Good morning. It's my pleasure to be here and to present to you some ongoing work using survey-based data to estimate risk factors for use in disease mapping surveillance. The Bayesian statistical approach is a foundation of this work. On the title slide, I present an outline for my talk this morning.

I am a Staff Scientist in the Research Unit at Cancer Care Ontario. Predominantly, my work has focused on using Bayesian methods for small-area disease mapping or risk analysis over the past 5 plus years. By small-area, I mean Census-based administrative units called Dissemination Areas having populations of 400 to 700 persons, on average. For people in the audience from America, these would be akin to “Census Blocks” in terms of geographical hierarchy.

I would like to acknowledge the important collaboration that we have developed between colleagues at the University of Toronto and Public Health Ontario as we work towards better understanding apparent clusters of liver cancer incidence within specific regions of Ontario.
 Advances in statistical methods for geographic surveillance have been paralleled by increased computing power, and utilization for geo-referenced cancer registry data has increased. Examples include NAACCR’s SEER*Stat Bridge and CINA+ Online, the Small Area Health Statistic Unit’s Rapid Inquiry Facility and the Public Health Agency of Canada’s Chronic Disease Infobase.

Such access to cancer registry data are useful for surveillance activities and hypothesis generation. However, there is presently a lack of granular risk factor data – an important consideration for either surveillance activities or hypothesis generation.
Background

- In the Greater Toronto Area (GTA), elevated liver cancer rates:
  - 2011 Population: Toronto = 2.6M; Peel = 1.3M; York = 1M

Within the Ontario context, it has been noted for some time that liver cancer incidence rates are elevated in the Greater Toronto Area. This has been attributed to the large immigrant population originating from parts of the world where hepatitis infection is endemic. The plot on the right side of the slide shows the relationship of age-standardized liver cancer incidence rates and the percent of immigrant population by public health unit. Three of the GTA health units are “outliers” on this graph: Toronto, Peel and York. You can see that these regions comprise a total population of about 5 million persons as of 2011.
The purpose of our study is to identify whether specific neighbourhoods of elevated liver cancer incidence exist and if so, to understand their association with relevant risk factors. For liver cancer, the risk factors are chronic hepatitis B or C infection, heavy alcohol consumption, non-alcoholic fatty liver disease, diabetes, smoking and exposure to aflatoxin B1. Three risk factors are derived from the Canadian Community Health Survey, or CCHS: heavy alcohol consumption, excess body weight and ever/never smoking status. Similar health surveys are conducted in other jurisdictions as listed on screen.

The primary objective of the CCHS is to gather health-related data at sub-provincial health regions. It is a cross-sectional survey and is currently collected annually with a sample of approximately 65,000. Prior to 2007, it was collected bi-annually with a sample of 130,000. Persons 12 years of age and over are sampled voluntarily with a strategy to provide equal importance to the health regions and provinces. Sampling is conducted by telephone, except for the first bi-annual survey which included in-person interviews. Note that the survey is not designed for small-area analyses, but the “share” version of the CCHS provides a respondent’s complete postal code.

In the United States, there is the Behavioural Risk Factor Surveillance System. This is an ongoing, telephone-based survey to understand and reduce behavioural risks for planning, initiating, supporting and evaluating health promotion and disease prevention programs. Sampling is conducted monthly of persons aged 18 and over and has been U.S.-wide since 1993. According to the U.S. department of health & human services, the zip code of respondent’s are available on the public use files (http://aspe.hhs.gov/hsp/06/catalog-ai- an-na/brfss.htm).
There are many challenges associated with estimating the prevalence of risk factors at the small-area level using statistical models, but none of these are “show-stoppers”. These include the granularity, accuracy and completeness of the data, population mobility, and in the Canadian context geographically locating survey respondents in rural areas. There may be changes in the survey design and data collection over different cycles of the surveys. Such changes are well-documented for the CCHS.

You may have to work with data access and confidentiality restrictions and there are such restrictions with the CCHS file used in this work. We are estimating risk factors at a more granular geographic unit than intended by the survey design. This translates to a small numbers problem which can cause implausible or highly variable estimates. To solve this problem, we employ a spatial, hierarchical Bayesian model that takes advantage of the spatial correlation often present in such data – a fact that would violate one of the fundamental assumptions of the classic statistical approach.
This slide lists the data sources and tools involved in the work. The data sources are:

- the Canadian Community Health Survey, or CCHS, for the risk factor estimates;
- The Census of Canada for population, socio-demographics and boundary files;
- the Ontario Cancer Registry, or OCR, for incidence cases of primary liver cancer;
- and, the Integrated Public Health Information System, or iPHIS, for data on registered cases of hepatitis occurrence. These data were not yet available for this presentation.

