Masters of Public Health Project (Health Stream)

Title: Public Health Program Evaluation Best Practices within the Canadian Federal Government

Student: Agata Stankiewicz

Project Supervisor: Dr. Ron Wall

March 23, 2009

TABLE OF CONTENTS

ABSTRACT

Program evaluation findings are reported in evaluation reports as part of Treasury Board of Canada Secretariat (TBS) funding requirements and are key information used to ensure accountability for planned results. This project critically appraises the Public Health Agency of Canada's (PHAC) eight program evaluation reports for their strengths and weaknesses – program evaluation planning, design and implementation, data collection and analysis, and reporting – for informing public health practice. First, these reports are appraised using a modified version of the review template obtained from the "Review of the Quality of Evaluations Across Departments and Agencies", developed by the TBS. These findings are then reviewed in light of public health program evaluation guidelines for compliance with the standards of public health evidence, as well as the current TBS Evaluation Policy (2001) for compliance with the standards of performance reporting. The project concludes with recommendations to advance public health program evaluation planning, design and implementation, data collection and analysis, and reporting in the joint context of public health practice and the Canadian federal government accountability for performance.

INTRODUCTION

"Health improvement is what public health professionals strive to achieve" (Centre for Disease Prevention and Control [CDC], 1999). This goal can be only achieved by evaluating public health operation impacts. As the targets of the public health sector's actions have expanded in areas such as infectious and chronic diseases, surveillance, emerging pathogens, threats of bioterrorism, and health inequality, so has the role of program evaluation within this sector (CDC, 1999). As a result of these refinements, guidelines for best practices of program evaluation in public health practice have been developed over the last decade in various public health sectors and disciplines (CDC, 1999; Health Canada, 2004; Porteous et al., 1997; Salabarría-Peña et al., 2004). These public health program evaluation guidelines, frameworks and toolkits have been developed in order to ensure that among the conversions in the public health sector, programs will remain accountable and committed to achieving measurable public health outcomes.

Program evaluation was first introduced into the operation of the Canadian Federal Government in the late 1970s to improve management practices and controls (Treasury Boards of Canada Secretariat [TBS], 2004a). Consistent with its objective to help the Government of Canada manage its human, financial, information and technology resources, the TBS sets standards for best practices in evaluation across Canadian departments (TBS, 2004b). In February 2001, the TBS approved the "Evaluation Policy and Standards for the Government of Canada" (TBS, 2004a), including defining the scope of evaluation to include programs, policies and initiatives and positioning the discipline of evaluation as a core competence for results-based management (TBS, 2004a).

The Public Health Agency of Canada's (PHAC) Centre for Excellence in Evaluation and Program Design (CEEPD) is charged with leadership in promoting high-quality and effective evaluation practices across the Agency (PHAC, 2008b), and in the context of the Agency's accountability to the TBS for grant and contribution public health programs, evaluates these programs using TBS guidelines (PHAC, 2008b).

Evaluation for accountability versus evaluation for public health decision-making

Program evaluation is "the systematic examination and assessment of features of an initiative and its effects, in order to produce information that can be used by those who have an interest in its improvement or effectiveness" (WHO European working group on health promoting evaluation, as cited in Kelly et al., 2006).

The goal of program evaluation is to produce timely, credible, relevant, and objective findings supported by valid and reliable data collection and analysis. These findings are the foundation upon which decisions about the future of the program are made (TBS, n.d.).

The TBS Evaluation Policy (2001) supports the generation of accurate, evidence-based and objective information to aid managers make sound, more effective decisions on their programs, policies and initiatives, and by doing so meet the needs of Canadians. The policy further supports the accountability aspect of evaluation, stating that all grants and contributions programs will be reviewed in order to evaluate issues related to relevance, results and cost-effectiveness (TBS, 2001).

One of the fundamental principles of the TBS Evaluation Policy (2001) is that "achieving and accurately reporting on results is a primary responsibility of managers within the federal government" (TBS, 2001). Program evaluation reports help to maintain the accountability of results (TBS, n.d.) and are crucial documents to Canadian federal department and agency reports to Parliament (TBS, 2004a). Therefore, it is mandated that Canadian federal departments and agencies support their findings reporting and program performance through documented evaluations (TBS, 2004a), which are then used by federal departments to guide future decisions regarding a program's design and implementation (TBS, 2004b). As outlined by the TBS "Guide for the Review of Evaluation Reports" (2004), evaluation reports should follow the format and be reviewed in light of a set of key review criteria (TBS, 2004b).

Program evaluation is also "essential to documenting and disseminating evidence-based practices" (Brownson et al., as cited in Kelly et al., 2006). In the Canadian federal government, evaluations are used in program renewals as a line of evidence and serve in re-focussing the program in key areas (TBS, 2004b). Within the PHAC, the evaluation recommendations are implemented ensuring programmatic change - therefore, the role of evaluation, apart from its accountability aspect as mandated by the TBS, also carries out a learning function and is used for evidence-based practices.

Project purpose

The purpose of this project - to facilitate guidance for conducting effective program evaluations within public health in the Canadian federal government - is carried out by conducting a critical analysis of public health program evaluation reports available from the PHAC. Detailed information on the project framework is provided in the "Project Framework" section of this report.

Work to date

A detailed assessment of 115 Canadian federal government evaluation reports found that almost one quarter were found to be inadequate in terms of their quality[1]. A quantitative comparison of evaluation reports completed before and after April 2002 demonstrated that the quality of the federal evaluation reports has improved on a number of criteria, such as applying methodological rigour, identifying alternatives, presentation of evidence-based findings, addressing cost-effectiveness issues and providing formal recommendations. However, there is still a pressing need for further improvement (TBS, 2004c).

LITERATURE REVIEW OF EXISTING PROGRAM EVALUATION GUIDELINES AND REQUIREMENTS

Literature Review Methods

The search engines used to identify program evaluation guidelines in public health were Scholars Portal (social sciences) 1995 to present, Ovid (databases selected: All EBM Reviews Full Text; Ovid Medline (1996 to 2008); EMBASE 1996 to 2008), the Canadian Evaluation Society Journal, the American Journal of Evaluation, and federal health department websites (Public Health Agency of Canada, Health Canada, Canadian Institute for Health Research, Centre for Infectious Disease Prevention and Control). The keywords used for the search were: 'public health' *or* 'surveillance systems' *or* 'population health' *and* 'program evaluation guidelines' *or* 'program evaluation framework' *or* 'program evaluation standards', other options included 'public health' *or* 'surveillance systems' *or* 'population health' *and* 'program evaluation' *and* 'guidelines' *or* 'framework' *or* 'standards'. One program evaluation guideline was identified by a specialist in the field.

The objective of the search was to identify generic public health program evaluation guidelines which could be applied to broad areas of the field. Evaluation guidelines which targeted specific areas of pubic health, for example, cancer, HIV/AIDS or oral health were not taken into consideration.

---

[1] Quality - the degree to which reports comply with the TBS guidelines on writing reports; these requirements align with the TBS Evaluation Policy (2001).

A. Program Evaluation Best Practices in Public Health

A high-level review of public health evaluation best practices is provided below, which will later serve as the framework in the review and critical analysis of public health evaluation reports. The review of program evaluation standards is divided by the public health field a given evaluation guideline applies to. It is in the discussion section of this report, where examples drawn from the evaluation report analysis will be provided, that existing best practices will be reviewed on a more micro level.

Public Health Program Evaluation

The Centre for Disease Prevention and Control's (CDC) Framework for Program Evaluation in Public Health (1999) is one of the most recognized and detailed frameworks for common understanding of program evaluation concepts and promoting the integration of evaluation in the public health workforce to date. Among the framework's purposes is to provide an overview of the essential elements of program evaluation; provide guidelines for conducting effective program evaluations; clarifying the steps in program evaluation; and addressing misconceptions relating the purposes and methods of program evaluation. The framework is based on a practical approach to evaluation that focuses on steps and standards suitable for public health settings (CDC, 1999). The steps outlined in the framework describe what evaluators do, whereas the standards outline what has to be achieved for an evaluation to be effective. An analysis of the framework's steps is followed by a description of its standards (Milstein et al., 2000). Due to the framework's generic nature, it is feasible to apply it to the design and implementation of specific evaluations across many different public health program areas (CDC, 1999).

As previously mentioned, the framework's main focus is a detailed outline of the steps and standards in public health program evaluation. The primary element of the framework, the six steps of public health program evaluation, include 1) engaging stakeholders; 2) describing the program; 3) focusing the evaluation design; 4) gathering credible evidence; 5) justifying conclusions; and 6) ensuring use and sharing lessons learned. These six steps are to facilitate the understanding of a program's context (e.g., the program's setting, history, and organization) and to improve designing and conducting most evaluations (CDC, 1999). Therefore, according to the framework, when engaging stakeholders (step 1), engage individuals involved with the program and primary users of the evaluation.

When describing the program (step 2), "provide a description of the need, expected effects, activities, resources, stage, context, and the logic model". Focusing the evaluation design (step 3) adheres to "the purpose, users, uses, questions, methodologies, and agreements". In gathering credible evidence (step 4) the framework discusses "indicators, sources, quality, quantity, and logistics of program evaluation", whereas justifying conclusions (step 5) focuses on "the standards, analysis and synthesis, interpretation, judgment, and making recommendations". The framework's sixth step, ensuring use and

sharing lessons learned, discusses "the evaluation preparation, feedback, follow-up, and results dissemination" (Milstein et al., 2000).

The second element of the CDC framework, i.e., a set of thirty standards for appraising the quality of program evaluation activities are categorized into four groups: utility, feasibility, propriety, and accuracy. These four standards determine whether the evaluation will be effective, as well as are recommended criteria for judging the quality of public health program evaluation attempts (CDC, 1999). The four categories mandate the following: "utility - serve the information needs of intended users; feasibility - be realistic, diplomatic, prudent, and frugal; propriety - behave legally, ethically, and with regard for the welfare of those involved and those affected, and; accuracy - reveal and convey technically accurate information" (Milstein et al., 2000).

In their review of the CDC Framework for Program Evaluation in Public Health, Milstein and colleagues (2000) identify the purposes for which the CDC framework is being used based on health care practitioners' feedback obtained since the implementation of the framework in 1999. The review states that the framework is being used for purposes such as clarifying program strategies, developing guidelines, policies, and practices for evaluation; guidance of specific evaluation projects; training public health professionals and students; writing funding proposals, and producing complementary resources to support evaluation actions (Milstein et al., 2000). These findings are supported by Davis (2006) who presents the use of the CDC Framework for Program Evaluation in Public Health for creating and teaching practical evaluation methods to master's of public health students (Davis, 2006). Furthermore, the practicality and timeliness of the framework in assisting health care practitioners with evaluation efforts have also been recognized by Laferty and Mahoney (2003) who applied the framework's six steps to the evaluation design of a comprehensive community health promotion initiative (Laferty and Mahoney, 2003).

Apart from the literature mentioned above, the CDC Framework for Program Evaluation in Public Health has also been used as the basis for developing a program evaluation plan (Francisco et al., 2000). The process for evaluation planning was set up with the assistance of the CDC Framework, the Joint Committee on Standards for Educational Evaluation, and the evaluation chapters in the community toolbox. The authors of this guideline, similar to that of the CDC Framework, outline that evaluation planning involves clarifying the program's goals and objectives, identifying who the key audiences for the evaluation findings include, posing relevant and useful evaluation questions based on ones own and the stakeholders' needs, using evaluation methods and designs that will answer those questions to the satisfaction of key stakeholders, and creating timelines for evaluation activities and response to the inquiry (Francisco et al., 2000).

