résultats de l'étude en question ; l'inclusion d'une population trop homogène fera interpréter les résultats avec prudence.

Représentativité des conditions d'utilisation

La **nécessité d'une représentativité des conditions d'utilisation** découle de l'exigence de pertinence. Par exemple, une étude de la fiabilité de l'interprétation de mammographies doit être menée sur une série de mammographies issues de cabinets de radiologie, représentatifs de tous les cabinets de radiologie dans lesquels les mammographies sont pratiquées. Une **source de biais de sélection** serait de limiter la série de clichés à ceux issus d'une clinique universitaire sous prétexte que les conditions d'application sont mieux maîtrisées par les investigateurs.

Sélection d'une population hétérogène

La **sélection d'une population hétérogène** permet une analyse pertinente des données de l'étude. Cette analyse consiste à décrire et à estimer la concordance des résultats de l'outil diagnostique appliqué dans différentes conditions. On peut concevoir, sans maîtriser les aspects statistiques des analyses de concordance discutés plus loin dans ce chapitre, qu'une population homogène donnera des résultats plus facilement concordants qu'une population hétérogène.

Exemple

Les possibilités de discordance dans l'interprétation d'un électrocardiogramme (ECG) sont *a priori* moins fréquentes si les auteurs ont inclus seulement des individus jeunes, sportifs et sains et des patients âgés ayant un infarctus majeur établi (population homogène), que s'ils ont pris soin d'inclure tous les états et âges intermédiaires (population hétérogène).

Quand l'outil est supposé mesurer une variable à deux modalités, **la sélection d'une population hétérogène consiste à forcer la distribution de cette variable à être équilibrée**. Par exemple, si on évalue la fiabilité d'un diagnostic exprimé par « maladie présente ou absente », on sélectionnera un échantillon dans lequel il y aura autant de malades que de non-malades.

La sélection d'une population représentative implique des **techniques d'échantillonnage aléatoire**. La nécessité d'une population hétérogène implique alors d'utiliser des techniques d'échantillonnage complexes telles que l'**échantillonnage stratifié**, qui permettent de représenter certains groupes de la population (par exemple, les malades) plus que d'autres, tout en gardant une possibilité d'extrapolation à l'ensemble de la population. Ces techniques sont évoquées au chapitre 20.

Conduite de l'étude

Modalités pratiques d'application du test

La **description des modalités pratiques** est indispensable pour savoir dans quelles conditions d'application le niveau de fiabilité rapporté par l'étude est atteint. Cette description permettra aussi au lecteur d'utiliser l'outil dans sa propre pratique ou dans une autre étude. Une étude qui ne rapporte pas les modalités pratiques d'application du test est inutile et peut être rejetée.

Observations indépendantes et, si possible, en insu

La description des modalités d'application doit aussi **indiquer que les mesures successives sont indépendantes les unes des autres**, pour garantir qu'aucune observation ou interprétation n'a été influencée par une observation ou interprétation précédente. Le meilleur moyen de garantir l'indépendance des observations est d'utiliser des procédures d'insu. L'utilisation de **procédures d'insu consiste à cacher les résultats obtenus par un observateur à tous les autres observateurs**. Ainsi, l'application de l'outil diagnostique se fait de manière indépendante par chaque observateur et l'interprétation des résultats ne peut pas être influencée par les résultats des autres applications. Quand l'outil est supposé mesurer un phénomène suggestif ou que l'interprétation est difficile et quand l'objet de l'étude est la variabilité entre observateurs, l'absence de procédures d'insu est une raison souvent suffisante pour rejeter l'étude.
Cependant, **les procédures d'insu ne sont pas toujours faisables**. Ainsi, dans une étude qui évaluerait la concordance entre deux outils supposés mesurer le même phénomène, par exemple un questionnaire administré par un professionnel de santé ou rempli par l'individu lui-même, l'individu connaît les résultats des deux questionnaires. Dans ces circonstances, le **tirage au sort de l'ordre d'administration** des outils est particulièrement important car il permet d'évaluer l'effet de l'absence d'insu. Les auteurs doivent alors évoquer l'absence d'insu dans la discussion et indiquer le degré plausible de surestimation de la concordance. **Le délai entre deux applications du test doit aussi être documenté** dans les résultats, pour garantir que la variabilité éventuelle des résultats n'est pas liée à un changement dans l'état des sujets observés.

Analyser les données de fiabilité

Une analyse statistique adéquate de la fiabilité repose sur l'**estimation de mesures d'accord ou de concordance**. De nombreuses mesures d'accord sont disponibles, mais seules **sont acceptables celles qui tiennent compte du fait qu'un accord entre deux mesures peut être en partie le fait de coïncidences**.

