Epidemiology and Infection

www.cambridge.org/hyg

Converting exposure proportions calculated from surveys based on one reference period length to another

Matt Hurst

Public Health Agency of Canada, Canada

Original Paper

Cite this article: Hurst M (2026). Converting exposure proportions calculated from surveys based on one reference period length to another. *Epidemiology and Infection*, **154**, e86, 1–8
https://doi.org/10.1017/S0950268826101745

Received: 09 December 2025 UTC
Revised: 18 March 2026 UTC
Accepted: 15 May 2026 UTC

Keywords:
Survey analysis; Reference periods; Case control; Converting proportions; Memory bias

Corresponding author:
Matt Hurst;
Email: matt.hurst@phac-aspc.gc.ca

Abstract

Commonly used components of health risk assessments and outbreak investigations are the proportion of the general population exposed to potential sources, often coming from national surveys collected with one reference period, such as 7 days. However, risk assessors may need this value for a different period of time (e.g. 1 day, a full year, etc.) and outbreak investigators may need a reference period that matches their case questionnaire. Use of these proportions is biased if the reference period does not match what is needed. Three surveys that collected exposure information with more than one reference group were used to fit six models that were developed to estimate the relationship between proportion and reference period length, and thus, convert proportions collected with one reference period to any desired reference period. A practical implementation of the best model is provided in a spreadsheet for health practitioners, such as those working in risk assessment or outbreaks who need to convert proportions to a new reference period. The conversion method can be used in any subject matter across the sciences, where surveys collect the proportion of respondents who experienced an event in a defined reference period.

Introduction

Enteric disease public health officials collect exposure information for a variety of purposes. Case questionnaires collect exposures to various food, animal, and environmental sources to pinpoint the origin of illness. The reference period for recall in these instruments varies with the maximum incubation period of the disease. These can range from 1 day to up to 25 days or more for some pathogens. Often, the next step in considering the importance of a particular exposure is to compare the case exposure proportion to one from the healthy population (specifically this proportion is a prevalence and we use the terms interchangeably). The source of this control information is usually large surveys, which, due to resource constraints, ask the exposure question for one reference period only. Seven days is often chosen, as it is close to the maximum incubation period of common foodborne pathogens, and is the period used for the Foodbook Study, Canada's national enteric exposure survey [1]. When the case and control reference periods do not match, epidemiologists are at a disadvantage because the comparisons are biased. A difference in proportions (or a case–control odds ratio), with all other things being equal including these reference periods, would be equal to zero (or one) if the exposure has no effect. However, if case questions are collected with a longer reference period than the controls, their proportions will be larger than they otherwise would be because the period of exposure is longer, biasing the comparisons to look more protective than they actually are.

Furthermore, risk assessment work often requires the reference period of the exposure proportion to be on a per day basis or a per year basis, which can require a conversion. For example, in a Canadian comparative exposure assessment of salmonella, to align the Canadian Community Health Survey estimates of the number of servings and the serving size in grams consumed in the past 24 h to the Foodbook Study's proportion of the population eating the food item in the past 7 days; the Foodbook results were converted from a weekly to a daily basis [2]. A Linear method of conversion was used in these studies, which is biased and is described in detail later in this document.

The current study considered six models to convert proportions. Four were chosen based on the literature and two were original works. The objective of this paper was to propose the most appropriate method for converting food consumption prevalences from one period to another, by assessing these models for their ability to fit the data, predict long period exposures, their simplicity of implementation in spreadsheet software, and their robustness to the use of parameters fit in other datasets. The conversion methods can be used in any subject matter where surveys collect the proportion of respondents who experienced an event at least once in a defined reference period.

© Crown Copyright - Government of Canada, 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

Methods

Data sources

Two of the data files analysed were collected by the Public Health Agency of Canada (PHAC). The Foodbook Study involved a national survey that asked questions on food, animal, and water exposures in 2014–2015 [1]. Respondents were asked exposure questions with a reference period of 7 days, with a separate group asked the same questions though with a reference period of 3 days. The Healthy Control Survey was also conducted by PHAC by FoodNet Canada (then C-Enternet), the integrated sentinel site surveillance network for enteric disease in Canada, in the pilot sentinel site in the Region of Waterloo. It asked questions about food exposures with three separate reference periods: 14, 7, and 3 days [3]. Also examined is the United States Population Survey collected in 2000–2001 by the Foodborne Diseases Active Surveillance Network programme in the Centres for Disease Control [4], which had a number of food exposure questions using both a 7 day and a month long recall period. A month was coded as 30 days.

In Table 1 of David et al., 2014, the columns showing observed proportions for the three reference periods of the Healthy Control Survey provide a concrete example of exposure proportions collected for different reference periods. Further details on the surveys, such as sample sizes and the questions used are in Supplementary Materials S2.