Another application is "The Program Evaluation Tool Kit" developed by Porteous and colleagues in the spirit of providing a common educational resource for public health program evaluation (Porteous et al., 1997). The Program Evaluation Tool Kit provides a simple five-step guide to "planning, conducting, and using program evaluation", and is presented in a series of short modules with practical descriptions and explicit tools

(Porteous et al., 1997). The Tool Kit contains a decision-oriented model of program evaluation and serves as a guide to a modest in scope, in-house, procedure and outcome evaluation. The Tool Kit is tailored towards helping managers integrate evaluation into their program management and help set a standard for evaluation and improve communication across health units (Porteous et al., 1997). The Tool Kit touches on points such as 1) focusing the evaluation; 2) choosing appropriate methodologies for answering evaluation questions; 3) developing or modifying data collection tools; 4) gathering and analysing the data; and last 5) interpreting the results and drawing conclusions to orient decisions about the program (Porteous et al., 1997). The Tool Kit also contains logistic guides for various data collection methods (e.g., surveys, focus groups, and case studies), which provide guidance on the methodological tasks' timelines and equipment supplies required, as well as responsible roles for the data collection (Porteous et al., 1997). Although the evaluation Tool Kit was designed to target public health protection and promotion programs, the approach is generic in nature and therefore can be applied to other program areas (Porteous et al., 1999).

An approach for critically appraising economic evaluations of health programs is discussed in "Methods for the Economic Evaluation of Health Care Programmes" (Drummond et al., 1997). The objective of Drummond's and colleagues' (1997) work is to provide guidance to multidisciplinary assessment teams (consisting of epidemiologists, clinicians, and economists) by providing them with a 'well-equipped tool kit' (Drummond et al., 1997). One of the main highlights of this economic evaluation methods text is the 'ten-point checklist' for assessing economic evaluations (Drummond et al., 1997), which is the foundation for the *British Medical Journal* guidelines for authors and peer-reviewers of economic studies and submissions (Drummond et al., 1996).

Public Health Program Evaluation - Surveillance Systems

The Updated CDC Guidelines for Evaluating Public Health Surveillance Systems (2001) address public health surveillance systems evaluations in order to make the evaluation process objective, inclusive, and explicit (CDC, 2001). These guidelines are tailored towards surveillance systems, however they may also be applicable to health information systems used for public health operations, pilot testing surveillance systems, and information systems at individual health care centers or hospitals. These evaluation guidelines could be also used for planning and creating systems, as well as effectively and efficiently monitoring a public health surveillance system (CDC, 2001).

The guidelines thoroughly outline and discuss each of the tasks involved in evaluating a public health surveillance system. The tasks' foundation are the six standards and four program evaluation steps obtained from the previously summarized Framework for Program Evaluation in Public Health (CDC, 1999), as well as from the elements in the original CDC Guidelines for Evaluating Surveillance Systems published in 1988. However, the uniqueness of these guidelines for evaluating surveillance systems lies in the fact that the guidelines outline the key surveillance system attributes, which should be considered when developing the evaluation questions, i.e., "acceptability, simplicity,

flexibility, data quality, predictive value positive, representativeness, timeliness, stability, compliance, efficiency, and effectiveness" (CDC, 2001). The Guidelines for Evaluating Public Health Surveillance Systems emphasize that not all activities under the evaluation tasks might be applicable for all surveillance systems, and therefore only those standards that are suitable to the system under evaluation should be used (CDC, 2001).

The CDC Guidelines for Evaluating Public Health Surveillance Systems (1988) were used by Sekhobo and Druschel (2001) to evaluate the New York State Congenital Malformations Registry (NYCMR), one of the largest state wide, population-based birth defects registries in the United States (Sekhobo and Druschel, 2001). In order to evaluate the NYCMR, the researchers assessed some of the NYCMR surveillance system's key attributes, as recommended by the CDC Guidelines, i.e., acceptability, simplicity, flexibility, predictive value positive, representativeness, timeliness, and sensitivity. Furthermore, they evaluated the cost of operating the system and its level of usefulness. (Sekhobo and Druschel, 2001). The study concluded that aligning the NYCMR evaluation and comparing it with the CDC Guidelines for Surveillance Systems was a useful approach and may be helpful for other congenital malformation systems evaluations. However, as the study authors point out, in line with the CDC Guidelines (2001), the congenital malformations systems have very unique issues, which should be recognized and included in evaluations (Sekhobo and Druschel, 2001).

Similarly to the Updated Guidelines for Evaluating Public Health Surveillance Systems (2001) having their guidelines based on the Framework for Program Evaluation in Public Health (1999), the Practical Use of Program Evaluation among Sexually Transmitted Disease (STD) Programs (2007) also sets its foundation on that same framework. This guide targets the evaluation of sexually transmitted diseases by providing a step-by-step approach and outlining the six evaluation steps provided in the Framework for Program Evaluation in Public Health (i.e., 1) engage stakeholders; 2) describe the program; 3) focus the evaluation design; 4) gather credible evidence; 5) justify conclusions; and 6) ensure use of evaluation results and sharing of lessons learned) (Salaberria-Pena et al., 2007).

A fourth evaluation framework identified in the area of public health surveillance is Health Canada's Framework and Tools for Evaluating Health Surveillance Systems (2004). This framework is tailored towards managers of public health surveillance systems in order to provide them with a standard approach for evaluating "the quality of the information that their systems produce; the effectiveness of their systems in supporting the objectives of the programs that they serve and in supporting informed decision-making, and; the efficiency of their systems" (Health Canada, 2004). The framework focuses on providing six steps in assessing public health surveillance systems: 1) establishing the context of the evaluation; 2) identifying the evaluation questions; 3) data collection and management findings, 4) collating and presenting the findings, 5) reviewing the report, and; 6) evaluation follow-up (Health Canada, 2004). This surveillance systems framework outlines the key surveillance system attributes, which should be considered when developing the evaluation questions (i.e., acceptability, simplicity, flexibility, data quality, predictive value positive, representativeness,

timeliness, stability, compliance, efficiency, and effectiveness). These same characteristics are also found in the previously mentioned CDC Updated Guidelines for Evaluating Public Health Surveillance Systems (2001) (CDC, 2001; Health Canada, 2004). The document also provides a template for reviewing evaluation reports, which outlines the key criteria for writing a public health-focused evaluation report (the evaluation report template was approved by Health Canada Audit and Evaluation Committee) (Health Canada, 2004).

Public Health Program Evaluation – Participatory Approach

The Guide to Project Evaluation: a Participatory Approach (1996) focuses on evaluation in the area of health promotion with a focus on a population health approach, although the guide states that it can be applied to other areas of public health, if modified and adjusted accordingly (Health Canada, 1996). The guide was developed in 1996 as the participatory approach seemed most consistent with the goals of Health Canada's strategies and programs. Participatory program evaluation is based on the recognition of the progression of change in skills, knowledge, attitudes and behavior, and as the guide points out, a participatory evaluation of a program is one that is never a "one-time, end-of-project event" (Health Canada, 1996).

The guide's main high-level focus is on a framework for project evaluation consisting of five key evaluation questions and five evaluation process steps; defining project work; developing success indicators; collecting evaluation data; analyzing and interpreting data; and use of evaluation reports. The guide to program evaluation also focuses on providing "tips and cautions" on various data collection methods used in assessing a population health approach program, such as surveys, focus groups, telephone interviews, observational studies, and program documents. These in turn are complimented by sample evaluation tools, i.e., questionnaires and interview guides, and framework worksheets for five key evaluation questions and success indicators (Health Canada, 1996).

The reviewed program evaluation guidelines in public health demonstrate that there are several guidelines, standards and tool kits that help in developing and conducting successful program evaluations in public health. Although each one of these guidelines has its own strengths and very useful recommendations on designing evaluations, engaging the right stakeholders, optimal data collection strategies or data analysis techniques, what none of them addresses is public health program evaluation within the Canadian federal government. Although public health programs should be evaluated according to public health program evaluation standards, they also must align and follow other federal government evaluation policies. Therefore, the next section of this literature review will focus on the evaluation policy requirements which Canadian federal government departments must adhere to in order to remain accountable for their programs.

B. Guidelines for Canadian federal government evaluations:

Treasury Board of Canada Secretariat Evaluation Policy (2001)

The Treasury Board of Canada Secretariat Evaluation Policy, last revised on April 1, 2001, clarifies the role of evaluation within the federal government's management framework. The policy ensures that the government has strategically focused, timely, objective and evidence-based information on the performance of its programs, initiatives and policies to produce better results for Canadians and meet their needs (TBS, 2001).

The TBS Evaluation Policy is based on three fundamental principles. The first principle states that achieving and accurately reporting on results is a primary responsibility of public service managers. Second, the Evaluation Policy outlines that rigorous and objective evaluation is a vital tool in helping managers to manage for results; and third, that departments with the support of the Treasury Board Secretariat are responsible to ensure that the discipline and rigour of evaluation are sufficiently spread out within their jurisdictions (TBS, 2001).

The Evaluation Policy's (2001) key requirements are:

- "Establishing an adequate evaluation capacity, including senior management;
- Encompassing a wider scope, including programs, policies and initiatives (including those delivered through partnership mechanisms);
- Developing strategic evaluation plans;
- Integrating evaluation with strategic decision-making and management;
- Placing greater emphasis on performance monitoring and early findings via:
    - ◊ Results-based Management and Accountability Framework (RMAFs) for new and renewed programs, policies, and initiatives;
    - ◊ Addressing issues regarding relevance, results, and cost-effectiveness;
    - ◊ Addressing issues regarding early administration and implementation; and
    - ◊ Continuous performance monitoring and measurement activities;
- Implementing Evaluation Standards of Practice"

(TBS, 2004a).

Whereas the TBS Evaluation Policy (2001) addresses the requirements that Canadian federal departments and agencies must adhere to when evaluating their programs – this also includes and applies to public health programs - its requirements are generic across government, and therefore only provide a very high level and narrow overview of program performance. Ideally, for the purposes of conducting program evaluations that contribute to the evidence-base for public health practice, it would be beneficial to have guidance and direction on how to conduct public health program evaluations which fulfill the requirements of the TBS Evaluation Policy (2001) and at the same time align with public health program evaluation guidelines and standards.

PROJECT FRAMEWORK

In light of the information presented in the literature review, the purpose of this project is to facilitate guidance for conducting effective program evaluations within public health in the Canadian federal government.

This will be done by critically analyzing program evaluation reports obtained from the Public Health Agency of Canada. The reports' analysis will be based on best practices for public health evaluations and best practices for federal government evaluations. The review and analysis of the public health evaluation reports will align with public health program evaluation guidelines and standards presented herein, as well as the current TBS Evaluation Policy (2001). This analysis will demonstrate how public health program evaluation within the federal government aligns with public health program evaluation guidelines and federal government evaluation requirements, as well as will fill in the missing gaps that the guidelines and requirements do not provide. Due to the fact that the program evaluation guidelines in public health provide a fairly detailed overview of evaluation practices, as demonstrated by the literature review, the analysis of public health evaluation reports will reveal missing gaps and areas for future improvement in conducting evaluation studies on a fairly micro level.

The reports' analysis will be based on the trends in the reports' strengths and weaknesses. The evaluation report analysis will capture the following information on public health program evaluation within the Canadian federal government:

- Provide an overview of the essential elements of public health program evaluation;
- Review existing public health program evaluation standards as applied in current federal government practice;
- Address strengths and weaknesses in the methodology, results, and recommendations made when evaluating a public health program, and;
- Provide recommendations on optimal measures in public health program evaluation.

The results of the analysis may be used to provide guidance and direction to public health program evaluators across the Canadian federal government on program evaluation planning, design and implementation, data collection and analysis, and reporting. Furthermore, the results of the analysis may be used as a potential source within the Canadian federal government should it be decided to draft best practices or tool kits in public health program evaluation. This would result in reliably and effectively assessed public health programs, and in turn would help managers and stakeholders take appropriate action to deliver programs aligning with the Public Health Agency of Canada's vision, mission, and mandate: to promote and protect the health of Canadians through leadership, partnership, innovation and action in public health, which is to result in healthy Canadians and communities in a healthier world (PHAC, 2008c).

