Predicting Geographical Human Risk of West Nile Virus – Saskatchewan, 2003 and 2007

Tasha Y. Epp, DVM, PhD,[1] Cheryl L. Waldner, DVM, PhD,[1] Olaf Berke, PhD[2]

ABSTRACT

Objectives: To detail the use of a model to predict areas of low, medium, and high risk of West Nile virus (WNV) in humans in both 2003 and 2007 in the province of Saskatchewan. To identify consistent high-risk areas from year to year as well as important environmental variables within those high-risk areas.

Methods: The number of laboratory-confirmed WNV individuals was obtained from Saskatchewan Health by rural municipality. The population at risk was obtained from Statistics Canada by rural municipality. Climate and habitat variables were incorporated into a discriminant analysis model with the production of risk maps as an end product.

Results: The discriminant analysis models had testing classification accuracies of 67% in 2003 and 44% in 2007. Climate and habitat variables remained important in all models while some habitat variables were less important in 2007. Risk maps from historically trained 2007 model revealed a southwest to northeast decreasing trend of risk.

Conclusion: The models could be useful for indicating areas of high risk on a year-to-year basis or based on historical data. High-risk regions are characterized by less rainfall in June and July followed by higher temperatures in July and August with less vegetation and water coverage than low-risk regions.

Key words: Zoonoses; arboviruses; Saskatchewan; public health

La traduction du résumé se trouve à la fin de l'article.

Can J Public Health 2009;100(5):344-48.

The introduction of West Nile virus (WNV) into North America sparked an interest in predicting where and when the virus would appear.[1] Predictive risk mapping is a process by which components of the disease cycle are used to create models and subsequent risk maps.[2,3] The methods have become more practical for a broader range of diseases and study locations because remote sensing can now provide environmental information at required spatial and temporal resolution.[4,5]

Vector-borne diseases, such as WNV, are particularly amenable to spatial and temporal analysis as they are highly influenced by regular, seasonal climate, and environmental changes.[3,6-8] During the season, mosquitoes become infected with the West Nile virus primarily through bird-blood meals and then retransmit the virus to any one of multiple bird species, a cycle which amplifies the virus. Governed by environmental conditions and host behaviours, infected mosquitoes can spread WNV to other incidental hosts, such as humans and horses.

Defining the risk of WNV infection is a key component to public health intervention strategies.[9] Prioritization of vector-borne disease programs in the overall public health budget is a juggling act, affected by limited funding availability. In Saskatchewan, interventions are prioritized largely based on environmental conditions conducive to mosquito development and surveillance for clinical disease in humans. Health officials could increase the cost effectiveness of control and surveillance programs with a method of predicting differences in regional risk of infection.

The primary objective of this study was to describe the application of a previously established model to predict areas of low, medium, and high risk of WNV in humans in both 2003 and 2007 in the province of Saskatchewan.[10] The second objective was to use historical surveillance data from 2003-2005 to make predictions of areas of risk of WNV in humans in 2007.

MATERIALS AND METHODS

WNV infection risk – 2003 and 2007

Human surveillance data were obtained from Saskatchewan Health as the number of laboratory-confirmed WNV individuals (which included WN fever, WN neurological syndrome and asymptomatic individuals, http://www.health.gov.sk.ca/wnv-surveillance-results-archive (accessed July 14, 2009)) per rural municipality (RM). In 2003, each RM with WNV individuals (sampled RM) was classified by category of WNV infection risk using the 25th and 75th percentiles: low-risk (0.0-0.09%), medium-risk (>0.09%-0.41%), and high-risk (>0.41%). This classification was repeated in 2007, with the following results: low-risk (0-0.14%), medium-risk (>0.14%-0.36%), and high-risk (>0.36%). The population at risk was deter-

Author Affiliations

1. Large Animal Clinical Sciences, Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK
2. Department of Population Health, Ontario Veterinary College, University of Guelph, Guelph, ON

Correspondence and reprint requests: Tasha Epp, Large Animal Clinical Sciences, WCVM, University of Saskatchewan, 52 Campus Drive, Saskatoon, SK S7N 5B4, Tel: 306-966-6542, Fax: 306-966-7159, E-mail: tasha.epp@usask.ca

Table 1. Results from Final 2003, 2007 and Historically Trained 2007 Multivariable Models