Models

Linear model

One method used to adjust for differing reference periods is a simple linear extrapolation. This involves dividing the prevalence estimate by the number of days of the reference period to obtain an estimate for 1 day (1). Predicting the prevalence for another reference period was done by simply multiplying the new 1 day estimate by the length in days of the desired reference period. It has been used in other Canadian work [2, 5], and elsewhere [6].

$$P_1 = P_x/x \tag{1}$$

where, P_x: the probability of ever being exposed in x number of days; P_1: the probability of ever being exposed in 1 day, and x: the length of the reference period in days (the number of days of exposure).

Geometric model I

Most enteric disease exposure proportions estimated from survey data can represent the probability of being exposed for the reference period in which it was collected. Some cannot, such as fixed daily exposures in which the same proportion of the population would say they are exposed, regardless of the reference period length. Examples include drinking water from a well on the property. For exposures that are not constant, a probability model is appropriate. Here, exposure probabilities represent the likelihood of 'ever being exposed' in the reference period. Examples include survey questions posed as 'have you eaten chicken in the last seven days', or 'have you had contact with your pet dog in the last seven days'. This was exploited mathematically. Quantitative microbial risk assessment guidelines have recommended this formula [7]. Treating the proportions as probabilities, the relationship between P_1 and P_x can be derived as follows:

$$\begin{aligned} P_x &= \text{observed probability of ever being exposed in x days} \\ &= 1 - \text{probability of never being exposed in x days} \\ &= 1 - (\text{probability of never being exposed in 1 day})^x \\ &= 1 - (1 - \text{probability of being exposed in 1 day})^x \\ &= 1 - (1 - P_1)^x \end{aligned} \tag{2}$$

Step three in the development of (2) assumed that the probability of never being exposed in a given day is the same for each time period. Equation (2) is the cumulative distribution function (CDF) of the geometric distribution, hence the name. Solving it for P_1 gave:

$$P_1 = 1 - (1 - P_x)^{(1/x)} \tag{3}$$

Table 1. Model results, untransformed parameter estimates, and indicators of model fit

	Model					
Dataset	Linear	Geometric I	Geometric II	Geometric III	Exponential	Weibull
Healthy Control						
Parameters			b: 0.93217122	b: 0.81946034	λ: 0.48396296	λ: 0.50481443, k: 0.90847478
RMSD	0.12290016	0.05432026	0.02375460	0.02339113	0.03148535	0.03143726
BIC	6,118	4,829	4,638	4,637	4,693	4,705
Foodbook						
Parameters			b: 0.87651122	b: 0.67128415	λ: 0.62528587	λ: 0.68036026, k: 0.82171348
RMSD	0.10996468	0.06174804	0.02029058	0.02052077	0.02423706	0.02423006
BIC	539,947	532,825	528,992	529,028	529,604	529,604
US Population						
Parameters			b: 0.96309458	b: 0.89564766	λ: 0.17422261	λ: 0.17567132, k: 0.69393968
RMSD	0.08823380	0.04447279	0.01037904	0.01060089	0.01865796	0.01865801
BIC	9,327	546	−3,391	−3,378	−2,773	−2,773

Note: Because the conversion formulas for many models use high order exponents, eight significant digits are provided and recommended in any implementation to maintain numerical stability.
BIC, Bayesian information criterion; RMSD, root mean squared deviation.

.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

With an estimate P_1 obtained, a prediction for any reference period was found using (2).

Geometric model II

Expanding on Geometric model I, we added a second equation and with it, a term that muted the growth of the predicted exposure proportion as the number of reference days increased. We dubbed this new parameter a 'decay' factor.

The new equation was:

$$P_1 = P_{1c}b^x \tag{4}$$

where, b: the decay factor; P_{1c}: 1 day exposure probability with the decay factor removed.

Combining (2) and (4) gave us:

$$P_x = 1 - (1 - P_{1c}b^x)^x \tag{5}$$

A worked example of this model is available (Supplementary Material S1).

Geometric model III

Another model we tested implemented the decay parameter b in a different manner. Each term in brackets in (6) represents the probability of not being exposed for each progressive day in the reference period. The decay factor becomes cumulatively stronger for each additional day as represented by its exponent increasing by one each day. This function flattens to an asymptote (The asymptote for Geometric model III can be found exactly using a Q-Pochhammer infinite product which is available in symbolic computation software products. For example, using the Q-Pochhammer function at www.wolframalpha.com will give the asymptote for P_{1c} = 0.0837 and b_{fit} = 0.819 using the input '1–Q-Pochhammer[0.0837, 0.819]/(1–0.0837)' which equals about 0.320. Practically, this asymptote can be approximated from spreadsheet predictions of long reference periods of 20 to 30 days for many common values of P_{1c} and b_{fit}). We hypothesized that this might fit the data more closely than Geometric model II.

$$P_x = 1 - (1 - P_{1c}b)(1 - P_{1c}b^2)(1 - P_{1c}b^3)\cdots(1 - P_{1c}b^x) \tag{6}$$

Geometric models II and III were original functions that we developed for this work and we are not aware of their use by other authors.

