Université de Montréal

L'intelligence artificielle pour analyser des protocoles avec alternance de traitements

Par

Emily Heng

École de psychoéducation

Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de maîtrise en sciences (M. Sc.)

en psychoéducation, option mémoire et stage

Août 2020

Résumé

Les protocoles avec alternance de traitements sont des protocoles expérimentaux à cas uniques utiles pour évaluer et pour comparer l'efficacité d'interventions. Pour l'analyse de ces protocoles, les meilleures pratiques suggèrent aux chercheurs et aux professionnels d'utiliser conjointement les analyses statistiques et visuelles, mais ces méthodes produisent des taux d'erreurs insatisfaisants sous certaines conditions. Dans le but de considérer cet enjeu, notre étude a examiné l'utilisation de réseaux de neurones artificiels pour analyser les protocoles avec alternance de traitements et a comparé leurs performances à trois autres approches récentes. Plus précisément, nous avons examiné leur précision, leur puissance statistique et leurs erreurs de type I sous différentes conditions. Bien qu'il ne soit pas parfait, le modèle de réseaux de neurones artificiels présentait en général de meilleurs résultats et une plus grande stabilité à travers les analyses. Nos résultats suggèrent que les réseaux de neurones artificiels puissent être des solutions prometteuses pour analyser des protocoles avec alternance de traitements.

Mots-clés : apprentissage automatique, erreurs de type I, protocole à cas uniques, protocole avec alternance de traitements, réseaux de neurones artificiels

Abstract

Alternating-treatment designs are useful single-case experimental designs for the evaluation and comparison of intervention effectiveness. Most guidelines suggest that researchers and practitioners use a combination of statistical and visual analyses to analyze these designs, but current methods still produce inadequate levels of errors under certain conditions. In an attempt to address this issue, our study examined the use of artificial neural networks to analyze alternating-treatment designs and compared their performances to three other recent approaches. Specifically, we examined accuracy, statistical power, and type I error rates under various conditions. Albeit not perfect, the artificial neural networks model generally provided better and more stable results across analyses. Our results suggest that artificial neural networks are promising alternatives to analyze alternating-treatment designs.

Keywords: alternating-treatment design, artificial neural networks, machine learning, N-of-1 trials, type I error

Table des matières

Liste des tableaux

Listes des figures

Liste des sigles et abréviations

AA : apprentissage automatique

ALIV : Actual and linearly interpolated values

ANN : Artificial neural networks

PAT : protocole avec alternance de traitements

RB : randomisation par blocs

RR : randomisation restreinte

RD : Ratio of distances

SMD : standardized mean difference

VSC : Visual structured criterion

Remerciements

Tout d'abord, je dois un énorme remerciement à mon directeur de recherche, Marc Lanovaz. Ce mémoire n'aurait jamais été possible si ce n'était pas de ton expertise, de tes conseils et de ton aide. À travers tous les imprévus rencontrés, je te remercie pour ton soutien du début jusqu'à la fin. Je suis très reconnaissante de toutes les occasions que tu m'as offertes, incluant la participation à de multiples projets de recherche, séminaires et conférences enrichissants. N'importe quel étudiant serait chanceux de t'avoir comme directeur !

Pour suivre, je souhaite remercier mes collègues de laboratoire avec qui j'ai partagé de nombreux fous rires et de nombreuses confidences. Vos conseils et votre soutien moral durant les périodes incertaines m'ont permis d'avancer et de me sentir moins seule. Les journées au labo n'auraient pas été pareilles sans votre présence. Une dédicace spéciale à Lydia, avec qui j'ai partagé pratiquement toutes les secondes de cette maîtrise et que je n'aurais jamais rencontré sans ce labo : merci d'avoir rendu cette expérience si plaisante, j'en suis déjà nostalgique !

Je continue en remerciant ma famille, mes parents et mon frère, pour leur support dans tous les sens de ce terme. Merci papa et maman de m'avoir encouragée depuis si longtemps et d'avoir toujours mis mes études en priorité… c'est enfin fini ! Merci Veng Jean pour tes conseils et tes explications lorsque j'en avais le plus besoin. Je tiens aussi à remercier mes amis qui m'ont accompagnée émotionnellement dans ce parcours. Votre écoute et votre intérêt dans tous mes projets m'ont constamment inspirée à avancer. Je termine évidemment en remerciant mon copain, Louis, qui a cru en moi du début jusqu'à la fin. Merci d'avoir été le premier à célébrer tous mes moindres succès et d'avoir été une source de motivation et d'énergie lors des moments plus sombres. Tes mots d'encouragement m'ont inspirée à persévérer et à me surpasser ; j'en suis si reconnaissante !

Ce mémoire de maîtrise suit le mode de présentation par articles. Ainsi, il est composé d'un article scientifique central rédigé en anglais. L'article est précédé d'une introduction générale et est suivi d'une discussion générale. Ces sections permettent de préciser comment les résultats de l'article empirique s'appliquent dans la recherche et la pratique en psychoéducation.

Contexte théorique

L'évaluation des effets en psychoéducation

Une des compétences du psychoéducateur[1] est de « mettre en œuvre une intervention (…) et d'en assurer le suivi » (Ordre des psychoéducateurs et psychoéducatrices du Québec, 2018, p.40). Ce suivi des effets s'insère dans l'évaluation post-situationnelle : l'opération professionnelle à travers laquelle le psychoéducateur s'assure que l'intervention soit bien adaptée aux besoins du client et à sa situation (Gendreau et collaborateurs, 2001). D'une part, l'évaluation des effets permet d'apprécier l'évolution du client ; d'une autre part, elle est nécessaire pour soutenir le professionnel dans la mise en œuvre de son intervention (poursuite, modification ou interruption ; Kazdin, 2019). Une évaluation juste des résultats et de l'efficacité d'une intervention permet d'éviter des dépenses inutiles de ressources en temps et argent. Pour le client, cette démonstration des progrès peut être une source de confiance et de motivation (Boswell et al., 2013).

Méthodes de suivi des effets

Ce mémoire met principalement l'accent sur les protocoles à cas uniques. Toutefois, il existe d'autres façons de suivre les effets d'une intervention et d'évaluer si les changements sont significatifs. La méthode de Jacobson et Truax (1991), par exemple, se fie aux seuils cliniques de l'instrument de mesure utilisé, afin de déterminer si le résultat post-intervention du client s'approche du niveau de fonctionnement d'une population non clinique.

[1] Le masculin est utilisé dans ce texte uniquement pour alléger la forme et ainsi faciliter la lecture.

Prétest-posttest

La méthode la plus simple pour faire le suivi des effets d'une intervention est de prendre une mesure pré-intervention et une mesure post-intervention, puis de les comparer. Cette méthode est facile à implanter et elle nécessite peu de ressources parce qu'elle ne requiert que deux prises de mesures, puis elle implique des analyses simples. Avec seulement deux mesures, il est toutefois difficile de reconnaître un changement réellement significatif et de déterminer si les effets sont le résultat ou non de l'intervention. Les facteurs historiques ou la maturation sont des exemples de variables confondantes qui peuvent expliquer les changements observés (Marsden et Togerson, 2012). De plus, le psychoéducateur obtient les résultats seulement lorsque l'intervention est terminée et ne peut pas apporter rapidement des modifications si celle-ci s'avère inefficace.

Protocole AB

Une autre méthode plus rigoureuse est l'utilisation du protocole quasi-expérimental à cas uniques AB. Ce protocole consiste en la prise de mesures répétées avant la mise en œuvre de l'intervention (phase A : niveau de base) et durant ou après l'intervention (phase B). Un minimum de cinq mesures en niveau de base (ou jusqu'à l'observation de mesures stables) est habituellement souhaité avant le passage à la phase d'intervention (Kratochwill et al., 2010). La répétition des mesures permet d'observer une certaine stabilité dans chaque phase et d'évaluer la possibilité qu'il existait déjà une tendance d'amélioration pré-intervention. Quoique le protocole nécessite plus de temps, l'exigence de mesures répétées en niveau de base demeure réalisable dans la majorité des milieux d'intervention où les psychoéducateurs s'allouent une période d'évaluation initiale. Ce protocole est dit « quasi-expérimental » parce qu'il n'implique pas une reproduction des effets. Pour cette raison, des enjeux de validité interne persistent et le professionnel ne peut inférer des relations causales entre les variables (Kratochwill et al., 2010), bien que certains chercheurs

explorent la possibilité d'introduire une randomisation (Michiels et Onghena, 2019) ou un seuil de taille d'effet (Lanovaz, Turgeon et al., 2019). Malgré ces limites, les caractéristiques du protocole sont avantageuses d'un point de vue pratique (aucun retrait de l'intervention, simple et rapide) ; c'est pourquoi le protocole AB demeure une option intéressante de suivi lorsque l'isolation des effets de l'intervention n'est pas la priorité. Dans le cas où la validité interne prime, d'autres méthodes de suivis peuvent être envisagées.

Protocoles expérimentaux à cas uniques

Pour un suivi plus rigoureux des effets, le psychoéducateur peut utiliser des protocoles expérimentaux à cas uniques (Lanovaz, 2013, Kazdin, 2019). Contrairement aux méthodes précédentes, ces protocoles impliquent la manipulation de l'intervention et la reproduction des effets expérimentaux (Kratochwill et al., 2010). La manipulation de l'intervention peut se faire en déterminant le moment ou la condition sous laquelle l'intervention sera introduite, retirée (ex., ABAB), décalée (ex., niveaux de bases multiples) ou modifiée (p. ex., changement de critères). Ces protocoles permettent d'observer le comportement ciblé sous différentes conditions (niveau de base, intervention, intervention alternative ou à plus grande intensité ; Kazdin, 2019). Autrement dit, ils permettent de tester la variable dépendante à différentes intensités de la variable indépendante. Une reproduction des effets a lieu lorsque le chercheur/intervenant répète la séquence de manipulation, créant ainsi une nouvelle occasion d'observer les changements souhaités. Ces deux caractéristiques des protocoles expérimentaux à cas uniques visent à écarter le plus de variables confondantes possibles, ce qui augmente leur validité interne (Kratochwill et al., 2010). En reproduisant les effets au sein d'une même personne, ces protocoles permettent d'obtenir une reproduction intra-participant et ne requièrent qu'un petit échantillon.