Remarque

Une erreur fréquente est de rapporter des mesures d'association ou les résultats d'**un test d'inférence**, par exemple un χ^2 pour l'accord entre deux mesures dichotomiques ou **un coefficient de corrélation** pour des mesures quantitatives. L'utilisation de ces mesures est fausse car, **si l'accord ne peut s'observer sans association, une association entre deux mesures, même forte et statistiquement significative, n'est pas synonyme d'accord**. Par exemple, il existe une corrélation parfaite, mais jamais de concordance entre deux thermomètres ayant systématiquement cinq degrés Celsius d'écart. Pensez aux graves conséquences, pour le traitement des patients atteints d'hyperthermie, qu'aurait le fait de dire que les deux thermomètres sont équivalents.

Tableau croisé et estimation du coefficient *kappa*

Parmi les mesures corrigées pour l'effet du hasard, la plus fréquemment utilisée est le coefficient kappa (κ) pour la concordance entre deux ou plusieurs mesures qualitatives (tableau 16.1). L'analyse des études de fiabilité considérant de nombreuses sources de variation peut aussi être fondée sur des modèles statistiques complexes (modèles log-linéaires ou analyses de variance).

Concordance des mesures quantitatives

Si les résultats de l'application de l'outil diagnostique évalué ne s'expriment pas sous forme d'une positivité et d'une négativité mais sous forme de résultats quantitatifs, **le κ peut se calculer après avoir fixé un seuil de positivité ou avoir défini des classes de valeurs croissantes**. Cette analyse fondée sur une transformation d'une variable quantitative en variable ordinale, bien que correcte, entraîne cependant une perte d'information. Dans cette circonstance, les chercheurs peuvent estimer le **coefficient de corrélation intraclasse pour la concordance de mesures quantitatives**. Cette analyse est habituellement complétée par une représentation graphique, incluant un **nuage de points** et le **diagramme de Bland et Altman**, qui permet de repérer les observations particulièrement discordantes (figure 16.2).

Tableau 16.1. Analyse d'une étude de fiabilité diagnostique.

Observateur 2	Observateur 1		Total
	Positif	Négatif	
Positif	a	b	$a + b$
Négatif	c	d	$c + d$
Total	$a + c$	$b + d$	n

Proportion d'accord observée : Po = $(a + d)/n$
Proportion d'accord attendue :
Pa = $\{[(a + b)/n] \times [(a + c)/n] + [(c + d)/n] \times [(b + d)/n]\}$/n
Kappa (κ) = (Po − Pa)/(1 − Pa)

Cas de données qualitatives à deux modalités (positif/négatif).

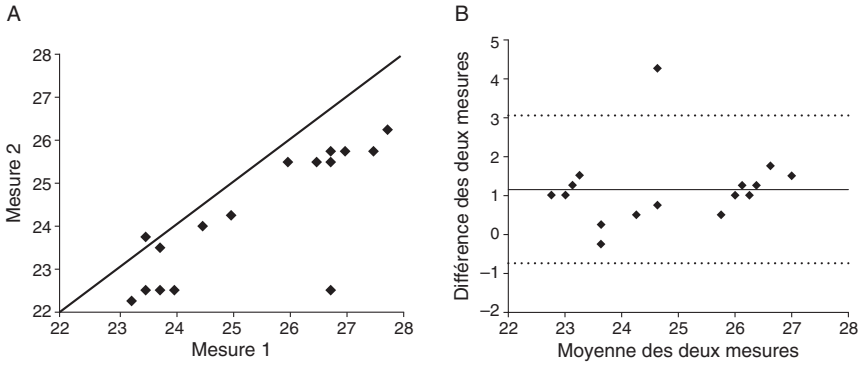


Figure 16.2. Exemple de diagrammes pour l'analyse de la discordance entre deux mesures quantitatives.

A. Nuage de points descriptif : la diagonale indique la ligne d'accord parfait. B. Diagramme de Bland et Altman : la ligne horizontale pleine indique la moyenne des écarts et les deux lignes horizontales pointillées délimitent la zone des écarts non statistiquement significatifs.

Source : Salmi LR. Lecture critique et communication médicale scientifique. 3ème édition. Elsevier Masson ; 2012.

Interprétation des résultats

Signification clinique de la concordance

Une difficulté habituelle dans l'interprétation des résultats est d'**appliquer la notion de signification clinique**. En pratique, il faut juger cette signification clinique sur trois éléments :

- la **précision des estimations** ;
- une estimation de la **précision relative de l'outil** de mesure ;
- la **nature des erreurs** (résultats discordants) **et de leurs conséquences**.

La précision des estimations est indiquée par la **largeur de l'intervalle de confiance autour des paramètres de concordance**, qui doit être fournie y compris pour des valeurs parfaites. Dans les études de fiabilité, c'est pour estimer le κ avec une précision suffisante que l'on calcule le nombre de sujets nécessaire.