		2003	2007	Historically Trained 2007
Model		148 sampled RMs	193 Sampled RMs	
RMs* sampled	Training	118 sampled RMs (observed categorization: 28 low-risk, 62 medium-risk, 28 high-risk)	154 sampled RMs (observed categorization: 35 low-risk, 77 medium-risk, 42 high-risk)	72 RMs with historical data (observed categorization: 22 low-risk, 32 medium-risk, 18 high-risk)
	Testing	30 sampled RMs (observed categorization: 9 low-risk, 11 medium-risk, 10 high-risk)	39 sampled RMs (observed categorization: 12 low-risk, 15 medium-risk, 12 high-risk)	193 sampled RMs (observed categorization: 47 low-risk, 92 medium-risk, 54 high-risk)
		Plus remaining 150 unsampled RMs	Plus remaining 105 unsampled RMs	Plus remaining 105 unsampled RMs
Model Accuracy	Training	67%	61%	100%
	Testing	60%	44%	45%
Significant Variables†	Function 1	Mean LST‡	NDVI	Tree coverage
		NDVI‡	Mean LST	NDVI
		Precipitation	Precipitation	Mean LST
		Tree coverage	Tree coverage	Precipitation
	Function 2	Water coverage	Water coverage§	Water coverage
		Wetland coverage	Wetland coverage§	Wetland coverage
				Precipitation
Eigenvalues	Function 1	0.544	0.446	7.55
	Function 2	0.203	0.020	2.495
Group membership probability	Low risk	85%	76%	91%
	Medium risk	67%	57%	91%
	High risk	74%	63%	96%

* RMs = rural municipalities
† Variables that were significant in the model are recorded by function in the order of importance for contributing to the function.
‡ LST = land surface temperature, NDVI = normalized difference vegetation index
§ Function not statistically significant but retained for increased model accuracy

mined for 2003 and 2007 using Statistics Canada 2001 and 2006 census data by rural municipality, respectively.

Environmental variables

Variables used in the analysis were the same as those identified in a previous study regarding WNV infection in horses in 2003.[10] Those variables that had multiple statistically significant time periods were condensed into one or two principal components with principal component analysis before use in the final models (SPSS 14.0, SPSS Inc., Chicago, IL, USA).[10,11]

Land Surface Temperature

Land Surface Temperature (LST) images (Moderate Resolution Imaging Spectrometer satellite (MODIS); Earth Observing System Gateway, National Aeronautics and Space Administration; http://lpdaac.usgs.gov) were provided as 8-day composites (1 kilometre resolution) beginning May 1st and ending September 13th for 2003 and 2007. The images were joined together and clipped to show only the province of Saskatchewan (PCI Geomatica 9, PCI Geomatics, Richmond, ON, Canada). The images included daytime (maximum) and nighttime (minimum) temperatures and were manipulated to give a mean LST. For each year, the mean LST averaged for each RM was calculated for each 8-day composite.

Precipitation

Precipitation values (mm) were obtained for 2003 and 2007 on a daily basis from Environment Canada. Eight-day composites (total precipitation per time period) were created to match the remotely sensed time periods. Interpolation among the 176 stations in the province was accomplished using Inverse Distance Weighted (IDW) method (ArcGIS 9.2, ESRI Inc., Redlands, CA, USA).[12] For each year, the averaged total value for each time period by RM was calculated.

Vegetation

Normalized Difference Vegetation Index (NDVI) (MODIS satellite; http://lpdaac.usgs.gov) is a simple index of vegetation cover which allows monitoring of seasonal changes in vegetation growth.[13] Images (500 metre resolution) were provided as 16-day composites starting April 23rd and ending September 13th for both 2003 and 2007. For each year, the average value per RM was calculated.

Land Cover

North Digital and South Digital Land Cover dataset based on satellite imagery from 2000 for the province of Saskatchewan was obtained from Information Services Corporation of Saskatchewan. Classifications were further aggregated to make a manageable number of categories for analysis. Those categories selected for consideration in the models included: water, wetland (which includes bog, marsh, fen, etc.), and treed (which includes pine, spruce, hardwood, softwood, etc.). The percentage of RM covered by each of the categories was calculated.

Statistical analysis

Overview

Discriminant analysis (SPSS 14.0, SPSS Inc., Chicago, IL, USA) was used to predict membership in the three mutually exclusive groups (low, medium and high risk).[14] The yearly datasets were divided into a) a training dataset (consisting of a random selection of RMs with data) and b) a testing dataset (consisting of the remaining RMs with data and any RMs without data) (Table 1). The training data were used to analyze the known differences between RMs with data; subsequently, these differences were then applied to the testing data to assess the accuracy of the predictions of remaining RMs with and without data.