Pour leur validité et leurs avantages pratiques en temps et en financement, les protocoles expérimentaux à cas uniques sont souvent utilisés en recherche pour l'évaluation de programme ou d'intervention (Manolov et Onghena, 2017). Toutefois, ces protocoles trouvent également leur utilité en pratique, même si un psychoéducateur s'intéresse davantage à l'observation de changements significatifs qu'à l'isolation des effets d'une intervention. Les données objectives et rigoureuses peuvent servir de compléments aux impressions cliniques du psychoéducateur, particulièrement lorsqu'il travaille auprès d'une nouvelle clientèle, d'un nouveau comportement ou lorsqu'il met en œuvre une nouvelle intervention ou activité (Krasny-Pacini et Evans, 2018). En outre, les clients peuvent eux aussi profiter du suivi de leurs progrès ; des études démontrent que des programmes d'évaluations continues en santé mentale rassurent les patients plus sceptiques et ont un impact bénéfique sur leur motivation et l'alliance thérapeutique (Boswell et al., 2013 ; Youn et al., 2012). Un suivi concret pourrait soutenir un client qui entreprend un changement difficile (p. ex., un parent qui ignore des demandes d'attention de son enfant pourrait avoir de la difficulté à réaliser que la durée des crises diminue petit à petit). Selon le type d'intervention, de comportement ciblé ou l'objectif du suivi, le professionnel peut se référer à des arbres décisionnels afin de choisir le meilleur protocole expérimental à cas uniques pour leur réalité (Lanovaz, 2013 ; Ledford et al., 2019 ; Krasny-Pacini et Evans, 2018).

Protocole avec alternance de traitements

Un exemple de protocole expérimental à cas uniques est le protocole avec alternance de traitements (PAT ; Barlow et al., 2009 ; Kratochwill et al., 2010). Ce protocole est utile pour comparer rapidement différentes conditions sur un comportement réversible. La Figure 1.1 présente un exemple de graphique d'un PAT comparant les effets de deux conditions (A et B) sur un comportement x. Le chercheur/psychoéducateur peut appliquer le PAT lorsqu'il s'intéresse à

l'évaluation de différentes interventions alternatives ou simplement à l'efficacité d'une intervention (en la comparant à son absence).

Avec un maximum de deux séances consécutives d'une même condition et d'un minimum de cinq répétitions de la séquence d'alternance, le PAT est considéré comme rigoureux selon les standards du What Works Clearinghouse (Kratochwill et al., 2010). En effet, l'alternance de traitements démontre que le comportement ciblé varie en fonction de la condition présentée et la répétition de cette séquence permet de reproduire les effets expérimentaux désirés. De cette manière, le PAT écarte les variables confondantes de maturation et toute tendance pré-intervention (Hayes et Blackledge, 1998). Le contrôle des tendances à l'intérieur du protocole est également assuré par l'absence de phases distinctes. Compte tenu des caractéristiques du protocole, les interventions ciblées doivent pouvoir être mises en place et retirées rapidement. Par conséquent, le comportement ciblé doit être réversible, c'est-à-dire qu'il pourrait retourner à son niveau de base en l'absence d'intervention. Un comportement irréversible est un comportement dont le changement induit par l'intervention serait permanent (impliquant souvent un processus d'apprentissage). Par exemple, un enfant pourrait recommencer à lancer ses jouets face au retrait de son système de récompense (réversible) mais il ne pourrait pas « désapprendre » à se servir d'un ustensile même si l'apprentissage s'interrompait (irréversible). Sans contredit, il est essentiel de s'assurer que le retrait de l'intervention soit acceptable d'un point de vue éthique (comportements dangereux) et pratique (temps, confusion des clients). Dans ces cas-là, d'autres protocoles expérimentaux peuvent être suggérés (ex., protocole avec alternance de traitements *adapté*, protocole à niveaux de bases multiples).

Le PAT s'applique de manière réaliste en psychoéducation, bien qu'il semble plus complexe que les méthodes présentées précédemment (Herrera et Kratochwill, 2005). Tout d'abord, le protocole n'exige pas de mesures pré-intervention. S'il le souhaite, le psychoéducateur

peut collecter ses mesures de niveau de base en alternance avec les mesures d'intervention. Cet aspect permet au protocole de s'étendre aux milieux d'intervention où la durée d'évaluation est plus restreinte ou lorsque les besoins doivent être adressés plus rapidement. Un exemple serait le milieu hospitalier où, en raison du mandat de gestion de crise, les interventions sont mises en place rapidement et l'évaluation se fait au fur et à mesure du séjour. De plus, en alternant rapidement les conditions, le psychoéducateur peut obtenir un aperçu des comparaisons recherchées très tôt. Des prises de décisions peuvent ainsi se faire plus promptement. Enfin, l'intervention ciblée n'est pas retirée pendant une longue période puisque l'alternance des conditions se fait après quelques mesures seulement. Par exemple, un enseignant pourrait mettre en place un système de récompense pendant seulement certaines périodes, obtenant alors des mesures de chaque condition en l'espace d'une seule journée. Avec le PAT, Adamson et Lewis (2017) ont pu tester et comparer rapidement les effets de trois interventions visant à augmenter l'engagement d'élèves du secondaire à l'intérieur d'un seul protocole. Une autre utilité du PAT est l'évaluation d'une composante spécifique d'un programme d'intervention. Bassette et Taber-Doughty (2016) ont examiné si la présence d'un chien d'assistance dans leur programme de lecture avait un effet sur les capacités de lecture, de compréhension et la motivation de 4 élèves. En utilisant le PAT, les chercheurs n'ont pas à se soucier aux différences inter-participants dans leur comparaison puisqu'ils peuvent administrer les deux conditions auprès de chaque enfant.

Méthodes d'analyse des données pour le protocole avec alternance de traitements

L'analyse visuelle

L'analyse visuelle a longtemps été utilisée avec les protocoles à cas uniques et elle représente le type d'analyse le souvent utilisé dans les études avec alternance de traitements entre 2010 et 2015 (Manolov et Onghena, 2017). Cette analyse permet de calculer ou d'apprécier

visuellement différentes dimensions du graphique. Le chercheur ou professionnel peut obtenir des informations en analysant chaque phase séparément ainsi que les changements d'une phase à l'autre. L'analyse visuelle permet en général d'évaluer ou comparer le niveau d'occurrence du comportement (ex., fréquence, durée, pourcentage), sa tendance, sa variabilité, le chevauchement entre les phases, l'immédiateté et l'uniformité des données (Kratochwill et al., 2010; Lane et Gast, 2013 ; Ledford et al., 2017 ; Ledford et al., 2019). L'analyse visuelle s'insère convenablement dans la pratique psychoéducative pour sa simplicité et sa rapidité d'utilisation. Néanmoins, cette analyse présente des limites. Lors de l'évaluation d'une intervention à travers un protocole avec alternance de traitements, une analyse visuelle permet de constater s'il y a présence d'une relation fonctionnelle entre les variables (observation répétitive de l'effet de l'intervention sur le comportement ciblé en fonction de la manipulation de l'intervention), mais elle ne conclue pas sur la taille d'effet (Ledford et al., 2017) tel que recommandé (Kratochwill et al., 2010 ; Ledford et al., 2019). Par ailleurs, Ninci et al. (2015) démontrent que les études utilisant l'analyse visuelle (seule) présentent une moyenne d'accord inter-juge faible (0,69). Quoique le changement soit nonsignificatif, la méta-analyse démontre une augmentation de l'accord-interjuge lorsque les études utilisent des soutiens visuels (0,75). Ces soutiens visuels consistent par exemple enune ligne d'estimation de la tendance des données. De ce fait, plusieurs analystes recommandent des soutiens visuels ainsi que des analyses statistiques comme compléments à l'analyse visuelle (Manolov et Onghena, 2017).

Le critère structuré visuel

Comme complément à l'analyse visuelle classique, Lanovaz, Cardinal et Francis (2019) ont développé le critère structuré visuel (VSC) pour les protocoles avec alternance de traitements. Cette méthode simple de deux étapes permet de comparer deux conditions en se basant sur la position

des points et trajectoires dans les graphiques. Cette méthode s'insère facilement dans la pratique professionnelle parce qu'elle implique des traçages simples et des seuils prédéterminés par les chercheurs. Cette méthode a été validée et reproduite sous différentes conditions. Lorsque le protocole a au moins 5 séances par condition, les chercheurs démontrent que le VSC contrôle suffisamment le taux d'erreur de type I ($p < 0,05$; Lanovaz, Cardinal et Francis, 2019 ; Manolov, 2019). En ce qui concerne la puissance statistique, les chercheurs démontrent que le VSC est meilleur lorsque la séquence d'alternance est systématique (ABABABABAB), plutôt que randomisée (Manolov, 2019). En effet, le VSC est en mesure de détecter des tailles d'effets de 2 à partir de 10 séances (5 par conditions) dans une alternance systématique (0.80). Or, le VSC atteint cette même puissance seulement à partir de 12 séances pour une randomisation par blocs (RB) puis à partir de 16 séances pour une randomisation restreinte (RR ; Manolov, 2019 ; voir Tableau 1.1 pour une présentation des types de randomisation).

Test de randomisation et ALIV

Or, l'introduction d'une randomisation augmente la validité interne et crédibilité scientifique du protocole en adressant des menaces tels les effets d'histoire et de séquence (Barlow et al., 2009 ; Kratochwill et al., 2010 ; Kratochwill et Levin, 2010 ; Manolov et Onghena, 2017). La randomisation s'applique en choisissant de façon aléatoire la séquence de présentation des conditions. Cette affectation aléatoire rend le test de randomisation une méthode d'analyse idéale. L'hypothèse nulle de ce test non-paramétrique suppose qu'il n'existe aucune différence significative entre les deux conditions comparées. Pour ce faire, le test permute la séquence de conditions du protocole en redistribuant les scores, créant une distribution d'échantillonnage à partir des scores observés. Le test de randomisation évalue la probabilité d'obtenir l'effet observé parmi cette distribution. La valeur-p illustre alors la proportion des permutations ayant une taille

d'effet égale ou supérieure (ou inférieure selon la direction souhaitée) à l'effet observé. L'avantage de ce test est la possibilité de l'appliquer sur n'importe quelle mesure de taille d'effet (ex., différence de moyenne, écart-type).

Afin de quantifier la taille d'effet dans les PAT, Manolov et Onghena (2017) proposent le *Actual and linearly interpolated values* (ALIV), qui adresse les limites de l'analyse visuelle et des autres statistiques habituellement utilisées. L'ALIV compare les valeurs de chacune des phases, en considérant à la fois les données réelles et à la fois les données interpolées. Les données interpolées servent d'estimation de données qui auraient été obtenues si les chercheurs n'avaient pas alterné de condition. Autrement dit, les valeurs interpolées de la condition A sont estimées aux temps où les valeurs réelles de la condition B ont été prises, en suivant la fonction linéaire entre les valeurs réelles de la condition A précédentes et suivantes. Les comparaisons sont faites pour chaque portion du graphique ce qui illustre comment les deux phases se différencient au fil du suivi. Cette méthode est avantageuse parce qu'elle donne un meilleur aperçu de l'évolution des données plutôt que de seulement prendre les données dans son ensemble (ex., la différence de moyenne). De plus, l'ALIV est en mesure d'analyser précisément des données même si elles sont non-linéaires ou présentent des chevauchements. Manolov (2019) démontre que le test de randomisation utilisant l'ALIV comme statistique présente un taux d'erreur de type I adéquat ($p < 0,05$) et une meilleure puissance statistique que le VSC ($\geq 0,80$), peu importe le type de randomisation utilisé (10 séances pour une RR et 12 séances pour une RB). Néanmoins, tout comme le VSC, cette méthode n'est puissante que pour des tailles d'effets de 2 et plus (Lanovaz, Cardinal et Francis, 2019 ; Manolov, 2019).

