PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A validation study of health administrative data algorithms to identify individuals experiencing homelessness and estimate population prevalence of homelessness in Ontario, Canada
AUTHORS	Richard, Lucie; Hwang, Stephen W.; Forchuk, Cheryl; Nisenbaum, Rosane; Clemens, Kristin; Wiens, Kathryn; Booth, Richard; Azimaee, Mahmoud; Shariff, Salimah

VERSION 1 – REVIEW

REVIEWER	Dennis Culhane UPENN/SP2 USA
REVIEW RETURNED	01-Apr-2019

GENERAL COMMENTS	An excellent paper. My primary concern relates to the misclassification of persons who had no health care visits around the index date for estimating homelessness. Since this was the group for whom misclassification was greatest, the authors should more clearly explain how the study addresses classification for people without recent health care experiences. For example, if the HICI dataset only includes data from which to estimate homelessness on persons with recent health services use, this could lead to a significant underestimation of homelessness by persons without recent health services use. At least that is this reader's inference. Please clarify. In other words, how can estimation be sufficient without data on the person from which to make the classification? This reviewer is unfamiliar with the Canadian data and may not understand. Any suggested explanation for the observed increase in homelessness over this period? Is there any concern that homelessness "caseness" is increasing in the admin data because of greater awareness by data entry persons? Some external validation of increasing homelessness rates would be useful to cite, if available. The .3-.5% annual rates are consistent with US rates of shelter use, FYI.

REVIEWER	Bisan A. Salhi Emory University USA
REVIEW RETURNED	07-Jun-2019

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript, in which the authors use health administration data to identify an episode of homelessness and to estimate the prevalence of homelessness in the general population. Overall, the paper presents important

information and is a worthwhile contribution to the literature on homelessness in the healthcare setting.

As a reader and a non-Canadian, however, I would have benefited from more contextual information in the paper. For instance, the introduction is rather short and misses an opportunity to better present the background of the study. Similarly, the authors make reference to ICES in page 4, line 51, but it would have been helpful to have the acronym spelled out and to have a short description of the institution. Finally, the HHiT study is briefly mentioned on page 5 and is appropriately referenced, but a bit more description would have been helpful.

The methods presented are appropriate, but more details on the algorithms and data analysis procedures would be helpful. On page 5, lines 29-31, the authors note that "the computer programs rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or may require modification." It's unclear what this means.

Finally, I have two copy editing questions:

1. Page 5, line 50: do you mean 2016?
2. Page 12, line 12: I would suggest changing "the homeless" to "homeless people."

Again, thank you for the opportunity to review this manuscript. I look forward to seeing it in published soon.

REVIEWER	Stuart Lee
	Swinburne University of Technology, Australia
REVIEW RETURNED	18-Jun-2019

| GENERAL COMMENTS | This study presents data validating the use of statistical algorithms using administrative data from multiple Canadian health services to estimate the prevalence of homelessness in the general population and use this to measure service use and clinical characteristics for this population and change in prevalence and characteristics over time. |

This paper is well written, and the reported study is unique in drawing from two robust data sources (a well characterised prospective longitudinal cohort of people homeless or marginally housed; an Ontario database of all people eligible for health) to both validate the predictive accuracy of developed statistical algorithms, and to then apply them to examine longitudinal trends in the studied population. Data is both informative for understanding homeless population health care needs and trends, and for demonstrating the potential benefits of such statistical algorithms when applied to a population that is often difficult and costly to measure.

Some aspects of this manuscript could be strengthened.

1) Introduction: It would help if previously generated counts of the Canadian homeless population are summarised (including measured prevalence changes over time) along with factors complicating the counting of this population to provide context for the need to develop and validate statistical algorithms to improve

the efficiency of measurement and prevalence and associated patient- and service-level data characteristic of this population.