Multivariable model selection was partially determined through overall classification or prediction accuracy percentage for both training and testing datasets. This was defined as the proportion of

Figure 1. Depiction of predicted group membership (low, medium or high risk of infection) for 2003 human dataset by RM, with indication of RMs with 75% or greater probability of group membership

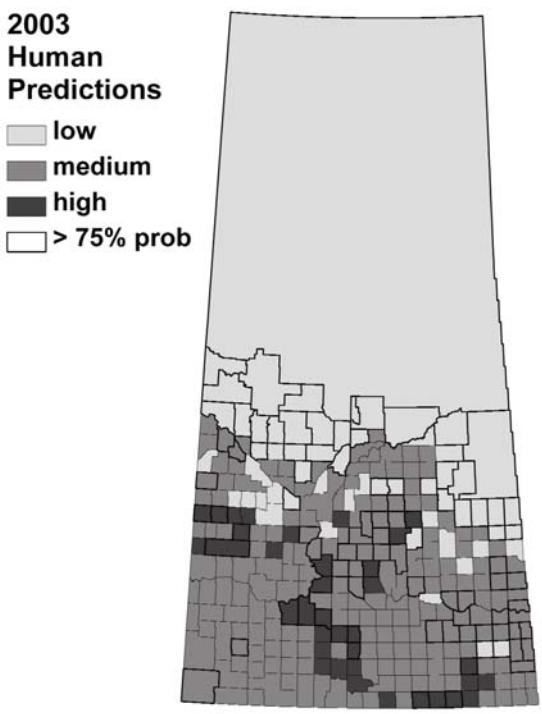
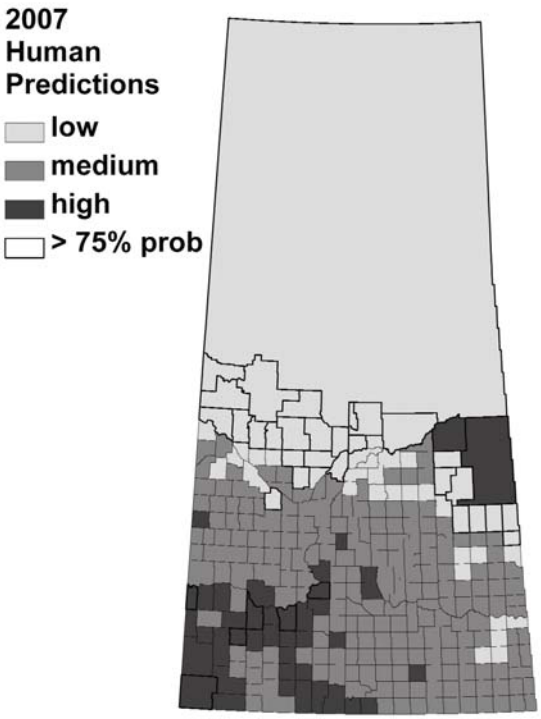


Figure 2. Depiction of predicted group membership (low, medium or high risk of infection) for 2007 human dataset by RM, with indication of RMs with 75% or greater probability of group membership



RMs with data correctly classified based on the observed risk category (pre-model classification based on proportion data) compared to the predicted risk category (post-model classification).[14] In addition, multivariable models were fit with an unequal weighting scheme to adjust the posterior probabilities to account for prior knowledge of observed group membership.[14] Separate matrices (to account for unequal group covariance matrices) were used when Box's M test was significant (p=0.05) and the prediction accuracy percentage changed substantially from a model that used a common matrix for all groups.[14] Ultimately, the final model was the one that produced the best overall classification with the least overlap of risk categorization between high- and low-risk areas.

The results of the discriminant analysis were twofold: a) providing functions or sets of variables by which the risk categories were discriminated by, how well each of these functions discriminated and which variables within the functions were most informative, and b) providing a set of three probabilities predicting the likeliness of membership in each of the three risk categories.[14] Individual RMs were classified (based on the functions) into one of the risk categories by predicting the group (low-, medium- or high-risk) to which the individual RM most likely belonged.[14] The categorization rule is less reliable for RMs with maximum probability of <75%. Therefore, maximizing the overall probability of group membership for all RMs in each group was used in final model selection. Chloropleth maps of the predicted risk categories were generated using ArcGIS 9.2.

Yearly Models (2003, 2007)

Yearly models were based on WNV infection risk and environmental variables by RM from within each year (2003 and 2007). Comparison of the 2003 and 2007 yearly models was done with the kappa statistic.