The key tools are:

- Statistic Canada’s postal code conversion file plus (PCCF+) to convert and link postal code data to Census administrative units;
- R software for statistical computing. Several packages or libraries are required to do this modeling: R-INLA, the Integrated Nested Laplace Approximation wrapper for R, geostatsinla providing functions to obtain exceedence probabilities and hyper prior diagnostics for the Bayesian approach, and mapptools and spdep to work with spatial data in R.
This flowchart displays the high-level process involved in the method. We start with the CCHS data, identify or derive the variables of interest, geocode them and pool multiple cycles or samples of the survey data to increase the statistical power. Most of the work is done in "R" as denoted by the grey box. This is where the modeling is performed once the data are imported and merged to the common Census administrative units.

<click for animation>

For a few minutes, let’s focus on the risk factor and cancer incidence models.
Both the risk factor and cancer incidence modeling employ the Besag, York and Mollié, or BYM, model. This is a spatial, hierarchical Bayesian model. Note that the BYM model can be used with different probability distributions – we will elaborate on this in a few slides. Another key point of the BYM model is that it pools neighbouring or abutting areas to take advantage of the spatial correlation that are often present in data. This may be referred to as a conditional autoregressive distribution or CAR and by performing something akin to a local mean, it stabilizes the resulting estimates or predictions. The model-based estimate for risk factors are input as covariates for a primary liver cancer incidence BYM model. The next slide elaborates on this.
Presented are the two formulae of the hierarchical Bayesian model known as the BYM model. First, let’s focus on the risk factors formula. Note that the top line specifies that this is using a Bernoulli distribution as we have discretized our risk factors into binary values, for example ever/never smokers, where \( i \) denotes a respondent in area \( j \). The next line contains the logit of the estimate and this is followed by the random effect specifications, where the \( U \) term is the independent random effect and the \( V \) term is the spatial effect, defined as a CAR. Let’s focus on the logit line of the equation.

The form of this equation may be familiar to you as it is that of a regression equation. The logit(\( P_{ij} \)) represents log-odds of a risk factor for a survey respondent \( i \) within an area \( j \). Alpha is the intercept, and the betas are the coefficients for the covariates age and cycle. But where we would typically have an error or residual, we instead have a \( U \) and a \( V \), previously defined. Together, these are akin to residuals in regression. They are the risk factor estimates unexplained by a respondent’s age group and survey cycle, by sex and within an area \( j \). Note that the estimates are for the areas of interest “\( j \”).

Next, let’s examine the cancer incidence formula. The top line states that the observed number of cases for an area \( j \) is Poisson distributed based on the risk and expected count. The next line is for the log of the relative risk. As before this is followed by specification of the \( U \) and \( V \) terms, the independent random effect and spatial effect, respectively. Now, let’s focus on the log of the relative risk.

Once again we have the familiar form of a regression equation. In this case, the log of the relative risk is comprised of an intercept, covariates and residuals. Note there are covariates for smoking, heavy alcohol consumption, excess body weight and hepatitis. As
before, U and V are the independent random effect and spatial effect, and together they comprise the relative risk for area j unaccounted for by the covariates and study area mean relative risk which is the intercept.

Specifically, we input the areal-level prevalence of the risk factor estimates unaccounted for by age group and CCHS cycle into the cancer incidence model, as denoted by the arrows.

After we have implemented the full cancer incidence model, the U and V terms together are the relative risk estimate for area j unaccounted for by the mean relative risk and covariates. Now, let’s move on to viewing some results.
For time constraints, I’ve elected to focus on results for males.
This map displays neighbourhood-level prevalence estimates of alcohol consumption of at least 4 drinks per day, on average, among men. The geographic boundaries of the dissemination areas are shown in dark grey, and highways are displayed in light grey. Dark brown areas indicate higher prevalence and these areas coincide with affluent areas such as Rosedale, Forest Hill and the Kingsway; but, they also coincide with some outer suburban areas in Scarborough, Mississauga and to the northern areas of Peel and York Regions.

The “exceedence probabilities” are a useful feature of the Bayesian approach -- these are the credibility limits for the estimates akin to confidence intervals in the classic statistical approach. We display these as cross-hatched overlays. In this example, we do not have any areas where we can be exceedingly confident that the estimated prevalence is significantly higher than the Ontario comparator within this study area.