2) Method: Clarify some aspects.
- What was the method for random sampling from the ICES Registered Persons Database?
- P 5: What is the timeframe for HHiT participant interviews (e.g. is this every 12 months for 2 years)?
- What was the sample of HHiT participants with available data in each of the validation testing phase years 2007-2014 and was there consistent accuracy in algorithm prediction over the validation phase years?
- P 6: Summarise in this paper the criteria for housed vs homeless classification rather than directing the reader to another paper to understand this. Were changes in accommodation status or other screening methods, apart from comparing postal code against known shelters, used to validate the general population participants as not being homeless? P 12 paragraph 2 states many of the general population false positives may have been homeless, but how was this determined.
- P 8: State the criteria used to determine the optimal algorithm. What method was used to correct for sensitivity in the population-prevalence estimates.

3) Results: Clarify the following:
- What was the mean length of participation of participants in the HHiT study used to generate the % of time being homeless? Similarly, what was the mean length of data availability for the included general population participants.

4) Discussion:
- P 12. It is stated that 238 general population participants who were false positives may have been misclassified. Did this mean that these participants were likely to have experienced periods of homeless and as a result should have been re-coded as homeless for the analysis. If so, would either their exclusion from the analysis or their recoding to being "homeless" in the analysis (which would add almost one-third to the homeless sample for analysis purposes) impact study findings?
- It would help to discuss further the observed under-identification of homeless prevalence in 2016 (e.g. 90,000 homeless individuals based on national counts vs 60,000 based on the chosen optimal algorithm) and whether the algorithm can be improved. First Nation populations were thought to have been under-sampled but are there also other homeless subpopulations (e.g. staying temporarily with family / friends; living in boarding rooming houses or overcrowded conditions) that were not sampled via this algorithm. Could housing provider data be included in the linkage to capture and match with health service data, people recorded by these services to be accessing crisis or shelter accommodation or housing case management.
- Figure 2 summarising annual prevalence rates identified an almost doubling of the Ontario homeless prevalence between 2007 and 2016. While this rate of growth is far higher than general population growth, is this estimate of growth different to other homeless population counts measured via other means (e.g. census or interview)? Perhaps add some discussion to examine what may have contributed, including whether change in clinician practice to identify and document homeless status in health service databases may have contributed?

	- Perhaps also add discussion on what would be needed for expansion or implementation of similar population homeless surveillance algorithms in other jurisdictions. How is the quality of coding of homeless / housing instability status by health services likely to impact algorithm prediction? 5) A final minor suggestion would be to ensure all acronyms (e.g. ICES; HHiT) are defined in text before their use.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 comments to authors
COMMENT 1.1: An excellent paper. My primary concern relates to the misclassification of persons who had no health care visits around the index date for estimating homelessness. Since this was the group for whom misclassification was greatest, the authors should more clearly explain how the study addresses classification for people without recent health care experiences. For example, if the HICI dataset only includes data from which to estimate homelessness on persons with recent health services use, this could lead to a significant underestimation of homelessness by persons without recent health services use. At least that is this reader's inference. Please clarify. In other words, how can estimation be sufficient without data on the person from which to make the classification? This reviewer is unfamiliar with the Canadian data and may not understand.

RESPONSE: Thank you for your positive feedback.

An inherent limitation in utilizing health services data for case definitions (e.g. diabetes [1]) is that those without a health care encounter are classified as not satisfying the case definition, and therefore considered free of the condition. As indicated, this results in a number of false negatives for individuals who did not experience any healthcare encounters during episodes of homelessness, in turn decreasing the sensitivity of the algorithm. We have clarified this in the Methods section, under "Case Ascertainment Algorithms and Data Sources" paragraph 3 as follows:
Reference housing episodes or calendar years without overlapping healthcare encounters were considered algorithm negative ("housed") by default, to reflect the administrative data's inability to identify homelessness for such reference periods.

We also reiterated the challenge inherent to this methodology in the Discussion section.

To quantify the potential misclassification due to lack of healthcare encounters, we provided the breakdown of false positives arising from no health care encounters in the Validation Results section as follows "Absence of a healthcare encounter during the reference period accounted for 64.5% (n= 1,825) of false negatives". This result is also elaborated in Supplement Table 6 for the optimal housing episode and optimal annual housing experience algorithms.