Ratio de distance

Le ratio de distance (RD) est une autre mesure de taille d'effet proposé par Carlin et Costello (2018) qui se base sur les différences/distances au sein des phases et entre les phases. Un

RD significatif (>1.2) permet d'écarter l'hypothèse que les différences observées sont dues seulement à de la variabilité aléatoire. Cette analyse est intéressante pour sa capacité d'adaptation à différents types de données : nombres de points différents par condition, présence de tendances (dans chaque phase ou dans son ensemble). De plus, cette statistique est comparable avec le *d* de Cohen, une mesure habituellement utilisée dans des essais randomisés. Enfin, le RD s'applique aisément en pratique étant donnée sa facilité d'utilisation : calculs simples et logiciel connu (Microsoft Excel). Cette analyse a été utilisée avec des protocoles à niveaux de bases multiples et ABAB (Deochand et al., 2019 ; Costello et al., 2019), mais n'a pas encore été appliquée avec des PAT. Carlin et Costello (2018) démontrent un contrôle adéquat des erreurs de type I ($p < 0,05$) ainsi qu'une bonne puissance statistique (0.94) avec des tailles d'effet de 2. Davantage d'études sont nécessaires pour reproduire ces résultats auprès de différents protocoles et tailles d'effet.

L'apprentissage automatique supervisé

Dans cette étude, il sera spécifiquement question de l'apprentissage automatique supervisée dans le but d'une classification binaire. L'apprentissage automatique (AA) est une sous-catégorie d'intelligence artificielle visant à estimer un modèle de prédiction en se basant sur des séries de données d'exemples. Dans l'AA supervisée, les séries d'exemples consistent à la fois des caractéristiques des données (ex., âge et sexe de la population) et à la fois de leur étiquette (ex., malade ou en santé). L'objectif du modèle est de prédire l'étiquette exacte (variable de sortie) de n'importe quelle donnée en se basant sur ses caractéristiques (variables d'entrées). Afin de produire un modèle de prédiction précis, les algorithmes d'AA analysent les séries de données d'exemples afin de reconnaître le schéma ou la fonction qui associent le mieux les variables d'entrées aux variables de sorties (entraînement du modèle). Une fois le modèle estimé, celui-ci peut être utilisé pour prédire de nouvelles données inconnues et non-étiquetées (généralisation du modèle). Cette logique est similaire à celle de l'utilisation en psychoéducation où l'intervenant enseigne des

concepts à son client dans le but qu'il les généralise à d'autres situations ou contextes de sa vie. Lors de l'enseignement d'habiletés sociales par exemple, l'intervenant offre à la fois les mises en situation (variables d'entrée) et à la fois les comportements adaptés associés (variables de sortie). Une fois que les habiletés sociales sont maîtrisées, le client doit identifier lui-même le comportement adapté à présenter selon la situation à laquelle il fait face (prédiction de la variable de sortie en fonction des variables d'entrée). Tel que son nom l'indique, une classification binaire consiste d'une variable de sortie catégorielle de deux classes possibles (ex., vrai ou faux).

L'AA représente une option intéressante d'analyse de données pour les PAT. Les modèles d'AA peuvent analyser les caractéristiques de chaque graphique afin de reconnaître les interventions ayant un « effet » ou n'ayant « aucun effet ». Ces modèles présentent plusieurs avantages, notamment parce qu'ils sont estimés à partir des séries de données choisies comme variables d'entrées par l'expérimentateur. Avec suffisamment d'exemples différents, le modèle peut continuer à s'améliorer et à s'adapter à différents types de données, telle que la présence de tendances non-linéaires. De plus, les algorithmes de l'AA peuvent considérer dans leurs analyses toutes les dimensions visées par l'analyse visuelle. Lanovaz et al. (2020) ont comparé l'utilisation de modèles de classification supervisée avec la méthode à double-critère pour analyser des protocoles à cas uniques AB et ont démontré la supériorité des modèles de l'AA en termes de précision, de puissance statistique et de contrôle des taux d'erreurs de type I. Ces résultats suggèrent la possibilité d'intégrer des modèles de classification supervisée afin d'analyser des données d'intervention psychosociale, telle qu'en psychoéducation.

Questions de recherche

Dans cet ordre d'idées, une reproduction de l'étude de Lanovaz et al. (2020) auprès de PAT permettrait d'évaluer la généralisation de leurs résultats à un protocole à cas uniques plus rigoureux. Par ailleurs, aucune étude n'a encore comparé les différentes méthodes d'analyse actuellement utilisées avec les PAT avec un modèle estimé par l'AA. Cette étude cherche alors à répondre aux questions de recherche suivantes :

1. Le critère structuré visuel, l'actual and linearly interpolated values avec le test de randomisation, et le ratio de distances permettent-ils d'analyser adéquatement les protocoles avec alternance de traitement ?
2. Un modèle estimé par un algorithme d'apprentissage automatique permet-il d'analyser les protocoles avec alternance de traitements avec autant de précision, de puissance statistique et de contrôle des taux d'erreurs de type I que les méthodes d'analyses statistiques précédentes ?

Références bibliographiques

Adamson, R. M. et Lewis, T. J. (2017). A comparison of three opportunity-to-respond strategies on the academic engaged time among high school students who present challenging behavior. *Behavioral Disorders, 42(2),* 41-51. https://doi.org/10.1177/0198742916688644

Barlow, D. H., Nock, M. K. et Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

Bassette, L. A. et Taber-Doughty, T. (2016). Analysis of an animal-assisted reading intervention for young adolescents with emotional/behavioral disabilities, *RMLE Online, 39(3)*, 1-20, https://doi.org/10.1080/19404476.2016.1138728

Boswell, J. F., Kraus, D. R., Miller, S. D. et Lambert, M. J. (2013). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25(1)*, 6-19. https://doi.org/10.1080/10503307.2013.817696

Carlin, M. T. et Costello, M. S. (2018). Development of a distance-based effect size metric for single-case research: Ratio of distances. *Behavior Therapy, 49*, 981-994. https://doi.org/10.1016/j.beth.2018.02.005.

Costello, M. S., Sheibanee, B. D., Ricketts, A., Hirsh, J. L. et Deochand, N. (2019). Exploration of social reinforcement for gambling in single case designs. *Analysis of Gambling Behavior, 12(1)*, 1-21. https://repository.stcloudstate.edu/agb/vol12/iss1/1

Deochand, N., Costello, M. S. et Fuqua, R. W. (2019). Real-time contingent feedback to enhance punching performance. *The Psychological Record, 70(1),* 33-45. https://doi.org/10.1007/s40732-019-00357-2

Gendreau, G. et collaborateurs. (2001). *Jeunes en difficulté et intervention psychoéducative.* Montréal, Québec : Éditions Sciences et Culture.

Herrera, G. C. et Kratochwill, T. R. (2005). Alternating treatments designs. Dans B. S. Everitt et D. C. Howell (dir.), *Encyclopedia of Statistics in Behavioral Science*. Wiley & Sons. https://doi.org/10.1002/0470013192.bsa017

Jacobson, N. S. et Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19. https://doi.org/10.1037/0022-006X.59.1.12

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy, 117*, 3-17. https://doi.org/10.1016/j.brat.2018.11.015

Krasny-Pacini, A. et Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicin, e61(3)*, 164-179. https://doi.org/10.1016/j.rehab.2017.12.002

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. et Shadish, W. R. (2010). *Single-case designs technical documentation.* What Works Clearinghouse. http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R. et Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15(2),* 124-144. https://doi.org/10.1037/a0017736

Lane, J. D. et Gast, D. L. (2013). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24(3-4), 445-463.* https://doi.org/10.1080/09602011.2013.815636

Lanovaz, M. J., Cardinal, P. et Francis, M. (2019). Using a visual structured criterion for the analysis of alternating-treatment designs. *Behavior Modification, 43(1),* 115-131. https://doi.org/10.1177/0145445517739278

Lanovaz, M. J., Giannakakos, A. R. et Destras, O. (2020). Machine learning to analyze single-case data: A proof of concept. *Perspectives on Behavior Science, 43,* 21-38. https://doi.org/10.1007/s40614-020-00244-0

Lanovaz, M. J., Turgeon, S., Cardinal, P. et Wheatley, T. L. (2019). Using single-case designs in practical settings: is within-subject replication always necessary? *Perspectives on Behavior Science, 42,* 153-162. https://doi.org/10.1007/s40614-018-0138-9

Ledford, J. R., Barton, E. E., Severini, K. E. et Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities, 124(1),* 35-56. https://doi.org/10.1352/1944-7558-124.1.35

Ledford, J. R., Lane, J. D. et Severini, K. E. (2017). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment, 19(1),* 4-17. https://doi.org/10.1017/BrImp.2017.16

Manolov, R. et Onghena, P. (2017). Analyzing data from single-case alternating treatments designs. *Psychological Methods, 23(3),* 480-504. https://doi.org/10.1037/met0000133

Manolov, R. (2019). A simulation study on two analytical techniques for alternating treatments designs. *Behavior Modification, 43*, 544-563. https://doi.org/10.1177/0145445518777875

Marsden, E. et Togerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education, 38(5)*, 583-616. https://doi.org/10.1080/03054985.2012.731208

Michiels, B. et Onghena, P. (2019). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*, 2452-2476. https://doi.org/10.3758/s13428-018-1084-x

Ninci, J., Vannest, K. J., Willson, V. et Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification, 39(4),* 510-541. https://doi.org/10.1177/0145445515581327

Ordre des psychoéducateurs et psychoéducatrices du Québec. (2018, Mai). *Le référentiel de compétences lié à l'exercice de la profession de psychoéducatrice ou psychoéducateur au Québec.* https://www.ordrepsed.qc.ca/fr/profil-des-competences/~/media/pdf/Psychoeducateur/Rf%20de%20comptences%20Version%20adopte%20par%20le%20CA%20duconseil%2017%20mai%202018.ashx?la=fr

Youn, S. J., Kraus, D. R. et Castonguay, L. G. (2012). The treatment outcome package: Facilitating practice and clinically relevant research. *Psychotherapy, 49(2),* 115-122. https://doi.org/10.1037/a0027932