Précision relative

Selon l'objectif de l'outil de mesure, l'expression des résultats par une mesure d'accord corrigée pour l'effet du hasard et son intervalle de confiance peut être insuffisamment informative. En effet, **une façon habile d'obtenir un bon accord est d'utiliser des outils imprécis**. L'analyse doit donc être complétée par l'estimation de la précision relative, définie comme le **rapport de la variation liée aux fluctuations de la mesure (le défaut de fiabilité) à la variation du phénomène que l'on veut être capable de détecter**. Si les auteurs ne fournissent pas d'informations sur la précision relative, empêchant ainsi de juger la capacité de l'outil à détecter les différences cliniquement importantes, l'étude doit être interprétée avec prudence.

Exemple

Deux thermomètres peuvent avoir une bonne concordance au degré, même s'ils ne sont pas toujours concordants au 1/10e de degré. Si l'objectif de ces thermomètres est de mesurer la température de malades, la parfaite concordance (graduation au degré) est inutile et la concordance imparfaite (graduation au 1/10e de degré) n'est pas gênante s'il est suffisant de mesurer la température au demi-degré près.

Description des mesures discordantes

L'analyse d'une étude de fiabilité doit être complétée par une **description des résultats discordants**. Que ces résultats soient fréquents ou non, l'analyse des sources possibles de désaccord est importante pour envisager des améliorations de l'outil. Cette analyse doit être fondée sur une description des observations discordantes et, éventuellement, sur des approches plus élaborées, par exemple une technique de consensus entre experts ou l'application d'un test de référence. Dans ce dernier cas, l'étude combine une évaluation de la fiabilité et de l'exactitude.

Les résultats de cette analyse doivent aussi permettre de **juger la gravité des conséquences des discordances**, en pondérant différemment les écarts proches (par exemple, un écart d'une unité) et les écarts importants (un diagnostic inversé, par exemple). Une étude de fiabilité qui ne rapporte pas la nature et les sources de désaccord peut être utile pour promouvoir un nouvel outil diagnostique, mais ne permettra probablement pas de faire progresser cet outil.

Le lecteur ne doit être convaincu d'utiliser un nouvel outil diagnostique, aussi fiable soit-il, que s'il apporte quelque chose par rapport aux outils existants. Le rapport d'une étude de fiabilité doit donc décrire, dans l'introduction ou la discussion, **quelles améliorations le nouvel outil apporte**, en termes de performance, de coûts, de risque pour le patient, de confort, de rapidité de résultats…

Formulaire de lecture critique d'une étude de la fiabilité diagnostique

La fiche 16.1 (en fin de chapitre) fournit un formulaire de lecture critique spécifique aux études de la fiabilité diagnostique. Ce formulaire comprend, entre autres, tous les critères attendus dans les normes de publication GRRAS (*cf.* références dans le tableau 4.1) pour les articles rapportant des études de la fiabilité diagnostique.

Exercices d'application

- Parmi les études sélectionnées dans l'exercice du chapitre 14, identifier les études dont l'objectif était l'évaluation de la fiabilité ; pour chaque étude, décrire les sources de variation envisagées et juger l'adéquation du schéma d'étude, de la population incluse, des procédures de conduite de l'étude et des analyses et de l'interprétation des résultats.
- Appliquer le formulaire de lecture critique (fiche 16.1) à un article rapportant une étude de la fiabilité diagnostique.

Fiche 16.1
Formulaire de lecture critique d'une étude de la fiabilité diagnostique

Référence

Auteurs : *Indiquer le nom des auteurs.*
Titre : *Indiquer le titre de l'article.*
Journal : *Indiquer le nom ou l'abréviation du journal.*
Année : *Indiquer l'année de publication.*
Tome : *Indiquer le tome (volume) de publication.*
Numéro : *Indiquer le numéro (issue) de publication.*
Première page : *Indiquer la première page de l'article.*
Dernière page : *Indiquer la dernière page de l'article.*

Résumé

Objectif ; maladie/phénomène à identifier ; définition et modalités pratiques des tests évalués

Formuler l'objectif de l'article en indiquant le(s) test(s) évalué(s), la maladie ou le phénomène à identifier, la population et le contexte.
☐ Répétition du test (test-retest)
☐ Accord entre observateurs
☐ Accord entre tests
☐ Autre : *Préciser.*

Tests évalués

Indiquer la définition et les modalités pratiques des tests évalués.

Schéma d'étude

Indiquer le schéma d'étude en précisant les sources de variabilité étudiées et les modalités du tirage au sort.

Population étudiée et nombre de sujets/nombre d'observateurs

Indiquer les critères d'inclusion et d'exclusion.
Indiquer le nombre de sujets effectivement étudiés dans chaque groupe.
Indiquer le nombre d'observateurs ou de répétitions d'observation.

Conduite de l'étude ; procédures d'insu

Indiquer les modalités pratiques de conduite de l'étude.