Exponential and Weibull

The literature also suggested an exponential distribution [8] for modelling the proportion of events remembered to the time between the event and the interview. This function form did not fit our observed data, however, the exponential CDF with some adjustment, did (7). We also tried the more flexible Weibull CDF, as it has an additional shape parameter that might better fit the data (8).

The Exponential model should fit the data identically to the Geometric model given it is its continuous analogue. It asymptotes to 1 as the reference period lengthens, which did not fit our observed data. To address this issue, we added a multiplicative shape parameter to both the Exponential and Weibull models that forced them to asymptote to any value in the interval between 0 and 1. These updated models fit the observed data more closely.

The exponential CDF function implementation was:

$$P_x = a\left(1 - e^{-\lambda x}\right) \tag{7}$$

where, a: a shape parameter that represents the limiting probability of exposure as the reference period goes to infinity and is thus the maximum proportion of the population that was exposed; λ: a shape parameter. The function increased faster and also levelled off faster with larger values of λ.

The Weibull CDF function implementation was:

$$P_x = a\left(1 - e^{-(\lambda x)^k}\right) \tag{8}$$

where, k: a shape parameter that affects concavity up to the first inflexion point as well as the shape of the curve after this point.

Assumptions and data preparation

We assumed that the exposures were average behaviours for the population on the whole, and thus there was no correlation between being exposed to a particular food, for instance, from one day to the next. Some individuals may not have eaten the same food from one day to the next and thus there may have been some negative correlation, though there may also have been people who did eat the same food day after day, adding positive correlation. We assumed these groups balance out, and for the population as a whole, we felt that the assumption of no correlation was reasonable. As well, we assumed that at the population level, the decay factor b is the same for each of the questions in our survey.

Case questionnaires and national control survey questionnaires have similar biases in the exposure proportions they estimate. The questionnaire flow and question wording within them are designed to be as identical as possible, so that any remaining biases are the same between the two instruments. In fact, this is a key assumption of case control analysis: when identical proportions are found in each tool (collected with the same reference period), and an odds ratio of 1 is calculated, an assessment of no association follows. The models assessed here are assumed to not change these biases. This is reasonable because they were fit to the control proportions that already contain them. Thus, the adjusted proportions from these models have the same biases as those from the case questionnaire and thus can be compared with them in the usual way. The potential introduction of new biases from the models presented here are discussed in the limitations section.

For each survey, questions were removed from the analysis if they represented constant exposures that were considered very likely to be identical from one reference period to the next. For example, having a private well as the main drinking water source in the home should be practically identical for reference periods in the time span of months, and certainly weeks. Also, questions in which the observed exposure proportions with smaller reference periods were larger than those of longer reference periods were removed as this was not logically consistent with longer exposure periods generating more exposure and this situation was likely do to sampling variation causing estimates to be too far off of their true value.

Estimation

Non-linear least squares was used in Stata [9] using evaluator functions and the nlsur command to model the systems of equations. The coefficients for the b term and the 1 day exposure proportion P_1, P_{1c} estimated by the models were constrained to be between zero and one using a logit transform. Probability weights were used in the Foodbook data models to ensure the sample with a 3 day reference period was given the same weight as the 7 day sample.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

The convergence criterion was adjusted from its default value to 1E-10 to improve model fit. The models were also run in MS Excel [10] using the solver add-on as a secondary method to verify optimal fit and check for coding errors.

We estimated parameters, P_1, P_{1c}, a, b, λ, and k, that best fit the proportions available for the different reference periods (two or three depending on the survey) and the questions analysed. P_1, P_{1c}, and a were question-specific parameters and so differed by question. The remaining parameters, b, λ, and k, were fit to the whole dataset for the survey and thus were not question specific.

The exposure questions were answered twice by separate groups of respondents, one for each reference period, generating two equations to fit for Foodbook. In the US Population Survey, there were also two equations. Though in this survey, it was the group of respondents who answered yes to having an exposure in the past month who were subsequently asked about having a shorter, 7 day exposure so the respondent groups were not independent. The Healthy Control Survey interviewed three separate groups, one for each reference period for which the exposure question is asked, generating three equations to fit.

Taking the Healthy Control Survey and Geometric model II as an example, we wrote out (5) for the three reference periods in this survey, which resulted in a system of equations:

$$\begin{aligned} P_{x1} &= 1 - \left(1 - P_{1c}b^{x1}\right)^{x1} \\ P_{x2} &= 1 - \left(1 - P_{1c}b^{x2}\right)^{x2} \\ P_{x3} &= 1 - \left(1 - P_{1c}b^{x3}\right)^{x3} \end{aligned} \tag{9}$$

The fitted b and P_{1c} coefficients, denoted b_{fit} and $P_{1c,fit}$, were chosen to minimize the residuals across all three equations and exposure questions within the survey. To convert an observed P_x collected from a survey with a reference period of x days to one with xnew days in the reference period using the fitted values for the parameters, the equation was:

$$P_{xnew} = 1 - \left(1 - P_{1c,fit}b_{fit}^{xnew}\right)^{xnew} \tag{10}$$

For Geometric model III, the equation was:

$$P_{xnew} = 1 - \Pi_{i=1}^{xnew}\left(p_{1c,fit}b_{fit}^{i}\right) \tag{11}$$

Model evaluation criteria

To choose the best model, four criteria were considered: (1) model fit, (2) ability to predict long exposures in excess of 20 days, (3) ease of implementation in spreadsheet software, and (4) how well the model fits with alternative parameters from the other datasets.