COMMENT 1.2: Any suggested explanation for the observed increase in homelessness over this period? Is there any concern that homelessness "caseness" is increasing in the admin data because of greater awareness by data entry persons? Some external validation of increasing homelessness rates would be useful to cite, if available. The .3-.5% annual rates are consistent with US rates of shelter use, FYI.

RESPONSE: We're unable to know to what degree improved awareness of homelessness increased case sensitivity province-wide during this period; however, while reviewing case sensitivity in our validation cohort on an annual basis between 2007 and 2014 we found little (less than 4%) variation and no consistent trend upwards in case sensitivity. Moreover, though there are no other counts in Ontario against which to compare our estimate, the State of Homelessness reports (2, 3) and a recent presentation by Employment and Social Development Canada (4) (both of which rely on Point in Time count data) indicate that counts have increased within Canada; the former show an increase of approximately 15% from 2013 to 2016, the latter describe a 14% increase between 2016 and 2018. Therefore, although we do believe the observed increase in homelessness to likely be a combination of increasing case sensitivity and actual increases, we believe a majority the change is due to an actual increase in homelessness in Ontario. We have altered the discussion to include this information.

Reviewer 2 comments to authors

COMMENT 2.1: Thank you for the opportunity to review this manuscript, in which the authors use health administration data to identify an episode of homelessness and to estimate the prevalence of homelessness in the general population. Overall, the paper presents important information and is a worthwhile contribution to the literature on homelessness in the healthcare setting. As a reader and a non-Canadian, however, I would have benefited from more contextual information in the paper. For instance, the introduction is rather short and misses an opportunity to better present the background of the study.

RESPONSE: Thank you for the positive feedback. While we wrote our paper to be applicable as much as possible beyond the Canadian context, we agree additional background about our healthcare system and homeless population would be helpful. We have adapted the introduction accordingly.

COMMENT 2.2: Similarly, the authors make reference to ICES in page 4, line 51, but it would have been helpful to have the acronym spelled out and to have a short description of the institution.

RESPONSE: Recently, the institute formerly known as the Institute for Clinical Evaluative Sciences formally adopted the initialism ICES as its official name. We have added a short description of ICES for those unfamiliar with the institution.

COMMENT 2.3: Finally, the HHiT study is briefly mentioned on page 5 and is appropriately referenced, but a bit more description would have been helpful.

RESPONSE: Thank you. We have added the following short description of the HHiT study to the quoted section:

The HHiT study was conducted between 2009 and 2014 in three Canadian cities (Toronto, Ottawa and Vancouver) and aimed to assess the impact of housing transitions on health. Participants were randomly selected at shelters, meal programmes, community health centres, drop-in centres, rooming houses, and single-room occupancy hotels and interviewed once per year until the end of the study or until the individual withdrew.

COMMENT 2.4: The methods presented are appropriate, but more details on the algorithms and data analysis procedures would be helpful. On page 5, lines 29-31, the authors note that "the computer programs rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or may require modification." It's unclear what this means.

RESPONSE: ICES analysts use a suite of unique SAS macros and templates to streamline and standardize routine data cleaning, processing and analysis tasks common to health administrative data research using ICES data. As most projects call upon these macros and templates at some point, project coding will not execute in other programming environments without access to this underlying code and datasets.

As such, we have added a Supplement Table (Supplement Table 2) that more explicitly describes the algorithms we tested.

COMMENT 2.5: Finally, I have two copy editing questions:
1. Page 5, line 50: do you mean 2016?
2. Page 12, line 12: I would suggest changing "the homeless" to "homeless people."

RESPONSE:
1. The HHiT cohort only extends to 2014, limiting the period for validation, however we tested our best performing algorithm for the 2007 to 2016 period, as data were available. We have clarified this distinction in the abstract and body of the manuscript.
2. Thank you, we agree and have edited the manuscript accordingly.

Reviewer 3 comments to authors
COMMENT 3.1: This study presents data validating the use of statistical algorithms using administrative data from multiple Canadian health services to estimate the prevalence of homelessness in the general population and use this to measure service use and clinical characteristics for this population and change in prevalence and characteristics over time.