Résultats principaux

Indiquer les résultats principaux : coefficients Kappa ou de corrélation intraclasse et leurs intervalles de confiance ; autres paramètres pertinents (précision relative, graphiques de Bland et Altman…) ; description et explication des discordances.

Discussion, forces et faiblesses

Indiquer les principales forces et faiblesses et les éléments à retenir dans la conclusion ; détailler sur la page 2.

Conclusion

(*Très bonne/Plutôt bonne/Faible mais éléments acceptables/Inacceptable/Pas qualifié pour juger*).
Choisir
Lecteur : *Indiquer votre nom et votre prénom.*
Date : __/__/____

Ce formulaire n'est pas couvert par un copyright : il peut être copié ou imprimé sans permission mais avec mention de la source – Salmi LR. Guide pratique de méthodes de recherche en santé. De la rédaction du protocole à la publication. 5e édition. Issy-les-Moulineaux (France) : Elsevier Masson ; 2026 ; la version modifiable est disponible auprès de l'auteur.

Formulaire de lecture critique d'une étude de la fiabilité diagnostique (suite)

Sélectionner niveau de respect du critère

(O = oui, I = incomplet ; N = non ; NA = ne s'applique pas ; NSP = ne sait pas)

Critère	Niveau	Commentaire
Schéma d'étude		
Formulation claire du type de question	*Choisir*	*Indiquer les éléments pas clairs.*
Hypothèse faite *a priori*	*Choisir*	*Indiquer si a posteriori ou pas claire.*
Tirage au sort de l'ordre d'application	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Tests évalués clairement définis	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Pertinence de toutes les sources de variation considérées	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Pas d'oubli de sources pertinentes importantes	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Procédures de sélection		
Critères d'inclusion (sujets) décrits et acceptables	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Critères d'exclusion (sujets) décrits et acceptables	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Critères d'inclusion (observateurs) décrits et acceptables	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Critères d'exclusion (observateurs) décrits et acceptables	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Modalités d'échantillonnage décrites	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Représentativité des conditions d'utilisation	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Population hétérogène	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Indication du nombre de refus avant l'étude	*Choisir*	*Indiquer les effectifs aux étapes clés.*
Procédures identiques dans tous les groupes	*Choisir*	*Indiquer les différences.*
Taille d'échantillon calculée et adaptée	*Choisir*	*Indiquer les nombres ou raisons de non-calcul.*
Règles éthiques respectées	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Conduite de l'étude		
Description des modalités d'application des tests évalués	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Indépendance des différentes applications	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Procédures d'insu (application et interprétation)	*Choisir*	*Indiquer les éléments pas clairs ou inadéquats.*
Retraits indiqués, expliqués et raisonnables	*Choisir*	*Indiquer les effectifs aux étapes clés.*
Retraits équilibrés entre les groupes	*Choisir*	*Indiquer les différences inadéquates.*
Analyse statistique et résultats		
Effectifs de la population incluse et analysée (diagramme)	*Choisir*	*Indiquer les éléments pas clairs et inadéquats.*
Description de la population incluse et analysée	*Choisir*	*Indiquer les changements ou éléments pas clairs.*
Quantification des délais et actions entre les deux tests	*Choisir*	*Indiquer les différences et éléments pas clairs ou inadéquats.*
Accord corrigé pour le hasard et intervalle de confiance	*Choisir*	*Indiquer les éléments pas clairs.*
Précision relative et son intervalle de confiance	*Choisir*	*Indiquer les éléments pas clairs.*
Description des discordances	*Choisir*	*Indiquer les éléments pas clairs.*
Analyse des causes de discordance	*Choisir*	*Indiquer les éléments pas clairs.*
Description des effets indésirables	*Choisir*	*Indiquer les éléments pas clairs.*
Prise en compte des variables importantes	*Choisir*	*Indiquer les techniques d'ajustement ou de design.*
Résultats vérifiables (tableau croisé ou diagrammes)	*Choisir*	*Indiquer les éléments pas clairs.*
Prise en compte des comparaisons multiples	*Choisir*	*Indiquer les techniques d'ajustement.*
Autres éléments		
Tests et résultats applicables hors du contexte de l'étude	*Choisir*	*Juger selon la question et du contexte.*
Information fournie sur la validité des tests évalués	*Choisir*	*Indiquer les résultats ou références.*
Apport des tests évalués par rapport aux tests disponibles	*Choisir*	*Juger selon le contexte.*

Ce formulaire n'est pas couvert par un copyright : il peut être copié ou imprimé sans permission mais avec mention de la source – Salmi LR. Guide pratique de méthodes de recherche en santé. De la rédaction du protocole à la publication. 5e édition. Issy-les-Moulineaux (France) : Elsevier Masson ; 2026 ; la version modifiable est disponible auprès de l'auteur.