Model fit was evaluated using the root mean square deviation (RMSD). RMSD is defined as $\sqrt{\sum_{x=1}^{nr}(Pxnew - Px)^2/nr}$, where n is the number of questions on the survey and r is the number of reference periods for which each question was asked (producing nr predicted proportions and observed proportions). As an example, the Healthy Control Survey had nr = 24*3 = 72 proportions used in the RMSD calculation. The Bayesian information criterion (BIC) was also used to assess model fit. A drop in the BIC of more than 10 is considered very strong evidence that the model fits the same dataset better [11]. A smaller RMSD implies better model fit. Since the outcome variable is on the same scale, zero to one, in all three datasets, comparing the RMSD across datasets is valid.

The robustness of the model fit to parameters from other datasets is important as the fitted parameters (b, λ, k) are not identical across datasets. To recommend one model over another for practitioners to use in the field with datasets collected with one reference period only, it is desirable that this model be the least sensitive to shape parameter misspecification. A RSMD score was developed to measure the increase in RMSD when a parameter from another dataset is used. These increases were tabulated for each of the six dataset–parameter source combinations for the Exponential model and Geometric model III. We then compared the difference in this score between Geometric model III and the Exponential model where the model with the lower score was considered to have performed better.

Results

Model fit

The six models were fit to the observed data from the three survey (Table 1). The Linear model did not fit the data well. Geometric model I fit roughly twice as well in terms of RMSD versus the Linear model and Geometric model II fit the data a further three times better or more for the Foodbook and US food Population Surveys. BIC and RMSD values indicated that Geometric model II fit the observed data the best in all three datasets, with Geometric model III a close second. The curves for these two models were nearly identical over the range of the reference periods to which they were fit (3, 7, and 14 days) using the Healthy Control Survey dataset (Figure 1). Confidence intervals for the fitted model parameters b and λ also suggest that the Geometric model II and Geometric model III fit the data reasonably well (see Supplementary Materials S3 for select models). All coefficients for each model are available in Supplementary Materials S4.

Robustness to parameter misspecification

The robustness of Geometric model III and the Exponential model fit to the use of parameters obtained from fitting the model to other datasets is available in Table 2. The increase in RMSD from Geometric model III using Foodbook data with the parameter estimated from this dataset (the reference) to one using Foodbook data and the parameter estimated from the Healthy Control Survey was

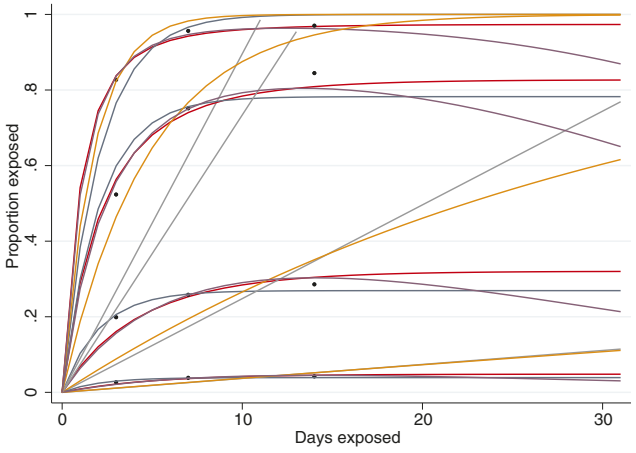


Figure 1. Five models fit to the Healthy Control dataset. Black dots are the observed proportions from four select questions (from top to bottom: q22a, q49, q20, q1_4) from the survey for the 3 day, 7 day, and 14 day reference periods. The curves are the fitted model equations for: Linear (light grey), Geometric I (orange), Geometric II (purple), Geometric III (red), and Exponential (dark grey).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

Table 2. Root mean square deviation of Geometric III and Exponential models for each dataset using parameters derived from those and other datasets

Dataset used	Parameter source			RMSD score	Score difference
	Foodbook	Healthy Control	USA Population		
Geometric III					
Foodbook	0.02052077	0.03027964	0.04223629	0.03147439	
Healthy Control	0.03156424	0.02339113	0.03109370	0.01587568	
US Population	0.03476763	0.02141165	0.01060090	0.03497748	
Exponential					
Foodbook	0.02423706	0.02692632	0.08923876	0.06769096	0.03621657
Healthy Control	0.03565337	0.03148535	0.07150202	0.04418469	0.02830901
US Population	0.03742245	0.03537497	0.01865797	0.03548148	0.00050400
Total					0.06502958

Note that the reference RMSD is along the diagonal where the dataset and parameter source are the same.
RMSD, root mean squared deviation.