METHODOLOGY

Evaluation reports:

The evaluation reports that were analyzed were the eight evaluation reports listed on the public PHAC website (http://www.phac-aspc.gc.ca/about_apropos/evaluation-eng.php):

- Evaluation of the Capacity-Building Component of the Canadian Breast Cancer Initiative;
- Summary Report of the Aboriginal Head Start Urban and Northern Communities 2003-2005 – National Impact Evaluation;
- Evaluation of the Canadian Diabetes Strategy 1999-2004;
- Evaluation of the Centres of Excellence for Children's Well Being – Summary Report;
- Evaluation of the Canadian Health Network;
- Evaluation of the Hepatitis C Prevention, Support and Research Program 1999 – 2006;
- Summative Evaluation of the National Health Surveillance Infostructure;
- Formative Evaluation of the Community Action Program for Children.

The evaluation reports listed above are all of PHAC's final evaluation reports to date and are available to the public. These reports that incorporated the CEEPD evaluator's feedback are denoted as revised versions of the original.

The reports' sections which were analyzed were the following: 1) evaluation methodology; 2) evaluation findings; and 3) conclusions and recommendations. As the main focus of the project was to analyze the methodologies and results analysis applied to Canadian federal government public health evaluations, the executive summary and introduction sections were beyond the scope of this analysis. However, when reviewing the evaluation reports' sections that fell within the scope of this analysis, strengths and weaknesses of actual evaluation reports were also taken into consideration.

Evaluation report review template:

The evaluation report review template (appendix 1) used to assess the eight evaluation reports is a modified version of the one found in the "Review of Quality of Evaluations Across Departments and Agencies" prepared by the TBS (TBS, 2004c). This review template was chosen as it was deemed to be a more detailed review guide, as opposed to the one developed by Health Canada (Health Canada, 2004). It must be noted that during the development of its review template, the TBS consulted among others the Health Canada review template (TBS, 2004c), and therefore the two review templates are similar (however the TBS review template allows for a more detailed and thorough assessment of the evaluation reports).

The ratings and scoring criteria assigned to the evaluation include:

- 'Poor': the score was assigned to an evaluation report criteria when information was substantially missing, the reader was experiencing difficulties in understanding the report based on the information provided. This rating was assigned to a given criterion requiring improvement;
- 'Adequate': the score was assigned when the report criterion met the requirements overall, however the information could have been expanded by inclusion of additional information and/or there were errors made;
- 'More than adequate': the score was assigned when a report presented information in a thorough and sufficient manner (the information was good) and no mistakes were noted;
- 'Non-applicable': this rating was assigned when information was not provided on a given criterion or was difficult to assess.

For clarity, the 'results' and 'discussion' section of this report follow the sequence of the review template used for the assessment of the evaluation reports.

Evaluation report review and assessment - rigour:

All eight evaluation reports were comprehensively reviewed against the guidelines on two separate occasions, typically one month apart. Afterwards, the evaluation reports' sections were compared with one another to assess compatibility of ratings assigned to the evaluation reports (i.e., as opposed to reading one full evaluation report once, sections based on a given set of criteria of all eight evaluation reports were compared with one another at one time).

Due to the scope of this project, only the evaluation reports were assessed; other additional compendiums that came with the reports were not considered.

RESULTS

1.0. Evaluation Methods

1.1. Evaluation Design and Methodology

Description of methodologies and design applied to the evaluation:

In relation to their description of the methodologies and designs applied to the evaluation, four out of eight evaluation reports were rated as 'more than adequate', three were rated as 'adequate' and one received a 'poor' score (table 1). Of those reports that were rated as 'more than adequate', the reports provided sufficient enough and detailed information on sample size, sample method, and instruments used to the extent the evaluation study could be replicated. The reports also provided information on how the instruments were developed and who was involved in their development. These reports would also link the methodology to evaluation issues and provide references to technical appendices, where

applicable. Two evaluation reports also pointed out that the evaluation instruments were developed through consultations with key stakeholders.

Four reports received an 'adequate' rating because some of the methodological information specified above was missing and incomplete. The general trend was that the reports had missing information on the sample size and sample method (e.g., respondent recruitment), and some of the methodological tools (e.g., questionnaires) were not explained. Also, the reports that provided the instruments in the appendices did not always reference these instruments in the text. Furthermore, some reports would provide a brief statement on a given methodological tool, however the instrument would not be provided in the report appendices. The reports' description of methodology also lacked at times an explanation of the linkage between methods and the evaluation issues. One report provided the methodology section as the last section of the report, which led to the reader having to guess the evaluation methodology when reading the report's findings.

The report judged as 'poor' lacked sufficient detail for the reader to understand the evaluation methodology either because no information was provided or they were referenced in an external source not included with the report.

The evaluation issues and questions are adequately addressed:

In relation to the evaluation issues and questions addressed, one evaluation report received a 'poor' rating, two reports received an 'adequate' score, and four reports received a 'more than adequate' rating (table 1). One evaluation report could not be assessed on this criterion, as it provided the information in a complementary document, not the report itself. The reports that were judged as 'more than adequate', apart from describing the evaluation issues and questions in text, also provided a matrix of evaluation issues, questions and lines of enquiry either in text or in an appendix. The three evaluation reports that received an 'adequate' rating either provided a description of the evaluation issues and questions in text or in a matrix. The matrix was found either in text or an appendix and was not accompanied by a description in the reports' text. Table 2 provides an overview of the issues being addressed in the evaluation reports. It must be noted that one evaluation report stated that the evaluation questions were developed by conducting focus groups with key stakeholders.

Table 1: Ratings of applied evaluation methodology and design in evaluation reports (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Description of the Methodology/ Design:				
Describes the methodologies and design applied to the evaluation	1	3	4	0
The evaluation issues and questions are adequately addressed	0	3	4	1

Table 2: Overview of key program areas addressed in the evaluation (N=8)

Evaluation Report	Key program areas addressed in the evaluation
Evaluation report 1	Relevance, success and cost-effectiveness
Evaluation report 2	Program role in supporting capacity development and capturing relevant outcome information
Evaluation report 3	Relevance, governance/ management, success, cost-effectiveness and alternatives
Evaluation report 4	Operational review (relevance), results/ achievement (implementation and effectiveness), strategic review (efficiency)
Evaluation report 5	Program reach and delivery, and to demonstrate to what extent the program has implemented a population health approach
Evaluation report 6	Relevance, progress/ success and cost-effectiveness
Evaluation report 7	Impact of program on target population
Evaluation report 8	Unable to assess

1.2. Multiple Lines of Evidence

The evaluation relies on more than one line of evidence to support its findings:

Six out of eight evaluation reports were rated as 'more than adequate' for employing multiple lines of evidence in their evaluation methodologies, whereas two evaluation reports received an 'adequate' rating (table 3). These same six reports ('more than adequate' score) used an appropriate balance between quantitative and qualitative methodologies, whereas the other two reports' ('adequate' rating) reliance was skewed towards qualitative methodologies. The multiple lines of evidence that were found in the evaluation reports from the most to least frequently occurring were: key informant interviews (7/8); document reviews (6/8); sample surveys (5/8); focus groups (3/8); 'other' (such as anecdotal user emails, research compendium, two year cohort study, evaluation studies of related initiatives, and expert panel review) (3/8); file and literature reviews (2/8); database reviews (2/8), and performance data analysis (1/8).

All stakeholder perspectives are included:

Three reports were rated as 'adequate' and five as 'more than adequate' in relation to including all stakeholder perspectives in the evaluation (table 3). It is recognized that it is not possible to include all stakeholder perspectives in one evaluation, however those evaluations that included a client and/or a program beneficiary at the same time including program management and other possible stakeholders were rated as 'more than adequate'. Therefore, two reports that were judged as 'adequate' included program management and other possible stakeholder perspectives, however failed to include program beneficiaries, whereas the remaining one report ('adequate' score) did not give a fair representation of program management, however included program beneficiaries. The stakeholder perspectives that were included in the evaluation reports from most to least frequently occurring were the following: program management and delivery (7/8); clients/ beneficiaries (5/8); partners (3/8); third-party deliverers (2/8); 'other' category (2/8), such as advisory committee members, and community members and professionals; and funding recipients (1/8). Only one report included qualitative evidence drawn from

key informants who did not have a stake in the program (community members and health professionals).

Table 3: Ratings of multiple lines of evidence applied in program evaluations (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Multiple Lines of Evidence:				
The evaluation relies on more than one line of evidence to support its findings (triangulation of results)	0	2	6	0
All stakeholder perspectives are included	0	3	5	0

1.3. Evaluation Limitations

Study limitations:

For describing methodological limitations three evaluation reports received an 'adequate' score, and five were assigned a 'more than adequate' rating (table 5). Among the five reports rated 'more than adequate', one report provided recommendations on how to avoid the study limitations, while another report provided examples of how the study was designed to avoid anticipated limitations. The reports rated 'adequate' failed to describe the study limitations in sufficient detail. The study limitations reported in the evaluation reports were in the area of data reliability (7/8) and actual and potential study bias (7/8). Table 4 provides an overview of the data reliability and bias limitations noted in the reports.

Table 4: Overview of bias and data reliability issues discussed in evaluation reports (N=8)

Criteria	Overview of biases and data reliability
Biases	▪ Self-report bias; ▪ Lack of inclusion of a specific group of participants in stakeholder survey (i.e., program beneficiaries); ▪ Only users of the evaluated tool surveyed, and not informants who do not use the tool; ▪ Participants not extensively surveyed due to unavailable resources; ▪ Use of internal evaluators.
Data reliability	▪ Low response rate; ▪ Lack of control group (cohort design); ▪ Representative group of key informants not obtained due to using random sample methodology; ▪ Lack of data compatibility (for cost-effectiveness) in reviewed reports; ▪ Documents used for evaluation were not related to the evaluation questions or did not span the appropriate evaluation period.

The constraints of the evaluation findings are made clear:

The constraints of the evaluation spanned from 'adequate' (2/8) to 'more than adequate' (5/8) and also included a 'non applicable' rating (1/8) (table 5). One report did not provide its evaluation constraints, hence was assigned a 'non applicable' score. The evaluation constraints identified in the reports were: data availability (4/8), high evaluation turn-around demand (3/8), budget (either limited financial resources or experiencing funding uncertainties) (2/8), and 'other' category (4/8), such as difficulty in attribution of program impacts (3/8); staff and management turn-over (2/8); changing and dynamic nature of the evaluated program's disease (2/8); experiencing difficulties in setting-up key informant interviews (1/8), and; unfamiliarity of key informants with the program (1/8).

Table 5: Ratings of evaluation limitations and constraints discussed in evaluation reports (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Evaluation Limitations:				
Limitations are described: actual and potential biases, reliability of data are identified and explained in relation to their impact on stated findings	0	3	5	0
The constraints of the evaluation are made clear	0	2	5	1

1.4. Methodological Rigour

Table 6 presents the factors used to appraise evaluation designs. One evaluation report did not provide sufficient details necessary to assess the rigour in key areas.

Table 6: Rigour identified in evaluation reports* (N=8)

Criteria	Yes (N)	No (N)	Unable to assess (N)	Examples of methodological rigour applied in evaluation reports
Survey of representative group of participants	3	4	1	▪Fairly high participant response rates; ▪Use of a broad scope of stakeholders.
Comparison group	2	5	1	▪Cohort group comparison; ▪Comparative review of cost-effectiveness with other organizations.
Comparison to baseline measures	2	5	1	▪Comparison of target audience at the beginning vs. end of one year cycle; ▪Comparison of program implementation phase vs. its development phase.
Comparison to norms/ literature/ other benchmark	6	2	0	▪Comparison of study results to literature results; ▪Past surveys and stakeholder consultations in tool development; ▪Study results comparison to census data.
Other	6	2	0	▪Pilot-tested instrument(s); ▪Use of independent expert panel review; ▪Use of objective 3rd party for data collection/ analysis; ▪Results reviewed by participants for interpretation accuracy.

*As identified by the evaluation report, not the reader

1.5. Analysis

Five reports were considered to be 'adequate', two reports were rated as 'more than adequate' and one was rated as 'non applicable' in relation to the data supporting the analysis as determined by response rates, significant tests and multiple lines of evidence.