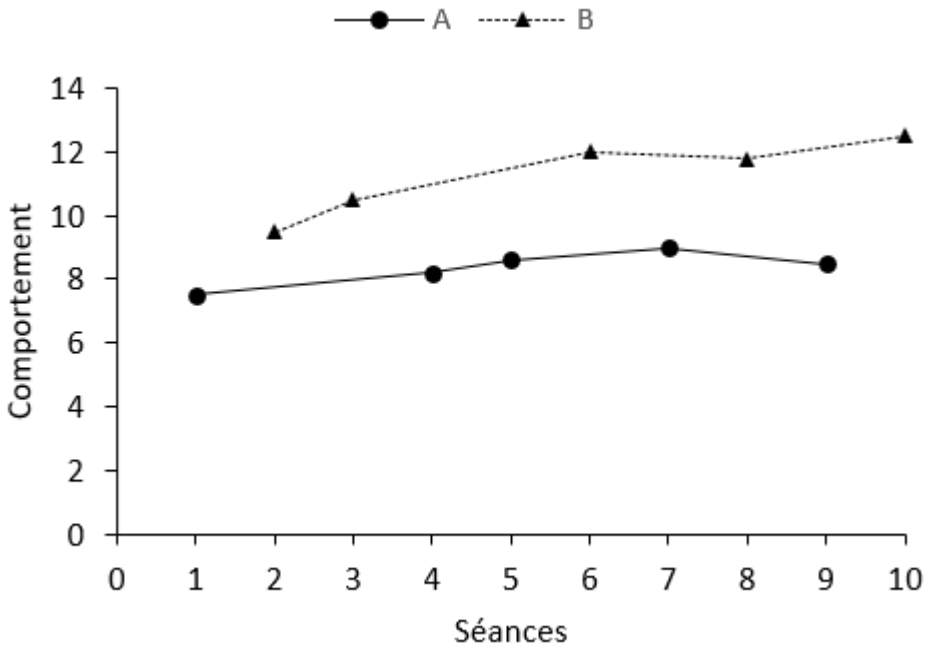
Tableau 1.1

Définition des types de randomisation dans les protocoles avec alternance de traitements

Types de randomisation	**Définition**
Randomisation par blocs	Pour cette randomisation, la séquence de conditions est séparée en paires. La première condition de chaque paire est choisie de manière randomisée. La deuxième condition de chaque paire est alternée en fonction de la première. Ex., AB-BA-AB-AB-BA.
Randomisation restreinte	La randomisation est faite avec la séquence de conditions dans son ensemble, avec une restriction de deux conditions identiques consécutives maximales. Ex., ABBAABAABB.

Figure 1.1

Exemple de graphique d'un protocole avec alternance de traitements classique.



Les pages suivantes présentent un article scientifique rédigé en anglais, dans lequel les questions de recherches ont été explorées. Par la suite, une discussion générale en français sera présentée afin de conclure ce mémoire de maîtrise en faisant les liens entre les résultats de l'article empirique ainsi que la recherche et la pratique en psychoéducation.

Artificial Neural Networks to Analyze Alternating-Treatment Designs

Emily Heng and Marc J. Lanovaz

Université de Montréal

Abstract

Alternating-treatment designs are useful single-case experimental designs for the evaluation and comparison of intervention effectiveness. Most guidelines suggest that researchers and practitioners use a combination of statistical and visual analyses to analyze these designs, but current methods still produce inadequate levels of errors under certain conditions. In an attempt to address this issue, our study examined the use of artificial neural networks to analyze alternating-treatment designs and compared their performances to three other recent approaches. Specifically, we examined accuracy, statistical power, and type I error rates under various conditions. Albeit not perfect, the artificial neural networks model generally provided better and more stable results across analyses. Our results suggest that artificial neural networks are promising alternatives to analyze alternating-treatment designs.

Keywords: alternating-treatment design, artificial neural networks, machine learning, N-of-1 trials, type I error

Artificial Neural Networks to Analyze Alternating-Treatment Designs

Researchers in psychology, education, and medicine often use single-case experimental designs to evaluate the effects of an independent variable (Lillie et al., 2011; Kazdin, 2019). While randomized controlled trials have many advantages in terms of control, power, and generalizability, these trials also require large groups of participants and substantial resources. Single-case experimental designs are a rigorous alternative when studying new interventions, uncommon behaviors, or rare populations (Krasny-Pacini & Evans, 2018). Furthermore, practitioners also seek to evaluate intervention effectiveness. Single-case experimental designs are applicable in clinical settings and offer benefits to both practitioners and clients. Researchers have shown that the continuous monitoring of one's progress provides comfort and encouragement for the client and can also improve the therapeutic relationship (Boswell et al., 2013; Youn et al., 2012).

The alternating-treatment design is a single-case experimental design that rapidly compares the effects of multiple conditions on a reversible behavior (Barlow et al., 2009; Kratochwill et al., 2010). Researchers frequently use this design to compare multiple treatments or to evaluate the effectiveness of a single intervention within a brief period of time. The alternating-treatment design presents practical advantages over other single-case experimental designs as it does not require baseline stability, which may reduce the number of sessions. This design meets the standards of the What Works Clearinghouse when it alternates from one condition to another after a maximum of two sessions and when it repeats the alternation sequence at least five times (Kratochwill et al., 2010).

Visual analysis is a well-established and recommended method in the analysis of single-case experimental designs (Kratochwill et al., 2010). In fact, researchers have used this method the most often in studies using alternating-treatment design between 2010 and 2015 (Manolov & Onghena, 2017). While the visual analysis of alternating-treatment graphs can assess various

dimensions of behavior change such as level, trend, variability, overlap, immediacy and consistency (Kratochwill et al., 2010; Lane & Gast, 2013; Ledford et al., 2017; Ledford et al., 2019), researchers have documented inadequate interrater agreement rates when using visual analysis alone (Lanovaz et Turgeon, 2020; Ninci et al., 2015). Guidelines in the use of single-case designs support the benefits of using visual analysis and statistical analyses jointly to evaluate both the presence of a functional relation as well as effect sizes (Kratochwill et al., 2010; Ledford et al., 2019).

Several statistical methods currently exist to analyze alternating-treatment design such as the mean difference and the percentage of nonoverlapping data (see Manolov & Onghena, 2017, for a review of existing analytical techniques). However, these methods require specific data characteristics, which produce poor performances in the presence of trend, overlap, or unstable data. To address this issue, researchers and analysts have come up with new analytical methods.

In a recent study, Manolov and Onghena (2017) developed the actual and linearly interpolated values (ALIV), which uses both actual and interpolated points in its computation. Interpolated points represent the values that would have been obtained if analysts did not alternate conditions. Unlike other measures, the ALIV performs well even with data showing overlap or non-linear trends. Moreover, this statistic makes comparisons within each portion of the graph, which provides a better representation of the evolution of data. To obtain statistical significance with this measure, researchers suggest the use of the randomization test (Manolov & Onghena, 2017; Manolov, 2019). This non-parametric test assumes random assignment; an assumption met when the design includes randomization in its alternating sequence. The test creates a distribution by permutating the observed values and computing new ALIV scores for each permutation. The test then determines the probability of obtaining the observed result within this distribution (i.e. the proportion of permutation showing scores equal or higher than the observed score). The

randomization test is advantageous for its ability to perform with random assignment and any measure. Validated on simulated data, the randomization test with the ALIV shows adequate control over type I error rates and strong statistical power, regardless of randomization type (needing a minimum of 10 sessions for the restricted randomization and 12 sessions for the block randomization. Nevertheless, this method is only powerful for effect sizes of 2 and larger (Manolov, 2019).

Similarly, Lanovaz et al. (2019) have developed and validated the visual structured criterion (VSC) for the analysis of alternating-treatment design, as a complement to classical visual analysis. Based on points and trajectories in a graph, the VSC aims to compare conditions and determine the presence of a statistically significant difference other than chance. With a minimum of 5 sessions per condition (10 total), this method shows sufficient control over type I error rates. The VSC also presents good statistical power for effect sizes of 2 and larger, when conditions are alternated systematically. However, Manolov (2019) has shown that the VSC's power decreases when randomization is present (needing 12 sessions for block-randomization and 16 sessions for restricted randomization). These results show a limitation for the VSC as the inclusion of randomization in a design increases its internal validity (Kratochwill et al., 2010; Kratochwill & Levin, 2010; Manolov & Onghena, 2017).

A third method that researchers have developed recently is the ratio of distance (RD), which is based on the distance both within- and between conditions (Carlin & Costello, 2018). The RD presents several advantages over commonly used analytical methods. Firstly, this approach performs well with various data characteristics including trends and different numbers of points per condition. Secondly, this method presents an easy computation and uses common analytical programs such as Microsoft Excel. Finally, analysts may directly compare the RD to the Cohen's *d*. While researchers have demonstrated adequate control over type I error rates and sufficient

statistical power for effect sizes of 2 and larger, the RD lacks replications. Besides, researchers have only used the RD with multiple baseline designs (Deochand et al., 2019; Costello et al., 2019), but have yet validated this method with alternating-treatment designs.

Given the previous limitations, machine learning appears as an interesting solution to the analysis of alternating-treatment designs. Machine learning is a subclass of artificial intelligence aiming to estimate a model that can accurately predict output variables (i.e., class labels) by analyzing input variables (i.e., features). In supervised machine learning, experimenters provide a series of example data (inputs and outputs) to train the model in recognizing recurring functions. Input variables consist of the data characteristics (i.e., a population's age and sex; an image's color and forms), while the output variables consist of the label/prediction sought (i.e., true or false, cat or dog, effective or ineffective). In other words, machine learning models "learn" series of examples (training) to be able to classify new data based on the input variables (generalization). This logic is similar to human learning as it uses known experiences/concepts to understand new similar situations.

Lanovaz et al. (2020) have recently tested the use of supervised binary classification to analyze a single case design (AB design) and have demonstrated its superiority over another well-established analytical technique (i.e., dual-criteria method). These results suggest the possibility of applying machine learning models with a single-case experimental design to support researchers and practitioners in accurately evaluating the effectiveness of interventions. The choice of machine learning algorithms to analyze alternating-treatment designs is appealing because its models are estimated by analyzing input variables specifically chosen by the experimenter. By doing so, machine learning algorithms can improve models to accurately make predictions with any type of data, which is the main limitation of current analytical techniques. Moreover, machine learning models can consider all dimensions of visual analysis simultaneously. To our knowledge, no

studies have yet compared current analytical methods with machine learning models in the analysis of alternating-treatment design. This study, therefore, seeks (a) to replicate findings of the new analytical methods presented for alternating-treatment designs (VSC, ALIV with randomization test, RD) and (b) to compare accuracy, power, and type I error rate with a model estimated by a machine learning algorithm.

Method

Procedure

Data Generation

For this study, we conducted a Monte Carlo simulation to explore type I error rates, statistical power, and accuracy score of the targeted analysis methods. A benefit of using simulated data for the evaluation of new methods or the comparison of existing ones is that experimenters have control over data characteristics (Morris et al., 2018). This control allows a determination of each method's strengths and limitations under various conditions. Moreover, simulated data improve confidence over the data's true values, that is whether the data present a significant difference between conditions or not.