0.03027964–0.02052077. The RMSD increase from the reference to the one using Foodbook data and the US Population parameter was 0.04223629–0.02052077. Adding these two differences together yielded an RMSD score of 0.03147439 for the Geometric model III using Foodbook data. Continuing in this way, the RMSD score for the Exponential and Geometric III models were calculated for each dataset. The score difference was the RMSD score for the Exponential model minus Geometric model III for the Foodbook dataset equal to 0.06769096–0.03147439 = 0.03621657. The total is the sum of these score differences.

A select number of exposure proportions for each dataset along with the fitted curves are provided for the two finalist models (Figure 2). The proportions were chosen to provide a broad range of values from under 0.05 to close to 0.9, where possible.

Discussion

Model fit

The Linear method has been used by practitioners, so it bears close scrutiny. It has the advantage of simplicity, but has a strong disadvantage for being inaccurate when all but small prevalences are converted. This error is explained using the prevalence of cheese consumption as an example.

From the Foodbook survey, the prevalence of eating cheese in a 7 day reference period was 88% [1]. Using the Linear method, one would calculate the per day prevalence as 88/7 = 13%. However, this is unreasonably low compared with the prevalence of 58% per day from the Canadian Community Health Survey [12]. Conversely, if converting in the opposite direction, from a smaller to a larger reference period, an 80% probability for 7 days would be converted to 160% per 14 days, which is not possible since prevalences cannot be higher than 100%. These examples make it clear there are very important drawbacks to the Linear model. The approximation can work in a pinch for small proportions of roughly 10% or less (see q1_4 in Figure 1) that are not extrapolated too far into the future, but it grows progressively worse for proportions larger than this. For instance, for pathogens with short incubation periods where a 3 day reference period might be needed, the comparable predicted value from the control data using the Linear model was usually less than half of the 3 day observed proportion values when the 7 day proportions were large; whereas the predictions from the other models were much closer to the 3 day observed proportions. Thus, for large 7 day proportions, 3 day values predicted from the Linear model were very poor. With these clear drawbacks and poor model fit (Table 1), the Linear method was excluded from further consideration.

Geometric model I was also a poor performer (Table 1). Though its predictions were always constrained between 0 and 1, it did not fit the data even a third as well as Geometric model II and Geometric model III in terms of RMSD. A visual inspection of Figure 1 gives an example of how the fitted curves for Geometric model I struggle to come near the datapoints for the two exposure questions in the top portion of the figure. Because of this, Geometric model I was also removed from further consideration.

The Weibull model fits the data almost identically to the Exponential model. Since the Exponential model has fewer parameters, the Weibull model was excluded from further consideration.

Long-term extrapolation

The Linear model and Geometric model II cannot perform long-term extrapolation. The Linear model can produce predictions larger than one for a proportion/prevalence, which is impossible. Geometric model II extrapolations beyond the largest reference period in the dataset decayed towards zero, which is not reasonable for the behaviour being modelled. Thus Geometric model II was excluded from further consideration.

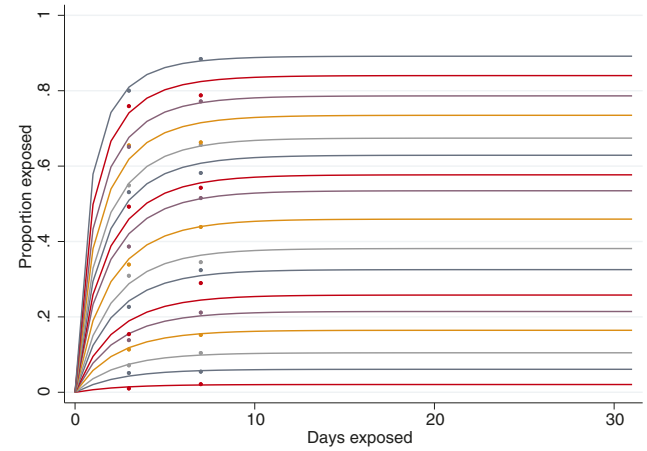
Spreadsheet software implementation

Of the remaining two models, the Exponential model was easier to implement in spreadsheet software. The model equation is straightforward and predictions for any reference day can be made easily. Geometric model III was more complicated to implement as it requires a bank of cells dedicated to calculating the predications. Though the Exponential model performed better on this criteria, it was not considered important enough to exclude Geometric model III based on this alone as it did clearly fit the data better than the Exponential model in terms of RMSD and BIC.