Five evaluations ('adequate' score) resulted from low participant response rates. Three out of the five reports addressed this low sample size. One of these reports stated that despite the low response rate and statistical insignificance, there was a fair representation of respondents from all geographic areas. Another report stated that in spite of the low response rate, the actual number of respondents was high, and supported its statement with a reliable margin of error. Another report declared that great care was taken to not only provide the percentage, but also the number of respondents, and concluded that the findings were supported by the respondent number. Two other reports ('adequate' score) did not provide caveats to the low sample size.

Two other reports ('adequate' ratings) included only one line of evidence in their evaluation of program's cost-effectiveness. One report did address this limitation by justifying that one out of the two lines of evidence led to inconclusive findings.

Two evaluations resulted in relatively high participant response rates ('more than adequate' rating) and one evaluation was difficult to assess based on the insufficient methodological information ('non applicable' score).

2.0. Results Analysis of Evaluation Reports

2.1. Evidence of Relevance

Evidence of need for the program and/ or evidence demonstrating program responsive to need:

In relation to whether the reports demonstrate actual need and provide evidence to demonstrate responsiveness to need, two of the evaluation reports were perceived as 'adequate', two were judged as 'more than adequate', whereas the remaining half received a 'non applicable' score (table 7). The information that was provided in the reports on this criterion would either meet the requirements or go into fairly sufficient detail. Four reports ('non applicable' rating) did not address this issue in their evaluation.

It must be noted that two evaluation reports which did not address relevance were formative evaluations. The TBS states that a formative evaluation (aka interim evaluation) examines the effectiveness of program implementation in order to facilitate improvement. A formative evaluation is normally conducted in mid-cycle of a program, usually within 2 years of the program's implementation. The purpose of a formative evaluation is to provide information to improve the delivery of the program, such as the quality of performance information and reporting systems. Formative evaluations assess outputs, early results, validation of program logic, and the likelihood of long-term results achievement (PHAC, 2008b; TBS, 2004b). On the other hand, summative evaluations assess impacts in order to allow decisions be made about overall program, policy or initiative effectiveness. These evaluations are usually carried out at the end of a four-year life cycle of a program when intermediate and long-term outcomes emerge. Summative evaluations perform an accountability function as they focus on the extent to which a program's desired outcomes have been achieved, and the degree to which the program has contributed to achieving these outcomes. Summative evaluations also provide recommendations on program design (PHAC, 2008b; TBS, 2004b).

Evidence to demonstrate continued relevance to government priorities:

The ratings for evidence to demonstrate continued relevance to government priorities were either 'more than adequate' (4/8) or 'non applicable' (4/8) (table 7). Two reports ('more than adequate' score) also provided findings to demonstrate not only relevance to government priorities, but also to PHAC's priorities and/or provincial and regional health requirements. The remaining four reports did not address this issue in their evaluation.

Throughout the assessment of the eight program evaluation reports, it was noted that one evaluation also addressed evidence to demonstrate program relevance to Canadians (deemed as 'more than adequate').

Evidence demonstrating that the program does not duplicate or otherwise work at cross purposes to other programs, policies, or initiatives:

The scores for evidence to demonstrate that the program does not duplicate or work at cross purposes with other programs, policies, or initiatives spanned from 'adequate' (1/8) to 'more than adequate' (2/8) and included 'non applicable' (5/8) scores (table 7). The one report that received an 'adequate' rating provided correct and sufficient information, meeting the criterion requirement. Two reports provided an in-depth analysis of this issue in question. However, the majority of reports (5/8) did not address this issue in their evaluation question.

Table 7: Ratings of results discussed in evaluation reports – relevance (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Evidence to demonstrate actual need and/ or evidence to demonstrate responsiveness to need	0	2	2	4
Evidence to demonstrate continued relevance to government priorities/ needs	0	0	4	4
Evidence to demonstrate that it does not duplicate or work at cross purposes with other programs, policies, or initiatives	0	1	2	5

2.2. Success

Clearly describes what has happened as a result of the program and articulates attribution of program, policy or initiative to success:

Seven reports were rated as 'more than adequate' on clearly describing what has happened as a result of the program and articulated attribution of program to success. One report was rated as 'adequate' (table 7). It must be noted that one of these reports ('more than adequate' rating), overall, presented the evaluation success findings in a non-objective manner, highlighting the program's positive aspects. The report that received an 'adequate' rating on this criterion was a report which emphasized the difficulty in attribution of changes in indicators directly to the program; the report stressed that indicators of success tend to change at different rates due to the changing and evolving nature of diabetes. As the evaluation report pointed out, these indicators of success often take many years, and are also influenced by complex issues of attribution, and hence it was difficult to attribute success to the program.

Table 8: Ratings of results discussed in evaluation reports – success (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Clearly describes what has happened as a result of the program and articulates attribution of program, policy or initiative to success	0	1	7	0

2.3. Results – cost- effectiveness

Presents a qualitative and/or quantitative assessment of cost-effectiveness:

The one report that received a 'poor' rating on assessing the program's cost-effectiveness, conducted a comparative review with five other organizations and key informant interviews (table 8). However, the comparative review did not provide adequate financial information to determine the cost-effectiveness of the program due to data availability and consistency issues. On the other hand, in the same report addressing cost-effectiveness, interviewees declared that the program is very cost-effective. Although the report did acknowledge that the comparative review did not lead to conclusive findings, the reliance on qualitative data (i.e., the interviews) alone to evaluate cost-effectiveness resulted in the report receiving a 'poor' score on this criterion. The second report that received a 'poor' score on cost-effectiveness evaluation, was a report that used only one qualitative source of evidence to assess this criterion (table 8).

The two remaining reports ('more than adequate' rating; table 8) both provided a detailed qualitative and quantitative assessment of program cost-effectiveness. It must be noted that the evaluators of one of these assessments have to be commended for the fact that although there was a lack of accepted methods, tools or indicators available for measuring the specific program's cost-effectiveness, the evaluation did carry it out fairly effectively.

The remaining four reports did not assess the program's cost-effectiveness, however one report provided a rationale as to why this was not assessed and another report was a formative evaluation (table 8).

One report ('non applicable' rating) described the allocation resources to different program components; however this was not considered an evaluation of cost-effectiveness.

Table 9: Ratings of results discussed in evaluation reports – cost-effectiveness (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Identifies the extent to which the program, policy or initiative could have been delivered by more appropriate, cost-effective methods to achieve its objectives - evidence supported by both qualitative and quantitative methods	2	0	2	4

2.4. Evidence-based findings

In relation to whether the findings were based on evidence drawn from the evaluation research, seven of the evaluation reports were rated as 'more than adequate' and one report was assigned an 'adequate' score. The report that was assigned an 'adequate' score

had gaps in the evaluation findings, which were stressed throughout the report as an evaluation challenge and were assigned to the difficulties in attribution and the changing and dynamic nature of the epidemic the program addressed.

3.0. Conclusions

Conclusions objectively answer the evaluation issues and are supported by the findings:

Five evaluation reports were judged as 'more than adequate' on this criterion, whereas three reports were considered to be 'non applicable' scores as they did not present conclusions in their reports (table 9). Of those that did not present conclusions in the report, one evaluation report had 'gaps and lessons learned' in place of the conclusions, however it must be noted that this section did support the evaluation issues and findings. Another evaluation report in place of conclusions had 'observations and recommendations of the expert panel', however this section went beyond of what is to be expected from the conclusions section. Finally, the third report that received a 'non applicable' rating, did not contain a conclusions section; only a summary preceding the findings section was found in the report.

Presents lessons learned about the program from the evaluation:

Four evaluation reports received a 'more than adequate' rating on the lessons learned, whereas the remaining reports did not contain this section in their reports (table 9). The lessons learned sections provided in the four reports spanned the domains of strategic alignment, operational framework, and results and achievements; management and program design and delivery; evaluation planning; and governance and management, collaboration and policy.

Table 10: Ratings of conclusions presented in evaluation reports (N=8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Conclusions objectively answer the evaluation issues and are supported by the findings	0	0	5	3
Presents lessons learned about the program from the evaluation	0	0	4	4

4.0. Recommendations

Recommendations are present:

All eight evaluation reports presented formal recommendations, but one which dispersed them throughout the findings section. One report's recommendations were that of the expert panel's committee (table 10).

Recommendations are operational, practical and realistically attainable:

In general, the recommendations for the eight evaluation reports were operational, practical and realistically attainable (six reports received a 'more than adequate' rating and two were judged as 'adequate') (table 10). Two reports ('adequate' score) provided a recommendation that went beyond the program's scope. Examples of such recommendations are the following: "the federal government could do more to address the challenge of some populations being at greater risk of diabetes", "there is a need for a legislated body outside the government that could report candidly to the Prime Minister and the people of Canada on the implications of population and public health promotion across the country", and "to ensure funds are available at the start of the fiscal year". Out of all eight reports one evaluation report also addressed the impact the proposed recommendations may have on the program.

Recommendations address significant evaluation findings:

Seven reports' recommendations addressed the evaluation significant findings ('more than adequate' rating) and one report's recommendations did not ('poor' score) (table 10). In the report that was judged as 'poor', it was very unclear which evaluation findings were being addressed by the recommendations and why the specified recommendations were being proposed.

No report presented evidence indicating that the public health program was not needed or not relevant to the Agency's mandate. It must be noted that the evaluation findings were accompanied by recommendations on implementing changes to the programs, however always in the context of the public health program being needed (TBS, 2004c).

Recommendations flow logically from findings and conclusions:

One evaluation report was judged as 'more than adequate' and five reports were considered to be 'adequate' in relation to their recommendations flowing logically from findings and conclusions. One evaluation report was rated as 'poor' and one received a 'non applicable' score (table 10).

The evaluation report that was rated as 'more than adequate' categorized the conclusions and recommendations by key evaluation areas (success, relevance, and cost-effectiveness) introducing flow into this section. This section of the report clearly outlined each of the key evaluation questions, where each evaluation question was followed by conclusions and where applicable, a recommendation.

Of those reports that received an 'adequate' rating for this criterion, one had its recommendations consist of "considerations and recommendations", where the considerations were too long and detailed, followed by recommendations, which on the other hand, were too short. Another report substituted the conclusions section with a "gaps and lessons learned" section. However, the logic flow from the findings and the "gaps and lessons learned" to the recommendations was maintained.

The report which was judged as 'poor' did not align its findings and conclusions with the proposed recommendations, and hence it was unclear why the specified recommendations are being proposed. Furthermore, the report did not number and organize the recommendations which did not have a good impact on the overall presentation of this section of the report.

The one report that received a 'non applicable' rating, did not contain a conclusions section in the report, hence it was not feasible to assess whether the evaluation recommendations flow logically from the conclusions.

Table 11: Ratings of proposed recommendations in evaluation reports (N= 8)

Criteria	Poor (N)	Adequate (N)	More than adequate (N)	N/A (N)
Recommendations are operational, practical and realistically attainable	0	2	6	0
Recommendations address significant evaluation findings	1	0	7	0
Recommendations flow logically from findings and conclusions	1	5	1	1

DISCUSSION

1. Evaluation Methods

1.1. Evaluation Design and Methodology:

Evaluation report review findings demonstrated that the methods applied to the evaluation, e.g., instruments, key informants, recruitment procedures, sample size and sample method were discussed to varying degrees of detail (table 1). The guidelines for reviewing evaluation reports outlined in the Health Canada Framework and Tools for Evaluating Health Surveillance Systems (2004) state that the design of the evaluation is described to the extent that the study can be replicated (Health Canada, 2004). This information is supported by the CDC Guidelines for Program Evaluation in Public Health, which state that all technical information should be provided in an evaluation report, for example in the appendices (CDC, 1999). As demonstrated by the evaluation report analysis findings, the description of the various methods applied did not always meet the specified standards (table 1); therefore, due to incomplete or missing information in the body of report or appendices, it would not have been feasible to replicate the studies. Furthermore, public health program evaluation guidelines clearly outline that the methods section in an evaluation report should come after the introduction (CDC, 1999; Health Canada 1999; Health Canada 2004; Porteous et al., 1999); this standard was not met by two reports, as one report provided the methodology section as the last section of the report (which led to the reader having to guess the evaluation methodology when reading the report's findings), and another provided this section in a summary report (and at an insufficient level of detail). Therefore, when disseminating program evaluation findings, it is advised to report methods with a high level of detail, to the extent the study can be replicated.