This paper is well written, and the reported study is unique in drawing from two robust data sources (a well characterised prospective longitudinal cohort of people homeless or marginally housed; an Ontario database of all people eligible for health) to both validate the predictive accuracy of developed statistical algorithms, and to then apply them to examine longitudinal trends in the studied population. Data is both informative for understanding homeless population health care needs and trends, and for demonstrating the potential benefits of such statistical algorithms when applied to a population that is often difficult and costly to measure.

Introduction: It would help if previously generated counts of the Canadian homeless population are summarised (including measured prevalence changes over time) along with factors complicating the counting of this population to provide context for the need to develop and validate statistical algorithms to improve the efficiency of measurement and prevalence and associated patient- and service-level data characteristic of this population.

RESPONSE: Thank you for your positive feedback.

While we have summarized the little other available information on Canadian homelessness counts in the discussion, we agree this same information is important context for the introduction. We have modified the introduction to include further information about homeless counts, as well as present a case for utilizing health administrative data for measurement of homelessness.

COMMENT 3.21: What was the method for random sampling from the ICES Registered Persons Database?

RESPONSE: We used the SAS function PROC SURVEYSELECT to randomly select, from a pool of eligible residents, 200 individuals for each HHiT participant. As this is a standard SAS function for random selection, we have not included the specifics in the manuscripts, but would happy to do so if requested by Reviewer or Editors.

COMMENT 3.22: What is the timeframe for HHiT participant interviews (e.g. is this every 12 months for 2 years)?

RESPONSE: The HHiT protocol indicates participants were interviewed once every year (every 12 months) until the end of the study (in late 2013/early 2014) or until the individual withdrew from the study. There were up to 5 surveys completed. This information has been added to the Methods section.

COMMENT 3.23: What was the sample of HHiT participants with available data in each of the validation testing phase years 2007-2014 and was there consistent accuracy in algorithm prediction over the validation phase years?

RESPONSE: There was variability in the number of HHiT participants in each of the validation testing years, as individuals were recruited and chose to drop out over the study period. This is very common in studies with homeless individuals. Additionally, each HHiT participant had a number of consecutive housing periods, often several per year, further complicating the yearly variability in episode counts being validated. We foresaw that this might cause inconsistencies over time in the algorithm validation statistics, so we reviewed the variability in algorithm validation statistics by year. We found little (less than 4% change in sensitivity, with no significant trend) variability in the algorithm's statistics over time. This information has been added to the discussion.

COMMENT 3.24: P 6. Summarise in this paper the criteria for housed vs homeless classification rather than directing the reader to another paper to understand this. Were changes in accommodation status or other screening methods, apart from comparing postal code against known shelters, used to validate the general population participants as not being homeless? P 12 paragraph 2 states many of the general population false positives may have been homeless, but how was this determined.

RESPONSE: We opted to avoid discussing the criteria for housed vs homeless classification used by the HHiT study in the body of the manuscript as it is fairly lengthy. However, we recognize that many will wish to judge the reference standard's quality and thus have added a high level review in the Methods section as well as added a detailed description of the HHiT cohort in Supplement 3: Databases Used.

Briefly, interview responses to housing status were initially classified into one of 25 types of residence, and then these categories were further classified into one of three mutually exclusive residence categories: housed, institution or homeless. To determine if periods of time spent in institutions (e.g. hospitals, prison, etc.) should be considered periods of homelessness or housing, housing status prior and subsequent to the period of institutionalization were reviewed, and institution housing episodes flanked by any period of homelessness was also considered homelessness, the

logic being that homelessness likely caused (or, conversely, was caused) by the institutionalization event.

Conversely, when sampling the general population, who were assumed to be housed, we excluded those with postal codes known to be associated shelters, and individuals already participating in the HHiT study. Unfortunately given the nature of our data, we did not have other methods to screen such individuals for homelessness, without using administrative data-generated homeless indicators, which would not have been appropriate to use, as this was the objective of the validation. As such, there is a chance that our general-population sample included a small number of individuals who were homeless at some point over the study period, but not living at a homeless shelter. We speculate that some of the 238 individuals who were positively identified as homeless during a health care encounter may truly have been homeless, but there is no method in which to confirm this.