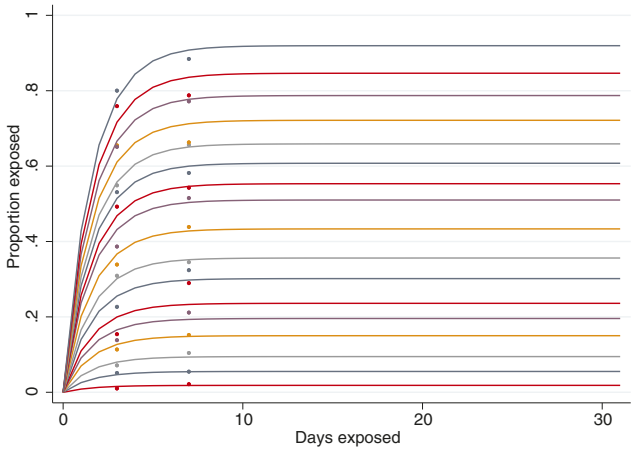
Robustness to parameter misspecification

How well the model can fit using parameters found from other datasets is also important. Some models may be more robust to parameter misspecification. To examine this, the RMSD from

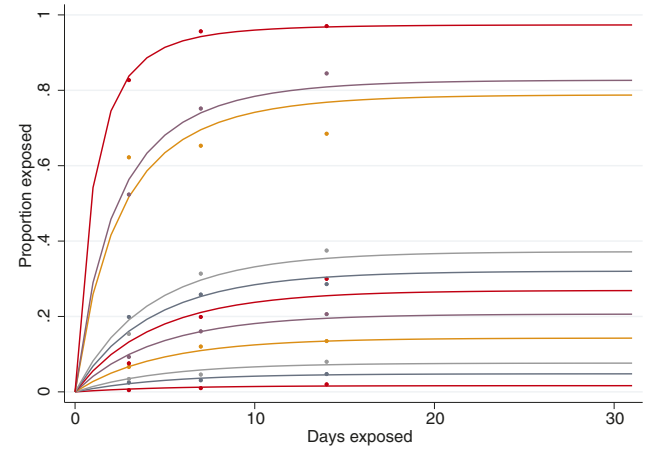
Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745



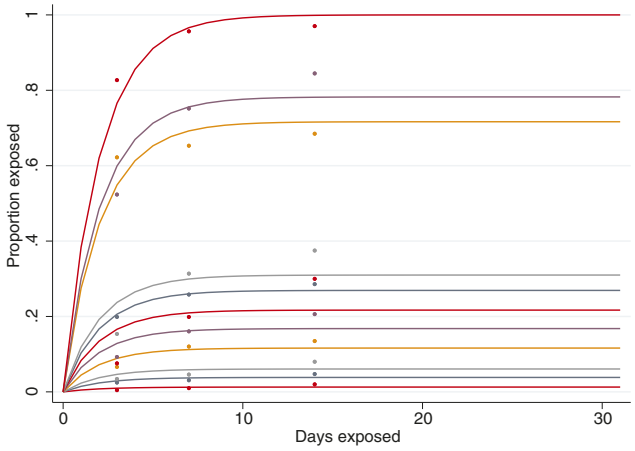
A1. Foodbook, Geometric modesl III.



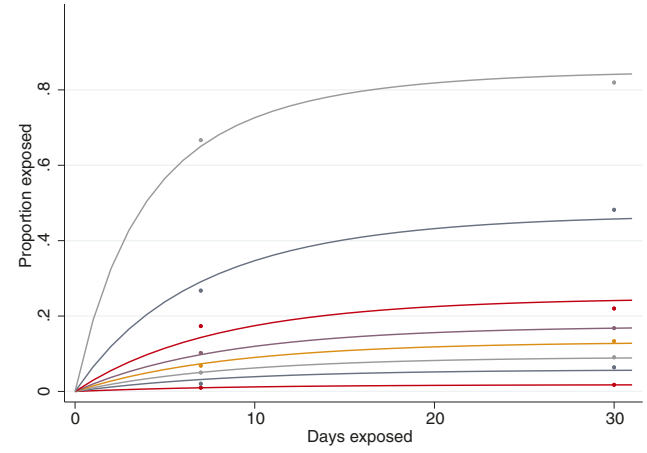
A2. Foodbook, Exponential model.



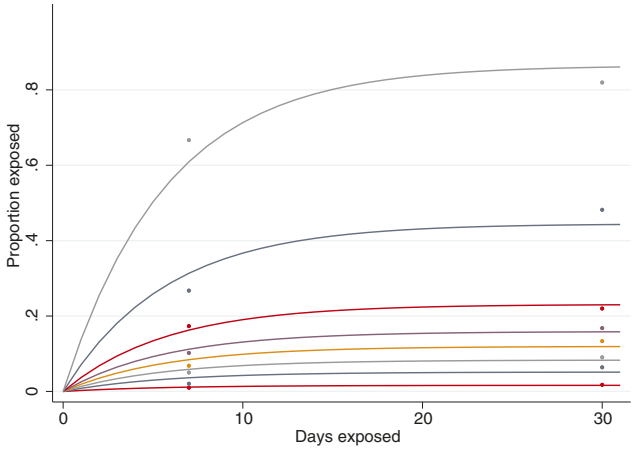
B1. Healthy Control, Geometric model III.



B2. Healthy Control, Exponential model.