This map addresses the prevalence of the risk factor taking into account the age group and cycle of the CCHS respondent. Since we adjust our liver cancer estimates for age, what about the risk factor? What does the distribution of heavy alcohol consumption look like unexplained by the respondent’s age group and cycle?
This map answers that question by displaying the residuals from the modeled heavy alcohol consumption. Specifically, these are the U and V terms combined from the BYM model, and graphed for each area. Note that these values are in the logit or log-odds scale. A value above zero on the log-odds scale indicates an odds above 50/50. This occurs in the highest class, displayed in dark brown. Similar to the prevalence estimates, we can calculate the exceedence probabilities for the residuals. Again, note that for our model-based heavy alcohol consumption estimates, we have no areas where we can be exceedingly confident that the log-odds for an area within the study region is significantly higher than the Ontario comparator. These estimates of heavy alcohol consumption among men unexplained by a respondent’s age group or CCHS cycle are then input as a covariate into a liver cancer model. This is done for each of the model-based risk factors, but for time constraints I’ve only shown heavy alcohol consumption in some detail. Next, let’s see what age-adjusted liver cancer incidence looks like.
Mapped here are the smoothed relative risk or standardized incidence ratios for primary liver cancer incidence among men within the study area from 2004 to 2008. There were a total of 1029 cases registered over this period within Peel, Toronto and York. Darker red hues indicate incidence rates higher than Ontario and grey hues are rates below Ontario, and yellow are similar. These SIRs are age-adjusted only using the BYM model. The cross-hatching indicates the exceedence probabilities, where there is statistical evidence that the relative risk estimates are significantly elevated compared to Ontario. We have several clusters of elevated risk within the study area including the downtown core of Toronto, northeastwards around Agincourt and Markham and in central Mississauga. Generally, the modeled SIRs decrease towards the north as shown in the inset map. For corroborative purposes, we ran the non-modeled observed and expected values through the Space and Time Scan Statistic. The significant spatial cluster detected by the scan statistic is shown in the blue ellipse, and there are about 50 excess cases here over the 5 year period. One would likely be interested in how much of the relative risk estimates are unexplained by age adjustment. Fortunately, since our relative risk estimates are model-based we can examine this.
Using the same colour and data classification scheme, this map displays primary liver cancer incidence among men unexplained by age adjustment alone. The dark red areas indicate rates higher than the Ontario comparator. Compared to the previous map, it is clear that where the relative risk estimate was high, we also have higher unexplained relative risk. As indicated by the cross-hatching, we have the most statistic evidence for excess incidence rates unaccounted for by age adjustment within the downtown core of Toronto.
Finally, this map displays the residuals or unexplained relative risk from a liver cancer model that takes into account the model-based risk factor estimates for heavy alcohol consumption, excess body weight and smoking. The colours and data value break points are the same as before. Compared to the previous map of the unexplained relative risk for age-adjustment alone, the pattern here is greatly attenuated. Two areas to the east and west of the downtown core of Toronto remain with rates 50% or greater than the Ontario comparator. Even after accounting for these risk factors, it is clear there are still areas of elevated liver cancer rates compared to Ontario, as indicated by the coinciding dark red and cross-hatched areas.

The next step is to build-in data on the reported cases of hepatitis B and C. Then, we will assess and map the associations between primary liver cancer incidence and each relevant risk factor. The coefficients resulting from the modeling are a key aspect of this approach and will allow us to quantify the strength of the association. Any remaining unexplained relative risk may be due to aflatoxin B1 exposure, limitations in the modeling or other unknown risk factors within this study area.
There are several limitations to creating model-based prevalence estimates of risk factors based on survey data. First, we have pooled multiple cycles of the CCHS thereby creating a population of people across almost a decade of survey data. This population does not truly exist as a cohort. Implicate with that first point is that the risk factor modeling assumes that these results applies to the true population, as we are using the survey data at a finer resolution than the survey design. Another limitation is the temporal period: the survey data are fairly recent and current cancer outcomes may be related to risk factors 15-20 years ago given the lag between exposure and onset of cancer.

At present, as we develop this work, we have elected to input the model-based risk factor estimates as a fixed effect in the cancer incidence model. Thus, we “loose” the variability present in the risk factor estimates. I’ll return to this again in a few slides.

And lastly, it’s important to recognize that this is ecological analysis. So, it’s very useful for surveillance activities and hypothesis generation. It’s also a fairly rapid method to build evidence towards larger grants for studies that address etiology.
Using survey-based data for risk factor estimation has several strengths. There is the availability of digital, relevant and high quality risk factor data. Regarding the utilization of the BYM model, it’s an approach that is designed to work with and take advantage of the geographical correlation that are present in the data.

The Bayesian statistical approach is a strength as it yields posterior distributions to calculate the exceedence probabilities or credibility intervals. This is particularly useful for detecting cancer clusters. Further, we have employed the Integrated Nested Laplace Approximation, or INLA, for the Bayesian inference. This approach is very rapid and does not require super computers. This allows you to devote more time to conduct sensitivity analyses for the Bayesian approach.

This approach is also widely applicable and extensible. By extensible, I mean that the results may be used for other models. By widely applicable, the methods can be used for any relevant health data far beyond cancer risk factors. Also, it’s evident that many jurisdictions have relevant health survey data.
In terms of future work, we are collaborating with Public Health Ontario to estimate hepatitis B and C occurrence by these small-area geographies. These results will also be input as a covariate into the liver cancer model.

As well, we hope to employ the methods as a “joint model” to allow the risk factor posterior distributions to inform the cancer model. This would take into account the variability in the risk factor estimates. For now we have approached this as a two step model. This is more straight-forward than the joint modeling but it means that we are not taking advantage of the variability present in the risk factor estimates.
As I wrap-up my presentation, I wanted to acknowledge our funding source and REB approval, listed on screen. I would be happy to take any questions, as time permits.
Questions?