The TBS Evaluation Policy (2001) provides a high level overview of the purpose of the evaluation by stating that a program, policy or initiative evaluation should focus on issues related to relevance, results and cost-effectiveness (TBS, 2001). Furthermore, the policy outlines that "the full range of evaluation issues (i.e., relevance, success, and cost-effectiveness) should be considered at the planning stage of an evaluation" (TBS, 2001). The reason for clearly identifying the broad evaluation goals from the very beginning is because otherwise one can lose sight of the evaluation's macro level picture (Porteous et al., 1997).

In their program evaluation toolkit, Porteous and colleagues (1997) identify the reasons for evaluating a program; the recognized purposes of evaluation span from identifying the program's strengths and weaknesses, measuring progress, sharing experiences, improving delivery and implementation, investigating the program's successes, or determining which aspects of a program should be continued and which should be discontinued (Porteous et al., 1997). This information is complimented by the CDC Guidelines on Program Evaluation in Public Health (1999) which provide a detailed outline of the four purposes of public health program evaluations in the guideline's third step ("focusing the evaluation design"). According to these guidelines, the first purpose is to gain insight, which takes place when assessing the feasibility of an innovative approach to practice. A second purpose for program evaluation is to change practice, which usually takes place in the implementation stage when an established program is being assessed on what it has accomplished and to what degree. Evaluations carried out for this purpose include efforts to improve the quality, effectiveness, or efficiency of program operations. A third purpose for evaluation is to assess program effects - this type of evaluation is applicable to mature programs which can state what interventions were delivered to what proportion of the target population. Finally, a fourth purpose, which is applicable to any level of program development, is based on using the process of evaluation inquiry to affect those who participate in the inquiry (CDC, 1999).

The key program areas addressed in the reviewed evaluations spanned from relevance and success to cost-effectiveness. The success evaluation area included issues such as program reach, delivery, and program impact on target population (table 2). The findings of the evaluation report review demonstrate that the assessed public health program evaluation purposes align with existing public health guidelines and the requirements of the TBS Evaluation Policy (2001).

Furthermore, one evaluation report noted that the evaluation questions were developed in consultation with stakeholders via focus group discussions. As outlined by public health program evaluation guidelines and supporting literature, the evaluation design can also directly involve stakeholders (CDC, 1999; Francisco et al., 2000; Porteous et al., 1997) - in such cases, both supporters and sceptics of the program may be consulted to ensure that the proposed evaluation questions are feasible and are responsive to the varied positions of different interest groups (CDC, 1999).

1.2. Multiple Lines of Evidence

The evaluation report analysis revealed that six evaluation reports used an appropriate balance between quantitative and qualitative methodologies, whereas two reports' reliance was skewed towards qualitative methodologies. Both qualitative and quantitative methodologies are recognized as valid, complementary approaches for data collection, with neither inherently superior (Health Canada, 1996; Porteous et al., 1997).

Each method has its merits, and the majority of researchers and practitioners accord that the integration of both provides a well rounded picture of the evaluation (Francisco et al., 2000). This statement is supported by the CDC Framework for Program Evaluation in Public Health (1999) which states that the combination of both quantitative and qualitative data can increase the chances that the evidence will be well-balanced (CDC, 1999). Rootman and colleagues (2001)

- Position qualitative knowledge as the foundation of all quantitative knowledge;
- Agree that qualitative methods provide preliminary measures underlying more sophisticated quantitative measures in that they i) facilitate the interpretation of quantitative data, ii) enable analysts to understand and investigate threats to their validity, and iii) assist in eliminating rival hypotheses; and
- Note that qualitative data are useful for assessing a study's generalizability

(Rootman et al., 2001).

Therefore, the use of both quantitative and qualitative approaches in public health program evaluation should be intertwined thereby increasing the chances of a well-rounded picture of evidence-based findings.

Analysis of PHAC's public health program evaluations finds multiple lines of evidence (ranked from most to least frequently occurring): key informant interviews; document reviews; sample surveys; focus groups; 'other' (such as anecdotal user emails, research compendia, two year cohort study, evaluation studies of related initiatives, and expert panel review); file and literature reviews; database reviews, and performance data analysis.

The CDC Guidelines on Program Evaluation in Public Health (1999) recognize a variety of selected techniques for gathering evidence, for example surveys, personal interviews, focus groups, group assessments, document analysis, portfolio reviews, and many more (CDC, 1999). The findings from the evaluation report analysis are also supported by Porteous and colleagues (1997) who state that a program's evaluation should incorporate data gathered from a variety of sources with varying perspectives (Porteous et al., 1997). Furthermore, in most cases, any single data collection method will not be fully sufficient for a program evaluation (TBS, n.d.), and although all types of data are characterized by weaknesses, an evaluation's overall credibility can be improved by using multiple data sources as it provides an opportunity to include different perspectives regarding the program (CDC, 1999). Therefore, when feasible, it is always beneficial to use several different data sources and data collection methodologies (TBS, n.d.).

The CDC Framework for Program Evaluation in Public Health (1999) recognizes three types of stakeholders: "those involved in program actions (e.g., sponsors, coalition partners, collaborators, administrators, funding officials, managers, and staff", those "served or affected by the program (e.g., clients, family members, neighbourhood organizations, academic institutions, professional associations, advocacy groups, elected officials, opponents, sceptics, and staff of common or competing organizations)", and "primary users of the evaluation (individuals who can take action or decide on something relating to the program; in practice, they are really a subset of all program stakeholders)" (CDC, 1999).

The stakeholder perspectives included to support the validity of the findings found in PHAC's evaluations from most to least frequently occurring were the following: program management and provider (7/8); clients/ beneficiaries (5/8); partners (3/8); third-party deliverers (2/8); 'other' category (2/8), such as advisory committee members, and community members and professionals; and funding recipients (1/8). Furthermore, five out of eight evaluation reports were recognized as being 'more than adequate' in engaging all stakeholder perspectives in the program evaluation (table 3). Although it was recognized that it is not feasible to include all stakeholder perspectives in one evaluation, those evaluations that included a client and/or a program beneficiary, at the same time including program management and other possible stakeholders received a higher rating ('more than adequate').

The CDC Framework for Program Evaluation in Public Health (1999) emphasizes that individuals or organizations affected by the program (either directly or indirectly) should be engaged in the evaluation to the extent possible (CDC, 1999). This includes engaging stakeholders when focusing the evaluation design as was discussed previously in the text, however it may also include the data collection phase, when the stakeholders are the interviewees who provide evidence-based findings relating the program. Furthermore, the framework points out that engaging with individuals who are antagonistic or openly sceptical towards the program may provide important information – that is, opening the evaluation to opposing views and engaging program opponents in the investigation may strengthen the evaluation's findings and its credibility (CDC, 1999).

Furthermore, the CDC Framework for Program Evaluation in Public Health (1999) also recognizes the value of including those stakeholders involved in program operations (CDC, 1999). Individuals or organizations involved in program operations have a stake in how evaluation activities are conducted as the program might be modified as a result of the evaluation findings. Although staff, partners, and funding officials work collaboratively on a program, they are not inevitably a group composed of single interests, and therefore these different subgroups are valuable in program evaluation for holding different perspectives. Furthermore, these stakeholders may provide an inside perspective of a program, which individuals affected by the program may not always be able to provide (CDC, 1999).

Only one evaluation was found to include qualitative evidence drawn from key informants who did not have a stake in the program, i.e., the overarching society

(community members and health professionals). The CDC Framework on Program Evaluation in Public Health (1999) states that engaging neutral observers is beneficial for an evaluation (CDC, 1999), and therefore, this approach may be worth considering as these informants may provide an unbiased perspective. However, these key informants are to be engaged only if they have sufficient knowledge and information relating the program, otherwise the data collected may either be unreliable or result in missing responses.

In conclusion, using multiple lines of evidence is vital in program evaluation, whether it is of stakeholders' perspectives or various data collection techniques. An inside perspective might be extracted from internal documents, such as program files, reports and databases, and comments from program managers or staff, whereas clients, neutral observers, or those antagonistic to the program might provide a different, but equally relevant point of view. Combining these provides a well-rounded and comprehensive picture of the program (CDC, 1999).

1.3. Limitations

The TBS Evaluation Policy (2001) requirement for evaluation reports is that the reports should "clearly expose the limits of the evaluation in terms of scope, methods and conclusions" (TBS, 2001). Several public health program evaluation guidelines provide guidelines for assessing evaluation reports (Health Canada, 1996; Health Canada, 2004) or checklists for evaluation report writing (CDC, 1999). Due to the fact that each evaluation method has its own biases and limitations (CDC, 1999), these guidelines state that the evaluation report should clearly describe the limitations and trade-offs of the methodologies, data sources and data used in the evaluation (CDC, 1999; Health Canada, 1996; Health Canada, 2004), and that "the constraints of the evaluation and the perspective from which the intervention is evaluated are clear and the reader can assess the validity of the evaluators' judgement" (Health Canada, 2004). The findings from the report review demonstrated that the evaluation reports would provide a description of the limitations (to varying degrees of detail, hence different ratings were assigned for this criterion, table 5). Furthermore, usually the reports would provide a description of the evaluation's constraints (table 5), which were identified to be data availability, evaluation time and budget constraints, and 'other' category, such as staff and management turn-over; experiencing difficulties in setting-up key informant interviews; unfamiliarity of key informants with the program; changing and dynamic nature of the evaluated program's disease; and difficulty in attribution of program impacts. Therefore, the evaluation reports did align with the TBS Evaluation Policy (2001) requirements and public health program evaluation guidelines in relation to describing study limitations and constraints. When describing the study limitations and constraints in evaluation reports, it must be kept in mind that no evaluation method is without limitations and every evaluation will contain some constraints, and therefore the goal is to meet the quality level that meets the stakeholders' threshold for credibility (CDC, 1999).

Although the TBS Evaluation Policy (2001) and public health program evaluation guidelines do not require the evaluation reports to provide recommendations on avoiding

methodological limitations in future evaluations, report review findings demonstrate that one evaluation report did discuss this issue. Following the example of this particular evaluation report, making recommendations on how to omit methodology limitations in future studies may be a suggestion for future evaluation reports. Furthermore, one evaluation report also discussed how potential methodological limitations, such as biases and data reliability were avoided, by foreseeing them in the evaluation planning and design stage. A couple of public health evaluation guidelines identify different strategies which provide examples of advantages and disadvantages of various data collection techniques. Applying these strategies in the evaluation design phase helps avoid study biases and data reliability issues (Health Canada, 1996; Porteous et al., 1997). Furthermore, taking into consideration evaluation constraints such as data availability, time and cost can also influence the selection of data collection methods (CDC, 1999). Therefore, when designing an evaluation, evaluators should weigh the various advantages and disadvantages of each method and take into consideration the various evaluation constraints as in the end these will impact the credibility of the results. Furthermore, providing examples of how the study was designed in such a way as to avoid encountering potential limitations may be useful when writing an evaluation report, as it helps the readers assess the methodological rigour applied, and hence have more confidence in the credibility of the findings.

1.4. Methodological Rigour

One of the three fundamental principles of the TBS Evaluation Policy is that "a rigorous and objective evaluation is an important tool in helping managers to manage for results" (TBS, 2001). Public health guidelines on program evaluation identify the application of methodological rigour by emphasizing the use of pilot testing of developed tools on small subgroups prior to the evaluation (Health Canada, 1996; Porteous et al., 1997) or by providing 'tips and cautions' on using different collection tools, such as focus groups, surveys, and interviews in order to yield higher response rates (Health Canada, 1996). Examples of such 'tips and cautions' include limiting the number of questions on a questionnaire, use of plain language, not biasing responses by posing certain questions in a specific manner, and paying attention to respondents' literacy level, language and visual capacity (Health Canada, 1996). However, apart from the examples provided above, the guidelines do not provide further information on applying rigour to program evaluations.