Historical Prediction Model

Information from modeling of horse and human surveillance data conducted in Saskatchewan in 2003-2005 was used to create a historical training dataset.[10] Selection of RMs (n=72) with suitable data was based on consistent predictions from previously established models where at least 2 of the predictions had probabilities of group membership of 75% or higher. The historical training dataset was used to train the modeling of the 2007 human dataset. Comparison of the 2007 yearly and historically trained models was done with the kappa statistic.

RESULTS

Predictive ability – 2003 human dataset

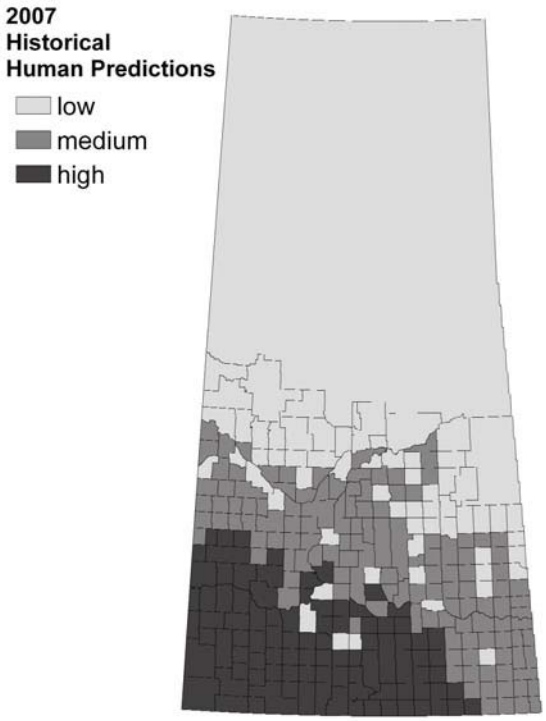
The numbers of predicted RMs in the three risk categories were: 59 in the low-risk group, 203 in the medium-risk group, and 36 in the high-risk group (Figure 1). The model used two functions to predict RM category (Table 1).

Predictive ability – 2007 human dataset

The numbers of predicted RMs in the three risk categories were: 59 in the low-risk group, 198 in the medium-risk group, and 41 in the high-risk group (Figure 2). The model used one function to predict RM category (Table 1). The second function was not statistically important in the prediction of RMs but was retained to maximize model accuracy.

The agreement in the classification result for individual RMs when compared between the 2007 and 2003 models was poor; kappa was 0.10 for high-risk RMs and 0.62 for low-risk RMs.

Figure 3. Depiction of predicted group membership (low, medium or high risk of infection) for historical training dataset model of 2007 human dataset by RM



Predictive ability – historical training data on 2007 human dataset

The number of predicted RMs in the three risk categories was: 65 in the low-risk group, 136 in the medium-risk group, and 97 in the high-risk group (Figure 3). The model used two functions to predict RM category (Table 1). Both functions incorporated precipitation information; function 1 used June values while function 2 used July values.

Comparison of the historically trained 2007 model with the original 2007 model revealed that both models classified the following number of RMs the same: 43 low-risk RMs, 118 medium-risk RMs and 33 high-risk RMs. The agreement, as calculated using kappa statistic, was 0.40 for high-risk RMs and 0.61 for low-risk RMs. The historically trained model clearly demonstrated a southwest to northeast trend of decreasing risk, which is not as clear from the other models.

DISCUSSION

The models in this study try to geographically predict which areas are at risk of infection (high, medium or low risk) by defining a set of criteria upon which to classify that risk. There appears to be a trend of high-risk areas concentrating in the south-central and south-western portions of the province. However, individual RMs did differ in their category of risk depending on the model and year, a reflection of the model's limitations. The models could have had error introduced due to training dataset selection, inaccuracies in the RM classification prior to entry into the model (i.e., location of human individual, not the location of exposure), and reliance on summarized environmental data. Predictions of individual RMs should be used with caution; instead, by applying a smoothing technique, the maps would indicate general but larger areas or trends of high risk of infection.