We address your final question about false positives in our response to Comment 3.41.

COMMENT 3.25: P 8: State the criteria used to determine the optimal algorithm. What method was used to correct for sensitivity in the population-prevalence estimates.

RESPONSE: Thank you, we have adjusted the manuscript to indicate that we defined the algorithm with the highest sensitivity, specificity and positive predictive value as optimal, and that considerations for scalability beyond Ontario (using health administrative data that were not specific to Ontario and available in other provinces) were also considered. We also now indicate how we corrected for sensitivity to derive the population prevalence estimates.

COMMENT 3.3: Clarify the following: What was the mean length of participation of participants in the HHiT study used to generate the % of time being homeless? Similarly, what was the mean length of data availability for the included general population participants.

RESPONSE: We have added to the results section the mean length of HHiT participation used to generate the proportion of time spent homeless (64 months), as well as the mean length of time followed for the general population sample (52 months).

COMMENT 3.41: Discussion: P 12. It is stated that 238 general population participants who were false positives may have been misclassified. Did this mean that these participants were likely to have experienced periods of homeless and as a result should have been re-coded as homeless for the analysis. If so, would either their exclusion from the analysis or their recoding to being "homeless" in the analysis (which would add almost one-third to the homeless sample for analysis purposes) impact study findings?

RESPONSE: As previously mentioned in the response to Comment 3.24, we do suspect some of our false positives may have in fact been true positives (i.e. actually homeless) but cannot know this for a fact, nor give some sense of how many are misclassified in our reference standard. We merely suppose this might be the case as our screening of the general population sample is not (and cannot be) foolproof. As our purpose is to validate administrative data indicators for homelessness, we could not re-code or exclude such individuals from our analysis on the basis of administrative indicators of homeless, as this would bias our results (increase the Positive Predictive Value and Specificity). We altered the manuscript discussion to better reflect these thoughts.

COMMENT 3.42: It would help to discuss further the observed under-identification of homeless prevalence in 2016 (e.g. 90,000 homeless individuals based on national counts vs 60,000 based on the chosen optimal algorithm) and whether the algorithm can be improved. First Nation populations were thought to have been under-sampled but are there also other homeless subpopulations (e.g. staying temporarily with family / friends; living in boarding rooming houses or overcrowded conditions) that were not sampled via this algorithm. Could housing provider data be included in the linkage to capture and match with health service data, people recorded by these services to be accessing crisis or shelter accommodation or housing case management.

RESPONSE: We (emphatically) agree that our data is limited in its ability to capture homeless episodes on its own. Ideally the Ontario Health Insurance Plan (OHIP) should introduce a method to capture homelessness in its mandatory data collection, as we believe this would improve algorithm performance. Additionally, ICES health administrative data sources should be linked to other important social service provider data like those you've mentioned (e.g. shelters, income assistance, law enforcement) to improve algorithm performance. As it stands, however, we are restricted by the data sources and housing status variables available to us (as an institution but also in Ontario generally): to our knowledge such other data sources are currently not available for use (particularly linked to other data, given its sensitive nature) in research in Ontario. In future, availability of new data sources with housing indicators may improve algorithm sensitivity.

We also agree that certain homeless subgroups are less well captured. We named those we couldn't capture due to OHIP eligibility, but it is correct to state that others are missed in the validation phase. For example, the HHiT cohort excluded youth <18 years (which misses an important and highly policy relevant subgroup). That being said, it would be incorrect to say we are not sampling individuals in unstable but not "traditional" homeless situations, as the HHiT study does include such transition-style housing statuses (for example, halfway houses, couch surfing, living in a campground, etc) in its classification, even though it ultimately resolves such statuses into Housed or Homeless. It is also important to emphasize that the 90,000 Ontario estimate is based on a number of assumptions. First, the 235,000 national prevalence estimate is a rough estimate, with some suggesting the true number may range between 150,000 to 300,000 (2). Furthermore, when estimating 90,000 in Ontario, we assumed that homelessness was proportionally distributed across Canada, which may not be the case.

We modified our discussion to more explicitly address these points.