We programmed a Python code to generate alternating-treatment datasets. We used block-randomization[2] to choose conditions order. We chose a randomization method that researchers and practitioners are most likely to use, considering it ensures a balanced design. We generated datasets using the following equations:

$$y_i = ti + dSMD + \varepsilon_i + c$$

$$\text{where } \varepsilon_i = ay_{i-1} + u_i$$

[2] For this type of randomization, the sequence of conditions is separated into pairs. Within each pair, the first condition is chosen at random and the second condition is alternated accordingly.

where y is the data point value, i is the time parameter or the session number, t is the trend parameter, d is the dummy variable representing the condition, SMD is the effect size parameter (i.e., standardized mean difference), ε is the error term, c is a constant variable, a is the autocorrelation parameter, and u is the random disturbance. The dummy variable d took the value of 0 for condition A and the value of 1 for condition B because the SMD was only added to points in condition B. We added a constant c of 10 to only have datasets with positive values. Finally, we added an error term that consists of an autocorrelation value and a random disturbance. First-order autocorrelation is the correlation between all measures of the same behavior across consecutive sessions. In other words, each measure of the same behavior is correlated to its last measures. The random disturbance took a random value following the normal distribution with a mean of 0 and a standard deviation of 1. The a, SMD, t, and number points varied across analyses (see Analyses section).

Artificial Neural Network

What Are Machine Learning and Artificial Neural Networks

The goal of using a machine learning algorithm is to estimate an accurate predictive model. This model should accurately predict the output variable (dependent) by analyzing the input variables (independent). In this current study, the output variable consists of two class labels (effect or no effect) and the input variables consist of the simulated data's descriptive variables. The input variables are also called features. To illustrate how machine learning algorithms work, we will consider the association between our variables to be a linear regression:

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

where y is the dependent variable taking a value of 0 (no effect) or 1 (effect), x is the independent variables, and a is the model's parameters. During training, experimenters use the example datasets

to estimate the weights associated with each feature. In terms of machine learning, the algorithm fits its parameters to the example datasets. The learning process in which the algorithm estimates the parameters are different from one machine learning algorithm to the other. For most algorithms, the experimenter can set the learning process parameters, called hyperparameters. For example, hyperparameters can be the number of times the algorithm repeats its process and adjusts its errors before returning its most accurate parameters. Experimenters can tune the algorithm's hyperparameters to optimize the learning process.

For our analyses, we chose to compare traditional methods with an artificial neural network. The neural network consists of an input layer, one or multiple hidden layers, and an output layer (see Figure 2.1; Park & Lek, 2016). Neurons from each layer are interconnected, which allows the information to pass through the network. Weights represent the estimated association between the outputs of the previous layer and the inputs of the next layer. To estimate the weights following the input layer, the algorithm analyzes the datasets, seeking for a recurring pattern/function. To estimate weights in the hidden layers, the algorithm first computes the activation value (weighted sum of the inputs) inside each neuron. Subsequently, the algorithm applies the chosen activation function (i.e. the rectified linear activation function; ReLU) to determine the output. Information passes through the networks until the last layer, where the model produces a prediction of the class label. This process is called the "forward-propagation".

During training, the algorithm improves its model by comparing each of its predicted value to the actual value and computing the loss (error). The algorithm then goes back through the network to update its weight to minimize the loss, using the gradient descent algorithm[3]. This process is called "backpropagation". Each forward-propagation, called an "epoch", results in a

[3] The gradient descent algorithm computes a loss function for each parameter and uses its derivative to identify the parameters' values for which the loss is the lowest. For more information about the algorithm, see Ruder (2016).

prediction. To produce the best model, the algorithm runs multiple epochs to improve each successive prediction. Theoretically, infinite epochs would produce a perfect predictive model. However, this model would only be accurate for the dataset it was trained on because parameters estimated would be too precise. Therefore, this model would fail to generalize to new datasets. Producing a too specific model is defined as overfitting a model. The number of hidden layers, the number of neurons per hidden layer, and the number of epochs are hyperparameters that experimenters can manually adjust to optimize the learning process.

Training the Model

To train the artificial neural network model, we generated 1,200,000 data series. We trained our models using series with variable numbers of points (a total of 10 to 20 points) and variable SMD values. We generated half of our datasets without an SMD to simulate the absence of an effect and half of our datasets with an SMD (600,000 with no SMD and 600,000 with an SMD greater or equal to 1). For the graphs simulating an effect, the SMD randomly took a value between 1 and 3 from a uniform distribution. Training data series presented no autocorrelation and no trend.

Before training our models, we proceeded to normalize our datasets. To transform our data, the second author adapted the code made available by Lanovaz et al. (2020). The code first transformed the data to z-score and secondly extracted the following features: (a) mean of condition A, (b) mean of condition B, (c) standard deviation of condition A, (d) standard deviation of condition B, (e) intercept of the least squares regression line for condition A, (f) slope of condition A, (g), intercept of the least squares regression line for condition B, and (h) slope of condition B. Subsequently, we split the data into a training set (80%) and a validation set (20%). Hence, we only fed the algorithm with 960,000 graphs when training the models. We kept the remaining 240,000 graphs to tune the hyperparameters.

Our neural networks model had one input layer of eight features, one hidden layer of four neurons, and one output layer of one class output (Figure 2.1). We chose four hidden neurons as it is between the number of inputs and the number of outputs, based on general recommendations made by Heaton (2015). Tuning hyperparameters consists of testing different values of a hyperparameter to determine the one that will create the most accurate model. In our case, we tested the number of epochs to run in order to obtain an accurate model without overfitting it. To do so, we included an early stopping hyperparameter which prevents the algorithm from training further after a maximum number of epochs in which the model accuracy did not improve. Using a separate set of data unknown to the algorithm to tune hyperparameters also helps to avoid overfitting our model. After multiple trials and errors, we attributed three times more weights to type I error rates than to type II error rates during training. To do so, we attributed more weights to negative labels although there was an equal proportion of both labels in the datasets. This method allowed us to sacrifice some statistical power for better type I error rates.

Testing the Model

To test the model, we applied it to the datasets generated in the analyses section. These datasets differed from those that we used during the training.

Actual and Linearly Interpolated Values

Consistent with Manolov (2019), we programmed a Python code to compute the ALIV with the randomization test. Figure 2.2 illustrates an application example of the ALIV to an alternating-treatment design graph. First, the algorithm computed the linearly interpolated values inside each condition (i.e. all points A that would have been obtained at the occasions that all points B were actually obtained, and vice versa). Secondly, the algorithm computed the difference between point B and point A, at each measurement occasion, excluding the first and last points. The ALIV is

defined as the average of all these differences. Subsequently, the algorithm applied the randomization test by permutating the order of the conditions. We used the same type of permutation as we used to generate our datasets (block-randomization). For each possible permutation, the algorithm applied the ALIV which created a distribution of ALIV scores. The randomization test ultimately provided a p-value. This value is the proportion of permutations showing an ALIV score greater or equal to the observed score (or lower, depending on the targeted direction).

Visual Structured Criterion

The second author wrote a Python code to compute the VSC, by adapting the code used in Lanovaz et al. (2019). The VSC also consists of comparing actual points and linearly interpolated points (called "trajectories") of both conditions at each measurement occasion, excluding the first and last ones (Figure 2.2). Instead of computing the difference, the algorithm counted the number of occasions in which condition B was greater than condition A. To determine the presence or absence of an effect, the algorithm used the cutoff values provided by Lanovaz et al. (2019).

Ratio of Distances

Finally, the first and second authors wrote a Python code to compute the ratio of distances described by Carlin and Costello (2018). Table 2.1 shows an example of a table used to compute the RD. To obtain the RD score, the algorithm calculated a quotient in which the numerator was the raw sum of pairwise distances between phases, and the denominator was the absolute sum of within-phase distances. A RD greater than 1.2 indicates the presence of an effect (Carlin & Costello, 2018). If a trend seems present in one of the conditions, the algorithm also offers a phase detrending option. If applied, the algorithm subtracts each value (n+1) by the preceding one (n) and averages these differences to get the "slope-correction factor". The algorithm then creates a

new matrix by multiplying each value by the slope-correction factor and by the value's index. Values' indices start with 0 (the first index being 0, the second index being 1, and so on). If the phase detrending option is applied, then the new matrix replaces the original one for the calculation of the RD.

Analyses

Outcome Measures

To compare all methods, we have programmed Python codes to compute three outcome measures, which include the accuracy score, the type I error rate, and the statistical power. The accuracy score is the ratio of correct predictions to all predictions. As such, accuracy includes both type I and type II errors. Type I error is the misclassification of an intervention as effective; it occurs when a predicted value is positive while the true value is negative. When computing type I error rates, the algorithm only looked at the portion of graphs that were simulated without an effect. For each method, the algorithm counted the number of times a set within this portion was categorized showing an "effect" (class label 1) and divided it by the number of graphs simulated without an effect. Statistical power is the method's ability to detect true effective interventions or graphs showing true effects. When computing statistical power, the algorithm only looked at the portion of graphs that were simulated with an effect. For each method, the algorithm counted the number of times a set within this portion was categorized as showing an "effect" (class label 1) and divided it by the number of graphs simulated with an effect.

Datasets and Parameters

For the analyses, we generated 350,000 datasets to test the effects of each parameter on our outcome measures. These datasets differed from those used during training the models. More specifically, we generated 20,000 graphs to test the effects of the autocorrelation, 140,000 graphs

to test the effects of the trend, 120,000 graphs to test the effects of the total number of points, and 70,000 graphs to test the effects of the SMD. Table 2.2 presents the values set for each parameter by the type of analysis.

First, we tested the effects of the total number of points because analysts have shown that the power of certain methods varied with the size of the data series (Lanovaz et al., 2019; Manolov, 2019). Secondly, we tested the effects of the autocorrelation as researchers have shown that it can affect a method's power and type I error rates (Levin et al., 2011). We tested multiple values of this parameter considering the high heterogeneity noted in autocorrelation values by Shadish and Sullivan (2011). We set the autocorrelation default value to 0.2 based on the same study. Thirdly, we tested the effects of a general linear trend, as it is commonly seen in alternating-treatment designs and because this characteristic is usually a limitation for traditional methods (Manolov & Onghena, 2017). Lastly, we added a standard mean deviation (SMD) to points in condition B to simulate a change in the behavior and to test performances with different effect sizes, particularly smaller ones. Values for the constant c, the dummy variable d, and the random disturbance u_i were set the same way as for the training datasets.

Results

Effects of the Number of Points

Table 2.3 presents the power, type I error rate, and accuracy score for all methods using data series with variable numbers of points. The ANN's model displayed accuracy score and statistical power that were stable and high with all number of points. Results for the RD decreased with large data series whereas results for the ALIV and the VSC decreased with smaller data series. In terms of type I error rates, the RD and the ANN presented respectable results when graphs

displayed at least 12 data points. For the VSC and the ALIV, the number of points did not affect their type I error rates, which remained below .05.

Effects of the Autocorrelation

Table 2.4 presents the power, type I error rate, and accuracy score for all methods using data series with variable autocorrelation values. Overall, the RD and the ANN presented similar results. Both methods presented relatively stable accuracy scores and power regardless of autocorrelation values, but their type I error rates were marginally greater than the 5% threshold when autocorrelation was null or negative. On the other hand, the ALIV and the VSC also presented similar results, but their power was significantly lower (i.e., nearly half) than the RD and ALIV. Their accuracy and power were positively associated with the autocorrelation and results remained low. However, the ALIV and the VSC always produced type I error rates below .05.