C1. US Population, Geometric model III.



C2. US Population, Exponential model.

Figure 2. Select exposure questions proportions, for the three datasets, fitted to the two finalist models. Each colour represents a different exposure question.

models fit on datasets using the parameters obtained from other datasets were analysed for the two models remaining for consideration: Geometric model III and the Exponential model (Table 2). The positively signed values in the score differences column imply that the Exponential models have higher RMSD scores than the Geometric models III and thus performed less well when parameter estimates from other datasets were used. The sum of these model comparison differences for the three datasets was 0.06502958, an overall positive score suggesting that Geometric model III was more robust to parameter misspecification.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

Model recommendations

There are two candidates for final evaluation, the Exponential model and Geometric model III. Summarizing the four criteria, we found that the Exponential model fared worse on model fit compared with Geometric model III; it made out-of-sample predictions that reach their asymptote in a reasonable manner similarly to Geometric model III; it was easier to implement because of its simpler functional form; and finally, it was not as robust to parameter changes. Overall, Geometric model III had better model fit and was more resilient to an incorrect parameter, which in our judgement made up for its more complex functional form and as such, it is the recommended model. An example implementation of this model is available in an xlsx file for converting prevalences collected with a reference period of 7 days to any reference period (Supplementary Material S6). This example is specific to outbreak practitioners, though it can be adjusted for any subject matter.

If converting proportions to long reference periods is important for the practitioner and Geometric model III is overly complex, then we recommend using the Exponential model. If Geometric model III is too cumbersome for implementation and extrapolating past the highest reference period is not needed, Geometric model II would be the next best option as it fits the data better than the Exponential model.

A discussion of recommended parameter values to use for these models, and the implications for designing a survey to collect exposures with multiple reference periods for those who want to estimate their own parameter values using the models, is provided in Supplementary Materials S5.

Interpretation of the a and b parameter

The b parameter in Geometric models II and III may represent the ability to remember the previous day's exposure. A value of 0.80 could be interpreted as there being a 0.80 probability of remembering yesterday's exposure, $0.80^2 = 0.64$ for 2 days ago, and so on. This interpretation suggests that one's memory gets progressively worse for each successive day that is added to the reference period. This accumulates to the point where, for b parameters we estimated here, very little was added to the probability of exposure past 10 days, and almost nothing past 20 days. For Geometric model III, with $b_{fit} = 0.80$ for the 10^{th} day would indicate about $0.80^{10} = 0.11$ is remembered. The P_{1c} parameter is then the 1 day exposure proportion with the effect of b removed.

Memory-based biases include telescoping (the respondent overstates the number of events for shorter time periods), and forgetting an episode entirely [8]. The literature delineated issues with recalling past events into four groups: the length of time prior to the interview that the respondent is asked to recall, the saliency or importance of event in a person's life, the difficulty of the recall task (such as the amount of information to be reported), and the motivation of the respondent to search their memories [6]. Personal characteristics such as age and income can also affect recall [13].

However, responses over longer periods of time can also be interpreted as normative statements about what one usually does [14]. This interpretation can also explain why probabilities do not converge to 1, but often much less than 1. This suggests that answers to questions with long reference periods more represent the population's long-term average preference. For instance, for a 100 day reference period, a person answering if they ate an apple in that time period are likely responding that they like apples and thus would likely have had at least one in that long time period versus specifically recalling the last time they had eaten one.

Conversely, answers of 'no' are expressing that the food item is not liked at all and thus there was never an exposure. Here, the b parameter in Geometric model III is the shape parameter of the function that produces the asymptote that represents the long-term exposure preference of the population. For the Exponential model, it is the a parameter. The lambda parameter produces an exponential curve with a specific concavity that fits the data best and would asymptote to 1 with an a parameter of 1. The question-specific a parameter compresses the function towards the x-axis so that the asymptote is exactly a, and thus represents the maximum long-term proportion exposed and has a direct interpretation as the long-term preference.

For Geometric model III, both interpretations for b may be valid, with the memory interpretation being more relevant with short reference periods of roughly less than 7 days as this is where the curve starts to move towards zero more quickly as the reference period length decreases. The preference interpretation may be more relevant for longer reference periods over 7 days.

We attempted to fit Geometric models II and III with a parameters representing the long-term exposure as in the Exponential model, as well as the b parameter, in the hope of separating the effect of these two factors, but they did not converge. We suspect there may be models using different functional forms that could separate the two factors. This may be a fruitful topic for future research.

Limitations

It is worth noting that collecting exposure information for controls with a directly comparable reference period to case data is always ideal, and should be done where time and resources allow. The methods described herein are designed to make the best use of exposure information that is not collected with the desired reference period and it is the only data source available.

Adjusted control proportions from the models presented here are assumed to contain the same questionnaire level biases as the unadjusted and as such, valid comparisons can be made with them to case exposure proportions under the common assumption that case and control data have the same biases. However, there are areas where the adjustment models may not be optimal and lead to biased comparisons.