In the model trained by the historical dataset, model predictions compared to the original risk categories based on passive surveillance were only 45% accurate; however, the model did maintain high probabilities that the predictions for each risk group were trustworthy. The predictive map clearly demonstrated a gradient of risk decreasing from south to north which mirrors what is found by mosquito trapping programs.[8]

The environmental variables included in this analysis were based on previous models built using horse surveillance data and included precipitation, temperature, vegetation and land cover, specifically wetlands, water and treed areas.[10] The present predictive models provided only marginally accurate predictions of risk geographically. Obviously, the complexity of the cycle is not completely explained by these variables and their interactions alone. Factors such as biodiversity, predators, parasites, food availability, human behaviour and spatial resources will affect interactions between the vector and hosts, while immune status of the hosts will become more important the longer the virus remains endemic in an area.[9]

Variables contributing to the model functions were fairly consistent between models. Precipitation and temperature were important in the prediction of risk of WNV in humans, particularly in 2003; decreasing rainfall into July and higher temperatures overall were associated with high-risk areas. *Culex tarsalis* uses standing water with increased organic content for oviposition, which would be washed away by increased rainfall.[8,15] Habitat was also highly important in the prediction of human WNV risk, particularly in 2007. In the present study, vegetation index was slightly higher on average in low-risk areas as was the percentage of RMs covered in trees, water and wetland. *C. tarsalis* prefers shallow, often stagnant water of high organic content with little tree cover surrounding the sites, such as water-filled hoof prints near livestock watering sites.[8] In the 2007 model, the percentage of water and wetland coverage was not statistically important to the prediction process. This could be influenced by the fact that actual wetland capacity was much higher in 2007 than 2003 owing to the few wet years that occurred between them (personal communication, P. Curry, Saskatchewan Health, December 2007).

By specifying what time periods should be incorporated into the model-building process, it could be used as a method to make early predictions or inform public health authorities about the development of high-risk areas as the season progresses. The usefulness of the models as a predictor of high-risk areas must be coupled with the knowledge of vector abundance and host population dynamics. Historically, maps of mosquito vectors consistently indicated high-risk areas in the southeastern portion of the province. If mosquito trapping data were available, these models could validate predictions made based on mosquito information or signal areas where mosquito data were required. With further research, greater accuracy in predicting WNV risk of infection will occur.

REFERENCES

1. Rogers DJ, Myers MF, Tucker CJ, Smith PF, White DJ, Backenson B, et al. Predicting the distribution of West Nile fever in North America using satellite sensor data. *PE&RS* 2002;68:112-14.
2. Brooker S, Hay SI, Bundy DAP. Tools from ecology: Useful for evaluating infection risk models? *Trends in Parasitology* 2002;18:70-74.
3. Kitron U. Risk maps: Transmission and burden of vector-borne diseases. *Parasitology Today* 2000;16:324-25.
4. Beck LR, Lobitz BM, Wood BL. Remote sensing and human health: New sensors and new opportunities. *Emerg Infect Dis* 2000;6:217-26.

5. Rogers DJ, Randolph SE, Snow RW, Hay SI. Satellite imagery in the study and forecast of malaria. *Nature* 2002;415:710-18.
6. Hay SI, Tucker CJ, Rogers DJ, Packer MJ. Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. *Ann Trop Med Hyg* 1996;90:1-19.
7. Mellor PS, Leake CJ. Climatic and geographic influences on arboviral infections and vectors. *Rev Sci Tech Off Int Epiz* 2000;19:41-54.
8. Curry P. Saskatchewan mosquitoes and West Nile virus. *Blue Jay* 2004;62:104-11.
9. Rainham DGC. Ecological complexity and West Nile virus: Perspectives on improving public health response. *Can J Public Health* 2005;96:37-40.
10. Epp T. West Nile Virus: From Surveillance to Prediction Using Saskatchewan Horses [thesis]. Saskatoon, SK: University of Saskatchewan, 2007. Available online at: http://library2.usask.ca/theses/available/etd-08012007-160845/ (Accessed April 1, 2009).
11. Dohoo I, Martin W, Stryhn H. *Veterinary Epidemiologic Research*, 1st ed. Charlottetown, PE: AVC Inc., 2003;322.
12. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*, 1st ed. Hoboken, NJ: John Wiley & Sons, Inc., 2004;494.
13. Jensen JR. *Introductory Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2005;311-19.
14. Klecka WR. *Discriminant Analysis*. Sage University Papers, Series on Quantitative Application in the Social Sciences, 07-019. Beverly Hills, CA: Sage Publications, 1980;1-71.
15. Shaman J, Day JF. Achieving operational hydrologic monitoring of mosquito-borne disease. *Emerg Infect Dis* 2005;11:1343-50.