The findings from the evaluation reports analysis demonstrated that the rigour applied in public health program evaluations applies to the following areas (table 6):

- surveying a representative group of participants, e.g., obtaining high participant response rates or using a broad scope of stakeholders;
- using a comparison group, e.g., cohort analysis;
- making a comparison to baseline measures, e.g., comparing a program's implementation phase versus its development phase;
- comparison to norms, literature or other benchmark, e.g., comparing study results to literature findings;

- using past surveys and stakeholder consultations in tool development; or comparing study results to census data; and
- 'other', such as:
    - ◊ pilot testing tools;
    - ◊ use of independent expert panel review or an objective third party for data collection and analysis; and
    - ◊ reviewing results by participants for interpretation accuracy.

As the public health evaluation guidelines do not discuss methodological rigour thoroughly, these findings may broaden the existing scope of information of the different types of rigour to be used in program evaluation, and perhaps serve as examples when applying rigour to future evaluation studies in public health.

1.5. Analysis

Public health program evaluation guidelines provide guidance on effectively summarizing, synthesizing, and interpreting data results; the guidelines provide direction on effectively detecting patterns and themes of the evidence, by either isolating the data or combining various data sources (CDC, 1999; Health Canada 1996). However, what the guidelines, and the TBS Evaluation Policy (2001), fail to address is the micro level of analyzing the data, as for example, approaches to data analysis when the response rate is low or when findings are supported by only line of evidence, as was found in the evaluation report analysis.

The Review of the Quality of Evaluations across Departments and Agencies (2004) suggests that data limitations should be supported by appropriate caveats before proceeding to analyze the results (TBS, 2004c). Examples of such caveats, as demonstrated by the evaluation report review findings, can include the following:

- Despite a low response rate and statistical insignificance, a fair representation of respondents from all geographic areas was present;
- In spite of the low response rate, the actual number of respondents was high (such a statement may be further supported with a reliable margin of error); or
- Great care was taken to not only provide the percentage, but also the number of respondents, and conclude that the findings were supported by the respondent number.

These are to serve as examples of addressing low response rates (an evaluation limitation noted most often in the reviewed evaluation reports) and therefore other justifications for data analysis are most acceptable. Out of the two reports that used one line of evidence to assess cost-effectiveness, one addressed this limitation by justifying that one out of the two lines of evidence led to inconclusive findings. Although the limitation was supported by a caveat, it is still considered to be insufficient, and the analysis, making conclusions and recommendations of such findings should be approached with great caution.

It is important to support the data analysis with appropriate and convincing caveats, as otherwise the validity of the data and the conclusions reached by using this evidence will be unreliable (as was the case with the evaluations which resulted in low response rates and which had one line of evidence to assess cost-effectiveness); this in turn is important as stakeholders must agree that the evidence and conclusions are justified before they will use the program evaluation findings with confidence (CDC, 1999).

2. Evaluation Findings Analysis

2.1. Relevance

The findings from the evaluation report analysis demonstrated that four out of the eight public health program evaluations demonstrated actual need and provided evidence to demonstrate responsiveness to need, as well as assessed program continued relevance to government priorities (table 7). The TBS Evaluation Policy (2001) states that an issue to be considered during an evaluation is whether "the program, policy, or initiative continues to be consistent with departmental and government-wide priorities and whether it realistically addresses an actual need (relevance)" (TBS, 2001). Addressing program relevance in an evaluation has also been identified by public health evaluation guidelines (Health Canada, 2004). Therefore, evaluating whether a program is needed as well as whether it is still relevant to government priorities are issues that should be implemented in the evaluation design.

The evaluation report review analysis also demonstrated that one out of the eight evaluations assessed whether the program meets the needs of Canadians. It is the TBS Evaluation Policy's (2001) primary objective to ensure that the government has "timely, strategically focussed, objective and evidence-based information on the performance of its programs, policies, and initiatives to produce better results for Canadians" (TBS, 2001). Furthermore, the TBS in its Program Evaluation Methods: Measurement and Attribution of Program Results outlines that continued relevance is one of the evaluation issues to be considered during an evaluation, and further explains that an evaluation should also investigate continued relevance in light of whether the program meets the needs of Canadians (TBS, n.d.). The CDC Framework on Program Evaluation in Public Health (1999) states that one of the purposes of the evaluation is to assess the needs of community members (CDC, 1999). Due to the fact that TBS puts a focus on results for Canadians in the Evaluation Policy's (2001) objective, as well as this issue is indirectly recognized by public health program evaluation guidelines, results for Canadians is worth taking into consideration when evaluating public health programs - especially that it was recognized as an evaluation issue in one of the reviewed evaluation reports.

The findings from the evaluation report analysis revealed that three out of the eight program evaluations assessed to varying degrees of detail whether the program does not duplicate or work at cross purposes with other programs (table 7). The Updated CDC Guidelines on Public Health Surveillance Systems (2001) outline that "an evaluation should assess the degree to which a public health surveillance system is integrated with other surveillance and health information systems (...) as taking appropriate action to

integrate programs within the public health surveillance network enables individual programs to meet specific data collection needs while avoiding the duplication of effort and lack of standardization that can arise from independent programs" (CDC, 2001). This information is further supported by the TBS Program Evaluation Methods: Measurement and Attribution of Program Results guidelines, which state that an evaluation should explore to what extent a program duplicates, compliments, overlaps, or works at cross-purposes with other programs (TBS, n.d.). Drummond and colleagues (1997) in their ten-point checklist for assessing economic evaluations, state that when assessing a program's evaluation, it should be noted whether a comprehensive description of competing program alternatives was provided (Drummond et al., 1997). Although evaluation of program duplication is not identified by the TBS Evaluation Policy (2001), based on the reviewed reports and provided literature evidence, this evaluation issue may be a criterion worth taking into consideration when evaluating a program.

2.2. Success

The TBS Evaluation Policy (2001) outlines that another evaluation question and issue that should be considered at the planning stage of an evaluation is whether the policy, program or initiative is effective in meeting its objectives, within budget and without unwanted outcomes, i.e., whether the evaluation questions and issues relate to a program's success factors (TBS, 2001). Public health guidelines for program evaluation on numerous occasions state that in order to assess program effects, the level of success in accomplishing program goals should be documented (CDC, 1999; Health Canada, 1996; Health Canada, 2004; Porteous et al., 1997). As demonstrated by the evaluation report review findings, all of the evaluations had taken into consideration this factor in their evaluation questions, addressing this issue in various forms (e.g., program reach, implementation, progress, results, achievements, and impacts) (tables 2 and 8).

The evaluation report review findings demonstrated that one evaluation report in general presented the evaluation success findings in a non-objective manner, highlighting the program's positive aspects as if to promote the program. In the opening paragraph of the TBS Evaluation Policy (2001), it is stated that the "policy supports the generation of accurate and objective (...) information to help managers make sound, more effective decisions on their policies, programs and initiatives" (TBS, 2001), and that "objective evaluation is an important tool in helping managers to manage for results" (TBS, 2001). This requirement is supported by the CDC Guidelines for Program Evaluation in Public Health (1999), which in its checklist for ensuring effective evaluation reports, states that reports should be unbiased (CDC, 1999). Furthermore, Framework and Tools for Evaluating Health Surveillance Systems (2004) state that findings should provide information on both successes and deficiencies of the system (Health Canada, 2004).

Therefore, based on the evaluation report review findings, and public health program evaluation guidelines and TBS Evaluation Policy (2001), the issue of success should be taken into consideration when evaluating a program. Furthermore, a program's success should be evaluated and reported objectively and in an unbiased manner (CDC, 1999), showing both the successes and deficiencies of a program (Health Canada, 2004).

2.3. Cost-effectiveness

The TBS Evaluation Policy (2001) states that "departments should introduce the discipline of evaluation into programs', policies' and initiatives' lifecycle management in order to evaluate issues of cost-effectiveness" (TBS, 2001). Furthermore, as outlined by the policy, the evaluation questions that should be considered are whether "the most appropriate and efficient means are being used to achieve objectives, relative to alternative design and delivery approaches (cost-effectiveness)" (TBS, 2001). Public health program evaluation guidelines outline cost-effectiveness as one of the key evaluation issues to be considered when evaluating a program (CDC, 1999; Drummond et al., 1997; Health Canada, 1996; Health Canada 2004; Salaberria-Pena et al., 2007) as assessments of cost-effectiveness provide a reference point for relating costs to program results (Health Canada, 2004).

However, the evaluation report review findings demonstrated that not all evaluations took this issue into consideration when evaluating the programs, although one evaluation did provide a justification as to why this issue was not analyzed, and another evaluation was a formative assessment. However, the remaining evaluations did not provide a cost-effectiveness analysis assessment, nor did they explain why this issue was omitted from the evaluation. The ratings for the evaluations that assessed a program's cost-effectiveness varied as some evaluations implemented only one line of evidence into their cost-effectiveness assessment (table 9). The use of multiple lines of evidence was discussed previously in the text, and it was concluded that all sources of evidence should use multiple lines of evidence, wherever possible, to support the validity of the findings.

Therefore, based on the public health program evaluation guidelines and the TBS Evaluation Policy (2001) findings, cost-effectiveness evaluation should be a standing item of a program's (summative) assessment. Otherwise, a rationale should be provided as to why this particular issue was not evaluated.

2.4. Evidence-based findings

The TBS Evaluation Policy (2001) states that "evaluation findings should be relevant to the issues addressed and follow from the evidence" (TBS, 2001). The guidelines for reviewing evaluation reports presented in the Framework and Tools for Evaluating Health Surveillance Systems (2004) outline that "all significant findings are presented, testable, and do not go beyond what the evidence will support", and that "findings are substantiated by the evidence, as described in the evaluation report" (Health Canada, 2004). The findings from the evaluation report analysis in majority align with the outlined program evaluation in public health guidelines and TBS Evaluation Policy (2001) requirements, as seven of the evaluation reports were rated as 'more than adequate' and one report was assigned an 'adequate' score in relation to whether the findings were based on evidence drawn from the evaluation research.

The one report that was assigned an 'adequate' score had gaps in the evaluation findings, which were stressed throughout the report as an evaluation challenge and were assigned

to the difficulties in attribution and the changing and dynamic nature of the epidemic the program addressed. However, in situations such as the one found in that public health program evaluation, the CDC Framework on Program Evaluation in Public Health (1999) recommends that when a program's activities are aligned with those of other programs functioning in the same environment, and hence certain outcomes can not be attributed solely to a particular program, the evaluation should gather credible evidence that describes each program's contribution in the combined change effort (CDC, 1999).

3.0. Conclusions

The TBS Evaluation Policy (2001) requires that evaluation of a program, policy or initiative "must produce timely, pertinent and credible findings and conclusions that managers and other stakeholders can use with confidence, based on practical, cost-effective and objective data collection and analysis" (TBS, 2001). This requirement was not always met in the reviewed evaluations as the findings from the evaluation report analysis revealed that three evaluation reports did not present conclusions in their reports (the remaining reports were judged to be 'more than adequate', table 10).

The TBS Evaluation Policy (2001) requirements outlined above align with that of both the CDC Framework for Program Evaluation in Public Health (1999) and the Updated CDC Guidelines for Public Health Surveillance Systems (2001) which state in agreement that conclusions from the evaluation are justified through "appropriate standards, analysis and synthesis, interpretation, and judgement of the gathered evidence" (CDC, 1999; CDC, 2001). Furthermore, public health program evaluation guidelines outline that program stakeholders should agree that the conclusions drawn from the evaluation are justified before they will use findings from the evaluation with confidence, and that the gathered evidence should be linked to their relevant standards for assessing the program's performance (CDC, 1999; CDC, 2001; Health Canada, 2004). Therefore, based on the information outlined in the public health program evaluation guidelines, and the TBS Evaluation Policy (2001), it is vital for program evaluation reports to outline the conclusions made based on the evaluation findings.

Although the TBS Evaluation Policy (2001) does not specifically require a discussion of lessons learned during a program evaluation, public health program evaluation guidelines state that the process of communicating either the procedures or the lessons learned from an evaluation to relevant audiences may be included in program evaluation reports (CDC, 1999, CDC, 2001; Salaberria-Pena, 2007). The analysis of the eight evaluation reports revealed that four of them thoroughly discussed lessons learned; these lessons learned discussed the strengths and weaknesses of the evaluation, what worked, and what did not, and the gaps that remain to be filled. From the findings from the evaluation report analysis, it may be worth considering discussing an evaluation's lessons learned. By disseminating this information via evaluation reports, future public health program evaluation practices may be optimized.