Effects of the Trend

Table 2.5 presents the power, type I error rate, and accuracy score for all methods using data series with variable trend values. For all outcome variables, the ALIV and the VSC's results did not vary with the trend and were stable. The VSC was also more accurate and more powerful than the other methods with extreme trend values (-0.3 and 0.3). However, the ANN presented higher accuracy and power for smaller trends. All methods exhibited appropriate type I error rates.

Effects of the Standard Mean Deviation

Table 2.6 presents the statistical power for all methods using data series with variable effect sizes. As expected, all methods performed better with larger SMDs. Power exceeded 90% for effect sizes of 2 and larger for the RD and the ANN. To attain the same power, the ALIV and the VSC required larger effect sizes (3 and 2.5, respectively). Results for the RD and the machine learning model were very similar. Both methods detected smaller effect sizes better than the ALIV and the

VSC. Figure 2.3 presents a decisional algorithm as a general guideline in the choice of an analytical method based on datasets characteristics.

Discussion

The current study explored the analytical strengths and limitations of different methods to analyze alternating-treatment designs and compared them to an artificial neural network. Our results replicated findings of previous studies and compared performances with a new model issued from machine learning. We also drew out new evidence, as we tested the effects of multiple parameters on each outcome variables.

Overall, the artificial neural network model presented the best accuracy (i.e., balance between type I and type II error rates) and power in comparison with the other analytical methods, which replicates Lanovaz et al. (2020) who had applied machine learning to AB designs. Trained on datasets presenting no autocorrelations and no trend, the model performed better with these characteristics. Therefore, power was low with large trend values (positive and negative trend greater than 0.2) and type I error rates were unsatisfyingly high with negative and low positive autocorrelations. However, the model's main strength is its stability through most tested conditions in terms of all outcome variables. The ANN and the RD presented similar results in general, but the artificial neural network model outperforms it with larger numbers of points and smaller effect sizes.

The ALIV with the randomization test showed strength by controlling over type I error rates, despite varying parameters values. However, power and accuracy varied positively with autocorrelation, number of points, and effect sizes, replicating findings from Manolov (2019). Adding to the existing evidence, our results indicated that outcomes for the ALIV along with the randomization test was not affected by the presence of trends in the data. Nevertheless, this method

presents issues with power when detecting smaller effect sizes and negative autocorrelation values. Power for graphs of 10 points only reached 80% with effect sizes of 2.5 and greater. Albeit respectable statistical power with larger effect sizes, results remained lower than all other methods.

Overall, the VSC presented similar results to the ALIV with the randomization test, which is not surprising considering they both used interpolated data points. The VSC's strength also relied on its control over type I error rates as results were adequate and stable across analyses. The VSC's power and accuracy were associated with the autocorrelation, the number of points, and the effect sizes, which is consistent with Manolov (2019) and Lanovaz et al. (2019). However, both research teams tested performances with datasets showing no trend. Therefore, our results contribute to the validation of the VSC by demonstrating no effects of the trend on statistical power. Finally, the VSC's power reached 80% when detecting an SMD of 2 (with 10 data points), which is consistent with Lanovaz et al. (2019) but inconsistent with Manolov (2019). Considering that the trend was null in both studies, this divergence may have been because of the way autocorrelation values were set.

Finally, our results regarding the RD's power in detecting effect sizes of 2 and larger were consistent with Carlin and Costello (2018). This conclusion validates the use of the RD with alternating-treatment designs. Moreover, the current study adds further evidence by testing performances in more conditions. The RD detected smaller effect sizes with more power than the two previous methods. However, the statistical power and type I error rates were affected by trend values: the RD did not perform as well with larger trends, but type I error rates nonetheless remained acceptable. The RD limitations include unacceptably high type I error rates with negative autocorrelations. Surprisingly and unlike other methods, the RD power was negatively associated with the total number of points, presenting low power with graphs of 14 points and more.

Our study compared multiple analytical methods for alternating-treatment designs, including a model derived from a machine-learning algorithm. Current findings provide a further understanding of each method and contribute to their validation. Our analyses drew out the analytical strengths and limitations of each method. Essentially, the VSC and the ALIV adequately controls type I error rates, but its power to detect smaller effect sizes is extremely low. On the other hand, the RD detects small effect sizes better but performs less well with graphs presenting a lot of points. For the most part, results indicated that the artificial neural network model was the most accurate and the most stable across conditions. Nevertheless, the ANN presents certain limitations including low power with large trends and high type I error rates with negative autocorrelations. By having a better knowledge of each method's analytical strengths and weaknesses, researchers or practitioners may choose the best method for their data.

This current study presents limitations that readers should consider. First, this experimental study only used simulated data. Simulated data are beneficial when comparing multiple methods and testing the effects of specific parameters (Morris et al., 2018). However, this type of data restricts the generalizability of our results because the data generation model makes datasets that share similarities. To replicate and extend our findings, future studies should use un-simulated alternating-treatment designs, as Lanovaz et al. (2020) have done with AB designs. Secondly, we tested the effects of a general linear trend. Considering that non-linear trends are present in approximately 87% of alternating designs (Manolov & Onghena, 2017), future studies should explore these analytical method's performances in the presence of non-linear trends. Lastly, when testing the effects of the number of points, the autocorrelation, and the trend, we simulated data with an effect size of 1. Considering that each method's performance is lower with smaller effect sizes, our results are not directly comparable to previous research (using effect sizes of 2).

References

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2013). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25(1)*, 6-19. https://doi.org/10.1080/10503307.2013.817696

Carlin, M. T., & Costello, M. S. (2018). Development of a distance-based effect size metric for single-case research: Ratio of distances. *Behavior Therapy, 49*, 981-994. https://doi.org/10.1016/j.beth.2018.02.005

Costello, M. S., Sheibanee, B. D., Ricketts, A., Hirsh, J. L., & Deochand, N. (2019). Exploration of social reinforcement for gambling in single case designs. *Analysis of Gambling Behavior, 12(1)*, 1-21. https://repository.stcloudstate.edu/agb/vol12/iss1/1

Deochand, N., Costello, M. S., & Fuqua, R. W. (2019). Real-time contingent feedback to enhance punching performance. *The Psychological Record, 70(1),* 33-45. https://doi.org/10.1007/s40732-019-00357-2

Heaton, J. (2015). *Artificial intelligence for humans, volume 3: Deep learning and neural networks.* Heaton Research Inc.

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy, 117*, 3-17. https://doi.org/10.1016/j.brat.2018.11.015

Krasny-Pacini, A., & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine, 61(3)*, 164-179. https://doi.org/10.1016/j.rehab.2017.12.002

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation.* What Works Clearinghouse. http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15(2),* 124-144. https://doi.org/10.1037/a0017736

Lane, J. D., & Gast, D. L. (2013). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24(3-4), 445-463.* https://doi.org/10.1080/09602011.2013.815636

Lanovaz, M. J., Cardinal, P., & Francis, M. (2019). Using a visual structured criterion for the analysis of alternating-treatment designs. *Behavior Modification, 43(1)*, 115-131. https://doi.org/10.1177/0145445517739278

Lanovaz, M. J., Giannakakos, A. R., & Destras, O. (2020). Machine learning to analyze single-case data: A proof of concept. *Perspectives on Behavior Science, 43,* 21-38. https://doi.org/10.1007/s40614-020-00244-0

Lanovaz, M. J., & Turgeon, S. (2020). How many tiers do we need? Type I errors and power in multiple baseline designs. *Perspectives on Behavior Science.* Advance online publication. https://doi.org/10.1007/s40614-020-00263-x

Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities, 124(1),* 35-56. https://doi.org/10.1352/1944-7558-124.1.35

Ledford, J. R., Lane, J. D., & Severini, K. E. (2017). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment, 19(1),* 4-17. https://doi.org/10.1017/BrImp.2017.16

Levin, J. R., Lall, V. F., & Kratochwill, T. R. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology, 49(1)*, 55-79. https://doi.org/10.1016/j.jsp.2010.09.002

Lillie, E. O., Patay, B., Diamant, J., Issel, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine, 8(2),* 161-173. https://doi.org/10.2217/pme.11.7

Manolov, R., & Onghena, P. (2017). Analyzing data from single-case alternating treatments designs. *Psychological Methods, 23(3),* 480-504. https://doi.org/10.1037/met0000133

Manolov, R. (2019). A simulation study on two analytical techniques for alternating treatments designs. *Behavior Modification, 43(4)*, 544-563. https://doi.org/10.1177/0145445518777875

Morris, T. P., White, I, R., & Crowther, M. J. (2018). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38(11)*, 2074-2102. https://doi.org/10.1002/sim.8086

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification, 39(4),* 510-541. https://doi.org/10.1177/0145445515581327

Park, Y.-S., & Lek, S. (2016). Artificial neural networks: Multilayer perceptron for ecological modeling. In S. E. Jørgensen (Ed.), *Ecological Model Types* (pp. 123-140). Elsevier.

Ruder, S. (2016). *An overview of gradient descent optimization algorithms*. PsyArXiv. https://arxiv.org/pdf/1609.04747.pdf

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971-980. https://doi.org/10.3758/s13428-011-0111-y

Youn, S. J., Kraus, D. R., & Castonguay, L. G. (2012). The treatment outcome package: Facilitating practice and clinically relevant research. *Psychotherapy, 49(2),* 115-122. https://doi.org/10.1037/a0027932

Appendices

Table 2.1

Example of a Table Used to Compute the Ratio of Distances.

		Conditions	A				B			
		Indices	0	1	2	3	0	1	2	3
Conditions	Indices	Values	3	4	1	2	-1	-3	-2	0
A	0	3	0	**1**	-2	-1	-4	-6	-5	-3
	1	4		0	**-3**	-2	-5	-7	-6	-4
	2	1			0	**1**	-2	-4	-3	-1
	3	2				0	-3	-5	-4	-2
B	0	-1	4	5	2	3	0	**-2**	-1	1
	1	-3	6	7	4	5		0	**1**	3
	2	-2	5	6	3	4			0	**2**
	3	0	3	4	1	2				0

Note. Bold numbers designate the numbers used to compute the slope-correction factor. The upper left and bottom right panels represent within-phase distances. The bottom left and upper right panels represent between-phase distances.

Table 2.2

Parameters Values set by the Type of Analysis

	Parameters			
Analysis	*a*	*t*	Number of points	*SMD*
Autocorrelation	[-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6]	0	10	[0, 1]
Number of points	0.2	0	[10, 12, 14, 16, 18, 20]	[0, 1]
Trend	0.2	[-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3]	10	[0, 1]
SMD	0.2	0	10	[0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4]

Note. a = autocorrelation. *t* = trend. *SMD* = standardized mean deviation. n = number of points.