For instance, the Healthy Control dataset had many more exposure questions with small proportions (< .5), which may have biased the estimation of parameters. The Foodbook dataset had a variety of proportion sizes, but it only had two reference periods to fit the models. More periods would provide more assurance of model fit and we look forward to future researchers collecting numerous reference periods to examine this. The US Population Survey asked respondents answering yes to exposure with the month long exposure period about their exposure with the 7 day reference period. This ordering could affect the way respondents answer the questions and this might bias the parameter estimates compared with surveys collected that do not have this dependency between the questions. For the Healthy Control and Foodbook datasets, questions were excluded if a proportion of a shorter reference period was larger than one in a longer reference period, which might have affected parameter estimates. Model parameters may change by age and this was not considered in the analysis. Though we could not find any differences, some groupings of exposures might have substantively different model parameters if

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745

estimated as a subset, and this along with age is another angle for potential future research.

Conclusions

A method to convert exposure questions from one reference period to another was described. Of the models examined, Geometric model III was the best overall model and was our top recommendation, though Geometric model II and the Exponential model might work for some practitioners depending on their specific needs. The ability to convert proportions is helpful in enteric outbreaks because the control proportions can be converted to match the case exposure proportions collected for the specific incubation period of the pathogen of interest. In particular, this may improve the accuracy of the control exposure proportions for short exposure periods such as 3 days when the only available data source is surveys collected with a 7 day reference period. It may also help risk assessors who wish to convert proportions to a more useful recall period such as 1 day. It has much broader applications in other fields with similar questions posed to respondents who must remember if an event ever occurred within a defined period of time where this period of time needs to be adjusted.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0950268826101745.

Data availability statement. Data analysed from the Foodbook Survey and the Healthy Control Survey will be made available upon request.

Acknowledgements. The author would like to thank Carolee Carson and Vanessa Morton for valuable feedback from reviewing a draft of the paper.

Author contribution. Conceptualization: M.H.; Data curation: M.H.; Formal analysis: M.H.; Investigation: M.H.; Methodology: M.H.; Project administration: M.H.; Software: M.H.; Validation: M.H.; Visualization: M.H.; Writing - original draft: M.H.; Writing - review & editing: M.H.

Funding statement. This work was funded by the Public Health Agency of Canada.

Competing interests. The authors declares none.

Disclaimer. The findings and conclusions in this report are those of the author and do not necessarily represent the official position of the Centres for Disease Control and Prevention or the Public Health Agency of Canada.

References

[1] **Public Health Agency of Canada** (2015) Foodbook report. Ottawa, ON, Canada: Government of Canada. https://www.canada.ca/en/public-health/services/publications/food-nutrition/foodbook-report.html

[2] **Christidis T**, et al. (2020) A comparative exposure assessment of foodborne, animal contact and waterborne transmission routes of *Salmonella* in Canada. *Food Control* **109**, 106899. https://doi.org/10.1016/j.foodcont.2019.106899.

[3] **David JM**, et al. (2014) Assessing multiple foodborne, waterborne and environmental exposures of healthy people to potential enteric pathogen sources: Effect of age, gender, season, and recall period. *Epidemiology and Infection* **142**(1), 28–39. https://doi.org/10.1017/S0950268813000770.

[4] **Centers for Disease Control and Prevention** (2001) *Population Survey Data, 2000–2001, Foodborne Diseases Active Surveillance Network*. Atlanta, GA: US Department of Health and Human Services.

[5] **Pintar KDM**, et al. (2017) A comparative exposure assessment of campylobacter in Ontario, Canada. *Risk Analysis* **37**(4), 677–715. https://doi.org/10.1111/risa.12653.

[6] **Groves RM** (ed.) (2004) *Survey Errors and Survey Costs*. Hoboken, NJ: Wiley.

[7] **World Health Organization** (2016) Quantitative microbial risk assessment: Application for water safety management. https://www.who.int/publications/i/item/9789241565370

[8] **Sudman S and Bradburn NM** (1973) Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association* **68**(344), 805–815. https://doi.org/10.1080/01621459.1973.10481428.

[9] **StataCorp** (2021) *Stata Statistical Software*. College Station, TX: StataCorp LLC.

[10] **Microsoft Corporation** (2025) *Microsoft Excel 365*. Redmond, WA: Microsoft Corporation.

[11] **Raftery AE** (1995) Bayesian model selection in social research. *Sociological Methodology* **25**, 111–163. https://doi.org/10.2307/271063.

[12] **Statistics Canada** (2015) *Canadian Community Health Survey Nutrition*. Ottawa, ON: Government of Canada.

[13] **Coughlin SS** (1990) Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology Elsevier* **43**(1), 87–91. https://doi.org/10.1016/0895-4356(90)90060-3.

[14] **Scott C and Amenuvegbe B** (1990) *Effect of Recall Duration on Reporting of Household Expenditures: An Experimental Study in Ghana*. Washington, DC: World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/598051468774854613

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 10:34:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0950268826101745