4.0. Recommendations

Recommendations are actions for consideration which result from the evaluation (CDC, 1999).

Compliance with the TBS Evaluation Policy (2001) requirement is for all evaluation reports "to include clear and actionable recommendations" (TBS, 2001). Out of the eight evaluation reports review here, seven reports made formal recommendations (one report's recommendations were made by the expert advisory committee), whereas one evaluation report had recommendations which were dispersed throughout the findings section of the report. Furthermore, findings from the evaluation report analysis demonstrated that some evaluations resulted in making recommendations which were beyond the evaluated program's control or influence, making it difficult for the recommendations to be practical, operational and realistically attainable.

The CDC Guidelines for Program Evaluation in Public Health (1999) note that recommendations should take into account the organizational context in which the decisions relating the program will be made and the political sensitivities which users and stakeholders can influence and control (CDC, 1999). Therefore, based on the public health program evaluation guidelines, as well as the TBS Evaluation Policy (2001), all evaluations should not only make recommendations about the future of the program and ensure that such recommendations are operational, practical and realistically attainable.

The CDC Guidelines on Program Evaluation in Public Health (1999) recommend stating the advantages, disadvantages and resource implications the recommendations will have on the program (CDC, 1999). Out of the eight assessed evaluation reports, only one report's recommendations stated the impact the recommendations will have on the program; based on these findings it would be worth considering providing the implications the recommendations will have on the evaluated program.

Modification and continuation of a public health program should be addressed by the recommendations drawn from the evaluation. In certain cases, conclusions from the evaluation demonstrate that the most appropriate recommendation is to discontinue the program. However, this type of recommendation should be weighed carefully before being implemented as the cost of renewing a program that has been discontinued could be significantly greater than the cost of maintaining it. The stakeholders involved in the evaluation should consider all possible consequences relating public health and other areas before discontinuing a program (CDC, 2001). To support this analysis, the review of the evaluation reports revealed that no report presented evidence indicating that the program was not needed or not relevant to the Agency's mandate. Furthermore, the evaluation findings were accompanied by recommendations on implementing changes to the programs, however always in the context of the public health program being needed (TBS, 2004c).

The TBS Evaluation Policy (2001) requires evaluation reports to "present the conclusions and recommendations so that they flow logically from the evaluation findings" (TBS, 2001). This is supported by the public health program evaluation guidelines which state

the "recommended actions or decisions should be consistent with the conclusions" (CDC, 1999), and that "findings should be used to support decision-making" (Porteous et al., 1997). The reviewed evaluation reports did not always align with the policy requirement and public health evaluation guidelines. One report did not align its findings and conclusions with the proposed recommendations, making it unclear why the specified recommendations were being proposed. Another report had its recommendations consist of "considerations and recommendations", where the considerations were too long and detailed, followed by recommendations, which on the other hand, were too short. Another report substituted the conclusions section with a "gaps and lessons learned" section. Therefore, based on the evaluation report review findings, and public health program evaluation guidelines, as well as the TBS Evaluation Policy (2001), an evaluation's conclusions and recommendations should flow logically from the evaluation findings in order to clearly outline the rationale of the recommendations which form the basis for the future of the public health program under evaluation.

Summary of evaluation report findings

Overall, the set of reviewed public health program evaluations complied with public health evaluation guidelines and the TBS Evaluation Policy (2001). Usually, a few out of the eight evaluations were characterized with areas for improvement on each criterion – focusing on these weaknesses, and taking examples from other evaluations deemed 'more than adequate' on a given criterion, as well as aligning these limitations with the guidelines from public health evaluations and the TBS Evaluation Policy (2001), will strengthen the credibility of evaluation findings. In certain cases, when aligning evaluations with existing public health evaluation guidelines and the TBS Evaluation Policy (2001), evaluations identified criteria for implementation in future evaluations, e.g., by evaluating whether a program meets the needs of Canadians, whether a program duplicates with existing programs or outlining 'evaluation's lessons learned'. The areas for improvement or implementation in future evaluations are identified in a summary format in the 'recommendations' section of this report, where recommendations on how to avoid weaknesses or implement additional criteria in future evaluations are found.

The findings from the PHAC's evaluations presented here will help in carrying out effective evaluations, resulting in reporting higher quality assessments when requesting renewal of a program's terms and conditions (TBS, 2004b). The findings from this report review will advance credible evidence-base evaluation results supporting decision-making processes relating to a program's future - managers will be able to determine progress made toward planned objectives and measure a program's results, effects and impacts, whereas the lessons learned from the evaluation will be used to re-define program goals, priorities, and management practices (TBS, 2004b). Sound evidence-based evaluation practices "will lead to program plans that are clearer and more logical" (CDC, 2001) and will result in optimal use of public health resources (CDC, 2001).

RECOMMENDATIONS

Based on the findings from the review and assessment of PHAC's eight final program evaluation reports, and the review of public health program evaluation guidelines and the TBS Evaluation Policy (2001), the following is a proposed list of recommendations that may serve as guidelines for evaluators across the Agency in designing and conducting effective evaluations and delivering high quality evaluation reports. The recommendations are divided based on improving evaluation methodologies and results analysis, and improving evaluation reports.

Improving evaluation methodologies and results analysis

- Create evaluation questions that cover all issues/ requirements as recommended by public health program evaluation guidelines and the TBS Evaluation Policy (2001), e.g., program relevance (program relevance to government priorities, relevance to Canadians and program duplication), success, and cost-effectiveness, and if it is not feasible to do so, provide a rationale as to why the issue was not assessed;
- Include a representative sample of key informants, e.g., include participants that were clients and/or beneficiaries of a program as well as program managers, staff, and other persons responsible for implementing the program. Random-sample methodology may not be the optimal approach as it will result in not including all stakeholders;
- Include a group of participants that do not have a stake in the program as they will be a source of unbiased information and will contribute to a well-rounded perspective of the program (CDC, 1999);
  Ensure that the evaluation relies on more than one source of evidence and that it uses a combination of both quantitative and qualitative approaches, e.g., combining a data review, stakeholder survey, key informant interviews, research reports and document review analysis would be considered a good multiple line of evidence approach which combines both qualitative and quantitative data;
- Minimize bias and data reliability issues as much as possible, e.g., do not use an internal evaluator or rely on self-reporting (where applicable); ensure that information collected from documents relates to the evaluation questions, spans the appropriate evaluation period and is compatible in comparison; introduce a control group (where applicable);
- Minimize bias by reporting findings in an objective and neutral manner;
- Ensure that rigour is introduced into the evaluation, e.g., include a peer-panel review, conduct a pilot study on the developed evaluation tools, conduct a comparison of the findings with the literature and/or other organizations, consider using past surveys and involve stakeholders in tool development, use a third party for objective data collection and analysis;
- Interpret with caution findings that resulted from low participant response rates. When surveys, interviews or focus groups result in low participant response rates, ensure that appropriate caveats are made, such as whether care was taken

to not only report the percentage, but the number of participants, or whether the representation of responses across different geographic regions was fair.

Improving evaluation reports

- Outline key evaluation questions in the main body of report, not in a separate companion document as otherwise the reader has to flip back and forth between documents;
- Provide a matrix of evaluation question, issues, and corresponding indicators for clarity purposes;
- Provide key details of the methodology, e.g., duration of the interviews, number of respondents, methods of collecting data and data analysis in the methodology section of the report and not in the appendix or separate companion documents;
- Provide the methodology section after the introduction, otherwise the reader will have to keep on guessing throughout the report what the applied methods were;
  Provide a description of limitations and trade-offs, and suggestions on addressing the limitations, i.e., provide a description of biases and data reliability, for example, was a representative number of stakeholders included in the evaluation? Were internal evaluators used? Was the response rate high? How would the methodological approach address these limitations in future evaluations so that they are not repeated?;
- Include a 'lessons learned' section, where applicable, as it helps guide and optimize future public health program evaluations;
- Provide a formal recommendations section in the report. Providing scattered recommendations throughout the results section of the report is not considered sufficient;
- Ensure that conclusions are clearly outlined and precede recommendations – combining conclusions and recommendations in one paragraph is not advised;
- Ensure recommendations flow logically from findings and conclusions and make sure that recommendations address the significant evaluation findings and conclusions; by aligning findings and conclusions with the proposed recommendations, it is clear why the specified recommendations are being proposed and hence it is easier to understand which program areas are being targeted by the recommendations and why;
- Propose recommendations that are practical and realizable, i.e., recommendations that do not go beyond the areas of program's operations. For example stating "there is a need for a legislated body outside the government that could report candidly to the Prime Minister and the people of Canada on the implications of population and public health promotion across the country" would be considered beyond the program's influence;
- Provide in the recommendations the potential impact of a proposed recommendation on the program evaluated (to the extent possible), e.g., state "these recommendations will decrease the prevalence of the disease in Canada and future costs to the healthcare system" or "this recommendation would transcend programmatic, jurisdictional and disease-specific boundaries and

leverage scarce resources, promote a coordinated domestic approach as well as a global understanding of infectious diseases, by linking this disease with other levels of government, community organizations and key stakeholders."

EVALUATION REPORT REVIEW AND ANALYSIS LIMITATIONS

The primary limitation of the evaluation report review and analysis was that at the time of conducting the evaluations and writing the evaluation reports, not all evaluators had the evaluation report analysis template available, and hence they were unaware of the evaluation and report assessment criteria. Although the purpose of this project was not to assess the quality of evaluations and evaluation reports, but to determine the trends in the evaluations' strengths and weakness, this still may have impacted the findings. Perhaps if some of the evaluators were aware of the criteria present in the evaluation report assessment template, they would have had a different approach to the evaluation's certain aspects.

Another limitation is the number of evaluation reports reviewed. Due to the nature and scope of a master's project and its timelines, it was not feasible to review program evaluation reports from, for example, Health Canada. However, the objective of this project was to facilitate conducting program evaluation in the field of public health within the Canadian federal government, and therefore it must be stressed that all approved evaluation reports from the Public Health Agency of Canada were analyzed. However, perhaps if more evaluation reports had been reviewed, other noticeable trends would have emerged.

Furthermore, it must be brought to attention that not all aspects of program evaluation were taken into consideration in this analysis. For example, logic model design or program description was not discussed. This is due to the fact that the purpose of the evaluation report review was to assess the methodologies and approaches applied to program evaluation. Although actual evaluation report strengths and weaknesses were noted when analyzing the evaluations through the reports, this was not the main purpose of the analysis. Analyzing the introduction section of the reports was judged to be an analysis of reports themselves as written and perceived by the evaluators, and not an actual analysis of the designed and implemented evaluation. Therefore, information which was found in the introduction section of the reports, such as logic models and program description, did not fall within the scope of the analysis. Furthermore, as this was supposed to be an assessment of evaluation methodologies and approaches, not all evaluation report weaknesses were discussed, e.g., evaluation report limitations such as report and various report sections' length, or ineffective presentation of data in tables or graphs were not mentioned herein, although noted throughout the review.

Another limitation of the evaluation report review is that although great care was taken to assess the evaluations and evaluation reports in a rigorous manner, as outlined in the methodology section of this report, reviewer interpretation bias may have also impacted the findings. Furthermore, the reviewer was new to the field of public health program

evaluation, and perhaps a better trained eye would have picked up other evaluation aspects presented in the evaluation reports.

CONCLUSIONS

Programs must be evaluated in order to distinguish between programs that promote health and prevent disease, injury, or disability from those that do not carry out this function. Program evaluation is used for introducing change in planning effective public health actions, showing the findings of resource investments, and improving existing programs. It emphasizes the common objective of public health programs, and assesses whether the amount and scope of investment compliments the tasks to be accomplished (CDC, 1999).