Table 2.3

Accuracy, Power, Type I Error Rates for all Methods on Data Series with Variable Numbers of Points.

	n	10	12	14	16	18	20
RD	Acc	.70	.69	.68	.68	.66	.66
	Power	.45	.42	.39	.38	.34	.32
	α	.06	.04	.03	.02	.01	.01
ALIV	Acc	.60	.67	.70	.74	.76	.78
	Power	.24	.40	.45	.52	.56	.62
	α	.03	.05	.05	.05	.05	.05
VSC	Acc	.63	.68	.66	.71	.68	.73
	Power	.29	.43	.35	.47	.39	.49
	α	.04	.06	.03	.05	.03	.04
ANN	Acc	.71	.72	.73	.74	.75	.77
	Power	.47	.49	.49	.52	.53	.56
	α	.06	.05	.04	.03	.03	.03

Note. n = number of points.

Table 2.4

Accuracy, Power, Type I Error Rates for all Methods on Data Series with Variable Autocorrelation Values.

	a	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5	0.6
RD	Acc	.68	.69	.69	.69	.70	.70	.70	.70	.70	.69
	Power	.42	.45	.45	.45	.45	.44	.44	.43	.42	.40
	α	.07	.07	.07	.06	.05	.04	.04	.03	.03	.02
ALIV	Acc	.57	.59	.60	.60	.62	.63	.64	.65	.66	.67
	Power	.18	.21	.23	.24	.27	.28	.31	.33	.36	.38
	α	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03
VSC	Acc	.59	.61	.62	.63	.64	.65	.66	.67	.68	.69
	Power	.23	.27	.28	.30	.32	.34	.36	.38	.40	.42
	α	.04	.04	.04	.04	.04	.04	.03	.03	.03	.04
ANN	Acc	.68	.70	.70	.70	.71	.71	.71	.71	.71	.70
	Power	.44	.47	.46	.47	.47	.46	.46	.46	.45	.43
	α	.08	.08	.07	.07	.06	.05	.04	.04	.03	.03

Table 2.5

Accuracy, Power, Type I Error Rates for all Methods on Data Series with Variable Trend Values.

	Trend	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
RD	Acc	.61	.65	.69	.70	.69	.65	.61
	Power	.23	.33	.41	.45	.42	.33	.23
	α	.01	.03	.04	.05	.03	.03	.01
ALIV	Acc	.63	.63	.63	.62	.63	.63	.62
	Power	.28	.29	.28	.28	.29	.29	.28
	α	.03	.03	.03	.03	.03	.03	.03
VSC	Acc	.65	.65	.65	.64	.65	.65	.65
	Power	.33	.33	.33	.33	.33	.33	.33
	α	.04	.04	.03	.04	.04	.04	.03
ANN	Acc	.63	.67	.70	.71	.70	.66	.62
	Power	.27	.36	.44	.47	.44	.35	.25
	α	.02	.03	.04	.05	.04	.03	.02

Table 2.6

Statistical Powers for all Methods on Data Series with Variable Effect Sizes.

SMD	1	1.5	2	2.5	3	3.5	4
RD	.45	.74	.92	.98	1.00	1.00	1.00
ALIV	.28	.52	.72	.87	.95	.98	.99
VSC	.34	.59	.80	.92	.98	.99	1.00
ANN	.47	.76	.93	.99	1.00	1.00	1.00

Figure 2.1

Our Neural Network with Eight Features, One Hidden Layer with Four Neurons and One Class Output.

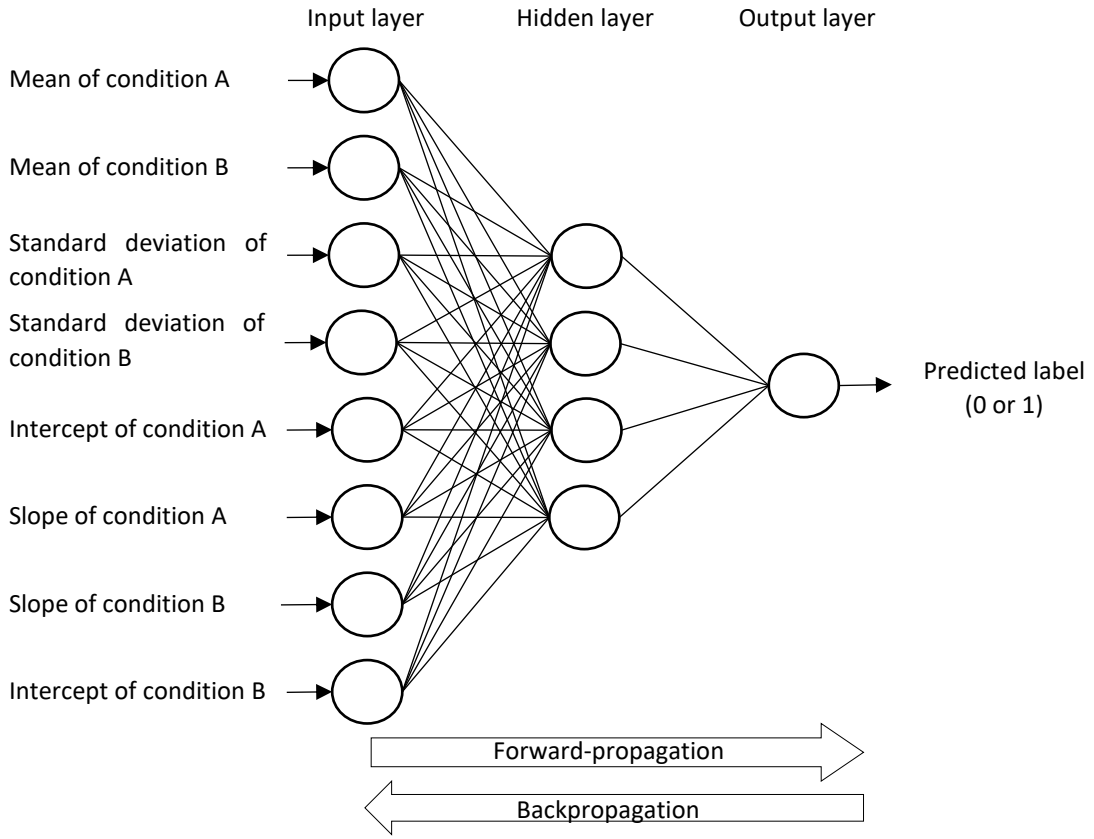
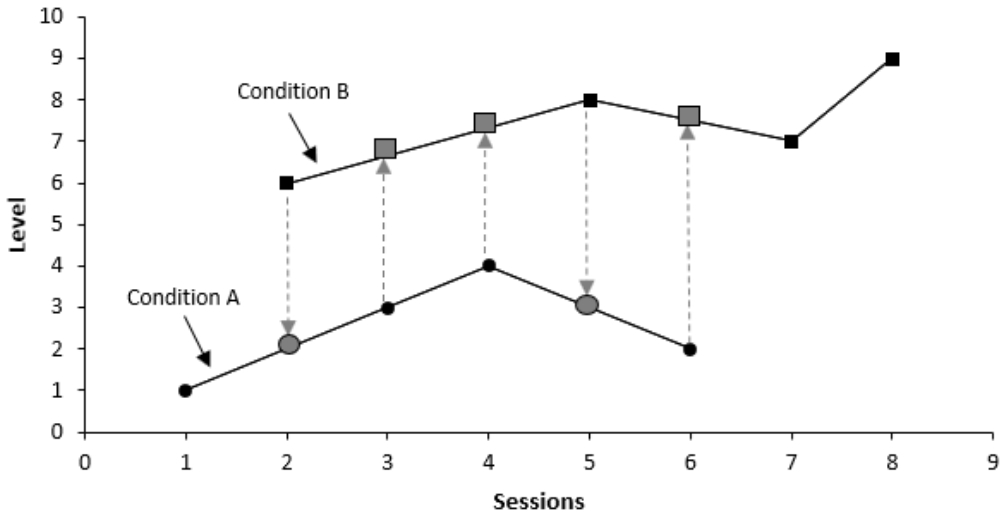


Figure 2.2

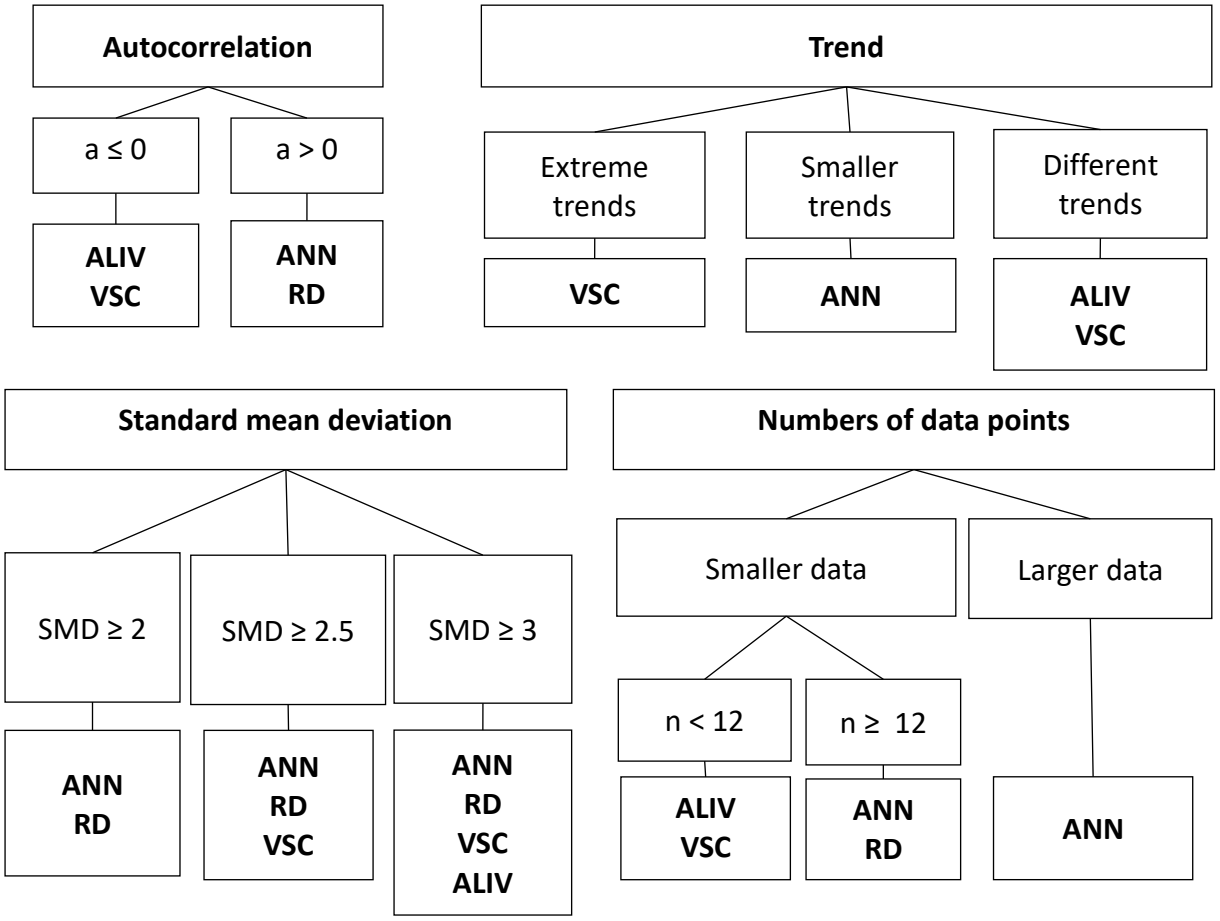
Example of a Graph Analyzed with the Actual and Linearly Interpolated Values or the Visual Structured Criterion.