COMMENT 3.43: Figure 2 summarising annual prevalence rates identified an almost doubling of the Ontario homeless prevalence between 2007 and 2016. While this rate of growth is far higher than general population growth, is this estimate of growth different to other homeless population counts measured via other means (e.g. census or interview)? Perhaps add some discussion to examine what may have contributed, including whether change in clinician practice to identify and document homeless status in health service databases may have contributed?

RESPONSE: Unfortunately, there are few other data sources to estimate the population size, let alone growth in Ontario (or indeed in Canada generally) (5). The only other population counts are estimates derived from shelter usage and coordinated Point in Time (PIT) counts – this is the data used by the oft-quoted State of Homelessness in Canada reports and from which our 90,000 estimate is derived. If one assumes PIT counts can reliably estimate growth (and many would object to this), they too show an increase over time, which if extrapolated to a period similar to our observation window would look rather similar (these other sources show ~15% change over 2-3 years (2-4).

It is possible that both the PIT methodology and our algorithm are subject to increasing sensitivity (in our case, increasing case sensitivity); however, as mentioned in our response to Comment 3.23, we verified that case sensitivity did not vary much across our validation period, suggesting the observed increase is, at least in part, real. We added this information to our manuscript to add context to our findings.

COMMENT 3.44: Perhaps also add discussion on what would be needed for expansion or implementation of similar population homeless surveillance algorithms in other jurisdictions. How is the quality of coding of homeless / housing instability status by health services likely to impact algorithm prediction?

RESPONSE: The quality of coding for housing status clearly impacts algorithm prediction. As such, establishing population surveillance algorithms for homelessness in other jurisdictions outside Ontario would require a) an assessment of the available data sources (for example, in Canada the standardized "CIHI data" sources are largely available across provinces and territories), b) ensuring there aren't any significant barriers (financial, particularly) to accessing said services, and c) ensuring there is no substantial differences in coding practices between jurisdictions, among the non-standardized datasets. We've added some discussion on this point to highlight these challenges to expansion of implementation.

COMMENT 3.45: A final minor suggestion would be to ensure all acronyms (e.g. ICES; HHiT) are defined in text before their use.

RESPONSE: Thank you, we have revised the manuscript to ensure all acronyms are defined in text before their use.

References cited in the responses:

1. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: Determination of prevalence and incidence of diabetes using a validated administrative data algorithm. Diabetes Care. 2002;25(3):512-516
2. Gaetz S, Donaldson J, Richter T, Gulliver T. The State of Homelessness in Canada 2013. Toronto: Canadian Homelessness Research Network Press.
3. Gaetz S, Dej E, Richter T, Redman M. The State of Homelessness in Canada 2016. Toronto: Canadian Observatory on Homelessness Press.
4. Duchesne, A. & Cooper, I. National Homelessness Data National Shelter Data 2005-2016 Point-in-Time Count 2018. http://datathatmakesadifference.com/wp-content/uploads/2019/06/Annie-Duchesne-Ian-Cooper.pdf Published May 28 2019. Accessed July 10 2019.
5. Ministry of Municipal Affairs and Housing. A place to call home: report of the expert advisory panel on homelessness. 2015. http://www.mah.gov.on.ca/AssetFactory.aspx?did=11038. Accessed July 15 2019.

VERSION 2 – REVIEW

REVIEWER	Bisan Salhi Emory University USA
REVIEW RETURNED	15-Aug-2019

GENERAL COMMENTS	Thank you for the opportunity to review the revised version of this manuscript. The authors have adequately addressed my concerns. The background information provided helps clarify their methods and findings. I look forward to seeing this in print soon.

REVIEWER	Stuart Lee Swinburne University of Technology, Australia
REVIEW RETURNED	28-Aug-2019

GENERAL COMMENTS	This is a previously reviewed manuscript reporting on a study that validated the use of statistical algorithms using administrative data from multiple Canadian health services to estimate the prevalence of homelessness in the general population and use this to measure service use and clinical characteristics for this population and change in prevalence and characteristics over time. My previous comments or suggestions have satisfactorily been addressed and I have no further suggestions to strengthen what is a well written, innovative and potentially highly impactful paper.