"No approach has yet been invented that can be implemented without tailoring and modifying it to the current situation" (Francisco et al., 2000). Although there are quite a few detailed public health program evaluation guidelines, none of them address program evaluation in a Canadian federal government context. Therefore, the intention of this review and assessment of the undertaken program evaluations at the Public Health Agency of Canada was to facilitate guidance and direction, as well as strengthen the current state of program evaluation planning, design, data collection and analysis, making recommendations, and reporting in public health within a Canadian context. This analysis should be treated as a broad generic overview of the existing public health program evaluation guidelines, as well as evaluation requirements for the Canadian federal government, and a potential standard for further improvement. Furthermore, it must be noted that this analysis did not cover every aspect of program evaluation, as it was based on a sample of eight program evaluations, and other potential aspects which were not part of this analysis should be taken into consideration when evaluating a program. However, the final recommendations which are a final summary of the analysis of the PHAC's evaluation reports to date may be used to facilitate creating guidelines or tool kits for program evaluators across the Agency or even other Canadian federal government health departments in order to improve and account for health promotion and disease prevention work, and hence improve public health for all. Ultimately, this may potentially facilitate the Public Health Agency of Canada in achieving its vision: healthy Canadians and communities in a healthier world (PHAC, 2008c).

REFERENCES

Centre for Disease Control and Prevention (2001). Updated Guidelines for Evaluating Public Health Surveillance Systems. *Morbidity & Mortality Weekly Report,* 50(RR13); 1-35. Retrieved November 15, 2008 from: http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm

Centre for Disease Control and Prevention (1999). Framework for Program Evaluation in Public Health. *Morbidity & Mortality Weekly Report*, 48, No. RR-11, 1-40. Retrieved June 26, 2008 from: http://www.cdc.gov/mmwr/PDF/rr/rr4811.pdf

Davis, M. (2006). Teaching Practical Public Health Methods. *Am J Eval* 27(2):247-256

Drummond M, Jefferson T (1996). Guidelines for authors and peer-reviewers of economic submissions to the BMJ. *BMJ* 313:275-283

Drummond, M.; Sculpher, M.; Torrance, G.; O'Brien, B.; Stoddart, G. (1997). Methods for the economic evaluation of health care programmes. 2nd ed. Oxford. Oxford University Press.

Francisco, V.; Capwell, E.; Butterfoss, F. (2000). Getting Off to a Good Start With Your Evaluation. *Health Promot Pract* 1(2):126-131.

Health Canada (2004). Frameworks and Tools for Evaluating Health Surveillance Systems. Retrieved June 26, 2008 from: http://www.phac-aspc.gc.ca/php-psp/pdf/i_Surveillance_Evaluation_Framework_v1.pdf

Health Canada (1996). Guide to Project Evaluation: A Population Health Approach. Retrieved November 15, 2008 from: http://www.phac-aspc.gc.ca/ncfv-cnivf/familyviolence/html/fvprojevaluation_e.html

Kelly, C.; Moehner, C.; Baker, E.; Brennan Ramirez, L.; Brownson, R. (2006). Promoting physical activity in communities: Approaches for successful evaluation of programs and policies. *Eval Program Plann* 29: 280-292.

Laferty, C. and Mahoney, C. (2003) A Framework for Evaluating Comprehensive Community Initiatives. *Health Promot Pract* 4 (1):31-44

Milstein, B.; Wetterhall, S.; CDC Evaluation Working Group (2000). A Framework Featuring Steps and Standards for Program Evaluation. *Health Promot Pract* 1(3): 221-228

Porteous, N.; Sheldrick, B.; Stewart, P. (1999). Enhancing Managers' Evaluation Capacity: a Case Study from Ontario Health Units. *Cad. J. of Pr. Eval.* 14: 1-18.

Porteous, N.; Sheldrick, B; Stewart, P. (1997) Program Evaluation Toolkit: A Blueprint for Public Health Management. Public Health Research, Education and Development Program, Ottawa-Carleton Health Department. Retrieved June 26, 2008 from: http://www.phac-aspc.gc.ca/php-psp/toolkit-eng.php

Public Health Agency of Canada (2008a). Retrieved January 17, 2009 from: http://www.phac-aspc.gc.ca/about_apropos/evaluation-eng.php

Public Health Agency of Canada (2008b). Evaluation Plan 2008-2009 to 2012-2013. Commitments to the Treasury Board of Canada Secretariat. Retrieved January 17, 2009 from: http://www.phac-aspc.gc.ca/about_apropos/evaluation_plan08-eng.php

Public Health Agency of Canada (2008c). About the Agency. Retrieved November 29, 2008 from: http://www.phac-aspc.gc.ca/about_apropos/index-eng.php

Rootman, I.; Michael Goodstadt, M.; World Health Organization Regional Office for Europe (2001). Evaluation in Health Promotion: Principles and Perspectives. WHO Regional Office Europe, No.92.

Salabarría-Peña, Y, Apt, B.; Walsh, C. (2007). *Practical Use of Program Evaluation among Sexually Transmitted Disease (STD) Programs*, Atlanta (GA): Centers for Disease Control and Prevention. Retrieved June 26, 2008 from: http://www.cdc.gov/std/program/pupestd/Introduction.pdf

Sekhobo, J. and Druschel, C. (2001). An Evaluation of Congenital Malformations Surveillance in New York State: An Application of Centers for Disease Control and Prevention (CDC) Guidelines for Evaluating Surveillance Systems. Public Health Reports. 116: 296-305

Treasury Board of Canada Secretariat (2001). Evaluation Policy. Retrieved May 30, 2008 from: http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=12309

Treasury Board of Canada Secretariat (2004a). Evaluation Function in the Government of Canada. Retrieved June 30, 2008 from: http://www.tbs-sct.gc.ca/eval/pubs/func-fonc/func-fonc_e.asp

Treasury Board of Canada Secretariat (2004b). Guide for the Review of Evaluation Reports. Retrieved June 30, 2008 from: http://www.tbs-sct.gc.ca/eval/tools_outils/4001752_e.asp

Treasury Board of Canada Secretariat (2004c). Review of the Quality of Evaluations Across Departments and Agencies. Retrieved June 30, 2008 from: http://www.tbs-sct.gc.ca/eval/pubs/rev-exam_e.asp

Treasury Board of Canada Secretariat (2005). The Health of the Evaluation Function in the Government of Canada. Report for Fiscal Year 2004-05. Retrieved May 15, 2008 from: http://www.tbs-sct.gc.ca/eval/dev/health-santé/pdf/hefgc-sfegc_e.pdf

Treasury Board of Canada Secretariat (n.d). Program Evaluation Methods: Measurement and Attribution of Program Results. 3rd Ed. Retrieved May 15, from: http://www.tbs-sct.gc.ca/eval/pubs/meth/pem-mep_e.pdf

APPENDIX 1: EVALUATION REPORT REVIEW TEMPLATE (TBS, 2004c)

Report Title	
Type of Report	
	o Formative Evaluation
	o Summative Evaluation
	o Other:
Date of Report	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
METHODOLOGY						
1.1. Description of the Methodology/ Design	1.1.1. Describes logical, valid, evidence-based methodologies that are linked to the evaluation issues explored OR there is a clear reference to a technical document for this information	Describes the methodologies and design applied to the evaluation	☐ Yes - describes ☐ Yes - only lists a few details ☐ No – reference to technical documents ☐ No – no reference to technical documents	☐ sample size ☐ sample method ☐ instruments ☐ links methods to issues ☐ reference to technical documents	Poor Adequate More than adequate N/A	
1.1. Description of the Methodology/ Design (continued)	1.1.2. The evaluation issues and questions are adequately addressed		☐ Yes ☐ No		Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Methodology (continued)						
1.2. Multiple Lines of Evidence	1.2.1. The evaluation contains multiple lines of evidence to support the validity of the findings	The evaluation relies on more than one line of evidence to support its findings (triangulation of results)		☐ focus group ☐ observational study ☐ key informant interviews ☐ census ☐ sample survey ☐ literature review ☐ document review ☐ file review ☐ secondary data analysis ☐ database review ☐ analysis of performance data ☐ case studies ☐ cost-benefit analysis ☐ other	Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Methodology (continued)						
1.2. Multiple Lines of Evidence (continued)	1.2.2. Is there an appropriate balance between qualitative and quantitative methodologies?		☐ Yes ☐ No			
	1.2.3. All stakeholder perspectives are included			☐ clients/ beneficiaries ☐ program management and provider (federal government) ☐ third-party deliverers ☐ partners ☐ experts ☐ funding recipients ☐ non-recipients ☐ other	Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Methodology (continued)						
1.2. Multiple Lines of Evidence (continued)		Qualitative evidence drawn from key informants who do not have a stake in the program	☐ Yes ☐ No ☐ Unable to assess			
1.3. Limitations	1.3.1. The limitations and trade-offs of the methodologies, data sources and data used in the evaluation are clearly articulated	Limitations are described: actual and potential biases, reliability of data are identified and explained in terms of their impact on stated findings	☐ Yes ☐ No ☐ No apparent limitations	☐ biases described ☐ data reliability explained	Poor Adequate More than adequate N/A	
		The constraints of the evaluation are made clear	☐ Yes ☐ No ☐ No apparent constraints	☐ budget ☐ time ☐ data availability ☐ other (specify)	Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Methodology (continued)						
1.4 Rigour	1.4.1. A comparison "point" exists	Survey of representative group of participants	☐ Yes ☐ No			
		Comparison group	☐ Yes ☐ No			
		Comparison to baseline measures/ norms/ literature/ other benchmark	☐ Yes ☐ No			
	1.4.2. Instruments were pre-tested prior to evaluation (pilot study) and/or consultations with specialists were conducted to develop instruments		☐ Yes ☐ No			

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
RESULTS						
2.1. Relevance	2.1.1. Presents findings related to establishing continued relevance and contribution to results achievement by linking results to societal need and government priority areas	Evidence to demonstrate actual need and/ or evidence to demonstrate responsiveness to need	☐ Yes ☐ No		Poor Adequate More than adequate N/A	
		Evidence to demonstrate continued relevance to government priorities	☐ Yes ☐ No		Poor Adequate More than adequate N/A	
		Evidence to demonstrate that it does not duplicate or work at cross purposes with other programs, policies, or initiatives	☐ Yes ☐ No		Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Results (continued)						
2.2. Success	2.2.1. Presents findings demonstrating whether or not the program, policy or initiative is producing results that support its continuation or renewal	Clearly describes what has happened as a result of the program and articulates attribution of program, policy or initiative to success	□ Yes □ No		Poor Adequate More than adequate N/A	
2.3. Cost-effectiveness	2.3.1. Cost-effectiveness is assessed: › qualitative assessment of cost-effectiveness and/or › quantitative assessment of cost-effectiveness		□ Yes □ No	□ qualitative assessment □ quantitative assessment	Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Results (continued)						
2.4. Evidence-based Findings	2.4.1. The findings are based on evidence drawn from the evaluation research	Demonstrates that the findings flow logically from the interpretation of the data and analyses	☐ Yes ☐ No		Poor Adequate More than adequate N/A	
CONCLUSIONS						
3.1. Presents clear, impartial and accurate evidence-based conclusions		Conclusions objectively answer the evaluation issues and are supported by the findings			Poor Adequate More than adequate N/A	
		Presents lessons learned about the program from the evaluation	☐ Yes ☐ No		Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
RECOMMENDATIONS						
4.1. Recommendations are present	☐ Yes – formal recommendations ☐ Yes – suggestions that are not called "recommendations" ☐ No					
4.2. Clearly states practical and realizable recommendations	4.2.1. Recommendations are operational, practical and realistically attainable				Poor Adequate More than adequate N/A	
	4.2.2. Identifies alternative scenarios (e.g., potential impact of recommendations on program) and takes into account practical constraints				Poor Adequate More than adequate N/A	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating	Qualitative Assessment
Recommendations (continued)						
4.3. Recommendations are supported by and flow logically from findings and conclusions		Recommendations address evaluation findings	☐ Yes ☐ No		Poor Adequate More than adequate N/A	
		Recommendations flow logically from findings and conclusions	☐ Yes ☐ No		Poor Adequate More than adequate N/A	