Note. Black shapes represent actual points while outlined gray shapes represent linearly interpolated points.

Figure 2.3

Proposed Decisional Algorithm for Choosing an Analytical Method Based on Datasets Characteristics.



Note. This algorithm only provides general guidelines in the choice of an analytical method based on datasets characteristics. For details in terms of accuracy, power and type I error rate, refer to the results section.

Discussion générale

Nous pouvons établir des liens entre le sujet de cette étude et la pratique psychoéducative. L'utilisation de protocoles à cas uniques en pratique est une approche rigoureuse qui permet une collecte de données et de suivi des effets dans le cadre de l'évaluation post-intervention. La cueillette de données est une compétence centrale en psychoéducation parce qu'elle s'insère dans plusieurs opérations professionnelles. Lorsqu'il est en contact direct, le psychoéducateur est constamment en observation : il réalise sa cueillette de données durant l'évaluation psychoéducative (phase de niveau de base), mais il la poursuit également durant la mise en œuvre de son intervention (phase d'intervention). Certes, une évaluation continue est cruciale pour s'assurer de l'efficacité de l'intervention et pour surveiller l'évolution des capacités adaptatives du client (Gendreau et collaborateurs, 2001). De plus, l'utilisation des protocoles à cas uniques se présente comme un outil essentiel dans le cadre d'une pratique basée sur des données probantes. Le professionnel utilise effectivement ses observations pour prendre des décisions cliniques significatives dans la vie de ses clients. Ainsi, il est essentiel de disposer de méthodes rigoureuses, précises et présentant le moins d'erreurs possibles. Bien que les protocoles à cas uniques soient principalement utilisés pour des comportements facilement observables, ces protocoles peuvent également être utilisés pour des comportements plus complexes ou moins observables si la clientèle le permet. Par exemple, un adolescent ou un adulte fiable pourrait remplir un journal de bord où il documenterait l'occurrence de ses obsessions, ses pulsions d'automutilation ou ses hallucinations. Le psychoéducateur pourrait également demander à un parent de documenter les comportements de son enfant d'une rencontre à l'autre.

Dans cet ordre d'idées, cette étude a exploré et comparé un modèle issu des réseaux de neurones artificiels avec diverses méthodes d'analyses des protocoles avec alternance de

traitements. Au-delà de la performance générale de chacune de ces approches, notre étude a examiné les effets de différents paramètres sur différentes caractéristiques de chaque méthode. Nous avons manipulé les paramètres afin de simuler des caractéristiques habituellement observables dans des données cliniques réelles. Lorsqu'un intervenant collecte des données d'un même comportement auprès d'une même personne, ces mesures présentent par défaut une certaine autocorrélation. De plus, les clients ne sont pas uniquement influencés par les services d'intervention puisqu'ils sont également sous l'influence de plusieurs autres aspects de leur vie. Pour cette raison, une valeur de tendance générale s'ajoute souvent sur les mesures collectées. Par ailleurs, la taille d'effet à l'issu du changement varie selon le type de comportement mesuré et le type d'intervention mise en œuvre. Enfin, le nombre de mesures qu'un intervenant pourra collecter diffère selon plusieurs facteurs (ex. : fréquence du comportement, disponibilité des acteurs, établissement déployant les services). Lorsque nous évaluons des méthodes qui seront utilisées en intervention, il est primordial de comprendre ainsi leurs forces et limites de façon détaillée.

Tout d'abord, nous avons analysé toutes ces approches avec une attention particulière aux erreurs de type I. Les psychoéducateurs travaillent avec des comportements pouvant avoir un impact significatif sur l'intégrité du client et de ses proches. Donc, conclure qu'une intervention a été efficace alors qu'elle ne l'était pas (faux positif) peut avoir des répercussions graves, particulièrement si la cible était un comportement dangereux (ex. : comportements d'automutilation, idées suicidaires ou homicidaires). Dans cette perspective, nos résultats ont démontré que l'ALIV et le VSC contrôlaient le mieux leurs erreurs de type I à travers les conditions examinés. Bien que les faux positifs aient possiblement des répercussions plus graves, les faux négatifs amènent également des enjeux en termes de ressources (temporel et monétaire) et de motivation pour le client. Ainsi, l'idéal serait un équilibre entre les erreurs de types I et de type II. Les faux négatifs sont représentés dans nos résultats sous la forme de puissance statistique. Dans

cette perspective, ce sont le RD et le réseau de neurones artificiels qui présentent les meilleures puissances statistiques et précisions (équilibre entre les erreurs de type I et type II). Néanmoins, si nous devons considérer toutes les conditions analysées ensembles (nombre de points, autocorrélation, tendance et taille d'effet), nos résultats démontrent que le modèle de réseaux de neurones artificiels est l'approche qui produit les résultats les plus élevés et les plus stables. Le modèle présente incontestablement ses propres limites, mais il apparaît comme la méthode qui s'adapte le mieux à tous les types de données. Sans contredit, le jugement clinique des psychoéducateurs demeure nécessaire pour transposer les conclusions statistiques en prises de décisions cliniques. Les résultats statistiques servent de soutien dans la prise de décision, mais ils doivent être accompagnés d'un jugement clinique qui permet d'évaluer les données sous un angle global.

Les résultats de notre étude présentent alors aux lecteurs quatre méthodes efficaces permettant d'analyser les protocoles avec alternance de traitements. Au-delà de leurs forces et limites analytiques, le psychoéducateur s'intéresse toutefois à la facilité d'application des méthodes d'analyse. Alors que toutes les méthodes présentent des calculs assez complexes, le VSC est la méthode la plus simple à appliquer pour un intervenant n'ayant pas d'expertise en analyse statistique. La méthode du VSC implique effectivement une analyse visuelle ainsi que le traçage de critères visuels simples (trajectoires). De plus, le VSC ne nécessite pas de calculs importants puisque le professionnel peut se référer aux seuils proposés par Lanovaz et al. (2019). Pour utiliser le VSC, le psychoéducateur doit faire le suivi des effets en mettant en place un protocole avec alternance de traitements. Une fois les données obtenues dans un graphique, le psychoéducateur trace des lignes (trajectoires) entre les points d'une même condition. Chaque temps de mesure où se trouve à la fois un point et une trajectoire représente un temps de comparaison. Pour un comportement dont on cherche à augmenter l'occurrence, le psychoéducateur compte le nombre

de fois que la condition B (intervention) se retrouve au-dessus de la condition A (niveau de base). Il peut ensuite se référer aux seuils prédéterminés par Lanovaz et al. (2019) afin de conclure si le changement est significatif ou non. Pour un psychoéducateur qui effectue déjà le suivi des effets de ses interventions, cette analyse nécessite peu de temps additionnel et aucun calcul.

En outre, une adaptation du VSC sous forme de calculateur en ligne pourrait rendre cette analyse encore plus facile d'usage, tel que Lanovaz et al. (2020) a élaboré pour les protocoles AB. Similairement, Carlin et Costello (2018) ont créé un document convivial, utilisant le logiciel connu Microsoft Excel, pour facilement calculer le RD. Dans les deux cas, le professionnel n'a qu'à entrer les valeurs dans les cases indiquées puis le calcul se fait automatiquement. À l'heure actuelle, l'ALIV et le modèle du réseau de neurones artificiels ne sont pas aussi facilement accessibles et adaptés pour le public général. Néanmoins, le concept de vulgarisation des méthodes d'analyse, tel qu'entamé par Carlin et Costello (2018) et Lanovaz et al. (2020), dégage une opportunité intéressante d'arrimer la pratique psychoéducative et les analyses statistiques rigoureuses issues de la recherche.

Limites de l'étude et recherches futures

Notre étude comporte des limites méthodologiques qui devraient être tenues en compte lors de l'interprétation de nos résultats. Premièrement, nous avons uniquement utilisé des données simulées pour réaliser nos analyses. L'utilisation de données simulées a ses propres avantages, considérant qu'elle permet la manipulation des paramètres et ainsi l'analyse de leurs effets. Étant toutefois générées avec un même modèle, toutes nos données présentent des similitudes qui empêchent nos résultats d'être complètement généralisables à de réelles données. Dans ce sens, les recherches futures devraient également inclure des données non-simulées afin de tester la généralisation de nos résultats. Lanovaz et al. (2020) a effectivement démontré la possibilité

d'inclure des graphiques non-simulées de protocoles à cas uniques dans son étude. Deuxièmement, notre étude a seulement exploré les effets d'une tendance linéaire générale sur nos résultats. En effet, notre paramètre ne représente pas complètement les tendances observées dans les protocoles avec alternance de traitements : Manolov et Onghena (2017) ont noté autant de protocoles avec des tendances non-linéaires que ceux avec des tendances linéaires (87%). Les recherches futures devraient alors explorer ce type de tendance et les effets qu'il pourrait avoir sur la performance des méthodes d'analyse. Enfin, nous avons généré des données avec une taille d'effet de 1 pour nos analyses. Sachant que toutes les méthodes performent mieux avec de grandes tailles d'effets, nos résultats en termes de puissance statistique sont faibles et directement incomparables aux résultats des recherches précédentes.

Conclusion

Notre étude avait pour premier objectif de reproduire et approfondir les connaissances sur trois nouvelles méthodes d'analyse de protocole avec alternance de traitements. À cet égard, nous avons examiné la puissance statistique, les erreurs de type I et la précision de ces méthodes sous différentes conditions. Nos résultats offrent ainsi un aperçu détaillé des forces et limites de chacune de ces approches en fonction de la taille du protocole, de l'autocorrélation, de la tendance et de la taille d'effet. Notre étude avait pour deuxième objectif de comparer chacune de ces méthodes avec un modèle estimé par l'algorithme du réseau de neurones artificiels. De manière générale, notre étude a indiqué que le modèle d'apprentissage automatique produisait des résultats plus élevés et plus stables que ceux des autres méthodes analysées. Les recherches futures doivent néanmoins examiner ces résultats avec des données non-simulées. Notre recherche explore ainsi une dimension centrale de la psychoéducation, la cueillette de données et l'évaluation continue, bien

que moins étudiée. Notre étude met alors à disposition quatre méthodes efficaces d'analyse de données pouvant être appliquées en pratique psychoéducative ainsi qu'en recherche clinique.