Received: December 4, 2008
Accepted: April 24, 2009

RÉSUMÉ

Objectifs : Expliquer l'utilisation d'un modèle de prévision des secteurs à risque modéré, moyen et élevé pour les humains de contracter le virus du Nil occidental (VNO) en 2003 et en 2007 dans la province de la Saskatchewan. Déterminer les secteurs à risque uniformément élevé d'une année à l'autre, ainsi que les variables environnementales importantes dans les secteurs à risque élevé.

Méthode : Nous avons obtenu auprès du ministère de la Santé de la Saskatchewan le nombre de cas séropositifs pour le VNO confirmés en laboratoire, par municipalité rurale. Statistique Canada nous a fourni le nombre de personnes à risque par municipalité rurale. Des variables climatiques et d'habitat ont été intégrées dans un modèle d'analyse discriminante afin de produire des cartes du risque.

Résultats : Les modèles d'analyse discriminante étaient exacts à 67 % en 2003 et à 44 % en 2007. Les variables climatiques et d'habitat sont demeurées importantes dans tous les modèles, mais certaines variables d'habitat avaient moins d'importance en 2007. Les cartes du risque produites à partir du modèle chronologique de 2007 montrent que le risque a eu tendance à diminuer en allant du Sud-Ouest vers le Nord-Est.

Conclusion : Ces modèles pourraient être utiles pour indiquer les secteurs à risque élevé d'une année à l'autre ou selon des données historiques. Les régions à risque élevé se caractérisent par une pluviosité relativement faible en juin et en juillet suivie de hausses de la température en juillet et en août, avec une couverture végétale et hydrique plus limitée que dans les régions à risque modéré.

Mots clés : zoonoses; arbovirus; Saskatchewan; santé publique

RECENSION

L'évaluation : concepts et méthodes

Astrid Brousselle, François Champagne, André-Pierre Contandriopoulos et Zulmira Hartz (Éds.), Montréal, QC : Les Presses de l'Université de Montréal, 2009; 304 pp., 49,95 $

Les auteurs ont respecté leur engagement maintes fois renouvelé, soit de publier le fruit de leurs réflexions et de leurs expériences de terrain en matière d'évaluation d'interventions. Le modèle développé depuis les années quatre-vingt par le Groupe de recherche interdisciplinaire en santé (GRIS), pour évaluer des interventions en santé, est la clef de voute du livre.

Après une analyse historique du développement de l'évaluation, les auteurs présentent leur cadre conceptuel. Ils insistent ensuite, à juste titre, sur l'importance de modéliser les interventions avant de les évaluer et décrivent la démarche à suivre pour ce faire. Dans la seconde section du livre, ils consacrent un chapitre à chacun des sept types d'évaluation de leur modèle qu'ils illustrent par des cas réels. Ils définissent les concepts, confrontent leurs définitions à celles d'autres auteurs et exposent brièvement, un peu trop selon moi, les méthodes pertinentes à l'atteinte des objectifs d'évaluation. Justement parce que l'évaluation doit faire appel à une diversité de méthodes et que le choix de celles-ci dépend d'une multitude de facteurs, le lecteur aurait besoin d'être mieux outillé. D'ailleurs, les cas mettent trop souvent l'accent sur les résultats des évaluations alors que sur le plan pédagogique, il aurait été plus utile de décrire davantage la démarche d'évaluation, les difficultés rencontrées et les décisions prises pour y faire face. En outre, les auteurs du chapitre consacré à l'analyse des effets auraient eu avantage à faire référence au livre de Shadish, Cook et Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* afin de mettre à jour les aspects méthodologiques.

La lecture de la dernière section est essentielle : elle permet de situer les paradigmes dans lesquels les travaux des évaluateurs s'inscrivent de même que les formes que l'utilisation des évaluations peut prendre et surtout, elle guide le lecteur qui désire produire des évaluations crédibles dont l'utilisation sera optimale.

Appuyé sur un nombre imposant de références, le contenu de ce livre est dense. Il s'agit d'un excellent outil pédagogique pour l'enseignement de l'évaluation en milieu universitaire, l'un des rares livres disponibles en français. Il sera également utile aux intervenants et aux chercheurs qui désireront améliorer leurs compétences dans le domaine de l'évaluation d'interventions, et ce, qu'ils travaillent dans le secteur de la santé et des services sociaux ou d'autres domaines d'activités tels que l'éducation ou l'économie sociale, le modèle du GRIS pouvant être mis à l'épreuve dans une grande diversité de contextes.

Diane Berthelette, Ph.D.
Professeure titulaire
Département d'organisation et de ressources humaines, Université du Québec à Montréal
Présidente-directrice générale, Centre de liaison sur l'intervention et la prévention psychosociales