BMJ Open

Patterns and predictors of high-cost users of the health system: a data linkage protocol to combine a cohort study and randomised controlled trial of adults with a history of homelessness

Kathryn Wiens ,[1,2] Laura C Rosella ,[1] Paul Kurdyak,[3] Stephen W Hwang[2,4]

To cite: Wiens K, Rosella LC, Kurdyak P, *et al*. Patterns and predictors of high-cost users of the health system: a data linkage protocol to combine a cohort study and randomised controlled trial of adults with a history of homelessness. *BMJ Open* 2020;**10**:e039966. doi:10.1136/bmjopen-2020-039966

► Prepublication history and supplemental material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-039966).

Received 01 May 2020
Revised 01 December 2020
Accepted 07 December 2020

For numbered affiliations see end of article.

Correspondence to
Ms Kathryn Wiens;
kathryn.wiens@mail.utoronto.ca

ABSTRACT

Introduction Homelessness is a global issue with a detrimental impact on health. Individuals who experience homelessness are often labelled as frequent healthcare users; yet it is a small group of individuals who disproportionately use the majority of services. This protocol outlines the approach to combine survey data from a prospective cohort study and randomised controlled trial with administrative healthcare data to characterise patterns and predictors of healthcare utilisation among a group of adults with a history of homelessness.

Methods and analysis This cohort study will apply survey data from the Health and Housing in Transition study and the At Home/Chez Soi study linked with administrative healthcare databases in Ontario, Canada. We will use count models to quantify the associations between baseline predisposing, enabling, and need factors and hospitalisations, emergency department visits and physician visits in the following year. Subsequently, we will identify individuals who are high-cost users of the health system (top 5%) and characterise their patterns of healthcare utilisation. Logistic regression will be applied to develop a set of models to predict who will be high-cost users over the next 5 years based on predisposing, enabling and need factors. Calibration and discrimination will be estimated with bootstrapped optimism (bootstrap performance—test performance) to ensure the model performance is not overestimated.

Ethics and dissemination This study is approved by the St Michael's Hospital Research Ethics Board and the University of Toronto Research Ethics Board. Findings will be disseminated through publication in peer-reviewed journals, presentations at research conferences and brief reports made available to healthcare professionals and the general public.

Trial Registration Number This is a secondary data analysis of a cohort study and randomized trial. The At Home/Chez Soi study has been registered with the International Standard Randomised Control Trial Number Register and assigned ISRCTN42520374.

Strengths and limitations of this study

- This protocol emphasises the innovative approach of combining data from a prospective cohort study and a randomised controlled trial with administrative health records to examine patterns and predictors of healthcare utilisation among adults with a history of homelessness.
- This approach is uniquely achievable due to the overlap in the eligibility criteria for the At Home/Chez Soi (AH/CS) and Health and Housing in Transition studies, and the similarities between the AH/CS intervention and community services available to participants.
- The study overcomes limitations of past research by joining individual-level demographic, housing and health status information with longitudinal administrative healthcare data.
- This will be the first study to develop a prediction model that identifies high-cost users of the health system exclusively among adults with a history of homelessness, which will supplement existing models in the general population.
- There are challenges to merging two distinct studies, including measurement and timeline differences, yet the ability to examine these cohorts together and separately will enable a more comprehensive examination of healthcare utilisation among a diverse group of adults with a history of homelessness.

INTRODUCTION

Homelessness is a global issue with a detrimental impact on physical and mental health.[1] Inability to obtain timely and adequate healthcare due to housing instability or mental health challenges may promote the progression of illness and lead to subsequent emergency department (ED) visits and hospitalisations that may otherwise have been prevented.

Homeless individuals are often considered frequent users of the health system, yet it is a small group of individuals who disproportionately use the majority of services.[1] Findings from a 4-year cohort study reported homeless

individuals were 1.76 (95% CI 1.58 to 1.96) times more likely to visit a physician, 8.48 (95% CI 6.72 to 10.70) times more likely to visit the ED, 4.22 (95% CI 2.99 to 5.94) times more likely to have a medical-surgical hospitalisation and 9.27 (95% CI 4.42 to 19.43) times more likely to have a psychiatric hospitalisation when compared with low-income controls over an average follow-up period of 3.9 years.[2] In this same study, 10% of the homeless cohort accounted for 43% of physician visits, 60% of ED visits, 80% of medical-surgical hospitalisations and 86% of psychiatric hospitalisations.[2]

Frequency of encounters is one component of healthcare utilisation. A number of studies examine factors associated with higher frequency of ED visits, with few studies examining hospitalisations and physician visits, among homeless individuals.[3–13] It has been reported that predisposing factors, such as age, gender, ethnic identity, housing stability, mental illness and substance use, are associated with hospitalisations, ED visits, and physician visits,[3–13] and criminalisation and victimisation with ED visits.[4 9–11 13] Enabling factors such as insurance and regular source of care are related to hospitalisations, ED visits and physician visits[3–7 9 11] while barriers to care,[12] social services[8] and competing needs (eg, food insecurity)[5] are associated with ED visits and hospitalisations. Need factors, such as perceived health and evaluated health conditions, are associated with hospitalisations, ED visits and physician visits.[3 4 7–11]

This collective literature highlights many factors to consider when examining healthcare encounters; however, these studies are not without limitations. First, many studies measured exposures and outcomes simultaneously, which made it difficult to establish temporality.[3–9 11] Second, healthcare utilisation was often measured by self-report without verification by medical charts.[3–13] This contributed to arbitrary and inconsistent definitions for healthcare utilisation being used across studies (eg, ≥1 ED visit,[3 4 6 9 13] ≥3 ED visits,[5 8] ≥4 ED visits).[4 7 12] Third, most of the literature focuses on ED visits, with less research on psychiatric and non-psychiatric hospitalisations or physician visits. To identify determinants of healthcare utilisation by adults with a history of homelessness, a next step is to model factors associated with frequency of utilisation ascertained from objective and complete administrative healthcare records.

Healthcare cost is another useful measure of the frequency and intensity of care utilisation. Risk prediction algorithms have been developed to predict who will become future 'high cost users', defined as the top 5% of users in the general population.[14–19] Many of the existing models were developed using a limited set of administrative variables that did not include pertinent individual-level characteristics (eg, sociodemographics).

Data linkage between survey data and administrative data is a promising way to obtain information on individual characteristics and healthcare utilisation. For instance, the Canadian Community Health Survey (CCHS) has been linked to administrative data to predict high-cost users in Ontario, Canada.[14] However, the CCHS is a household survey that does not represent homeless individuals. By excluding a group of individuals who are considered frequent users of health services, these models are likely to underestimate future healthcare resources and system costs. Since a minority of homeless individuals (10%) use a majority of services (>50%),[2] it is necessary to develop a model that can identify individuals who will be high-cost users in the next 5 years. In doing so, targeted efforts can be made at the individual level to improve health (or disease management) and reduce avoidable acute service use.

Objectives

1. To describe the distribution of healthcare utilisation by adults with a history of homelessness, and to identify the predisposing, enabling and need factors associated with hospitalisations, ED visits and physician visits in the following year.
2. To describe the patterns of healthcare costs by adults with a history of homelessness, and to characterise higher cost users of the health system by predisposing, enabling and need factors.
3. To develop and validate a risk prediction model to identify high-cost users or recurrent high-cost users over the next 5 years among a cohort of adults with a history of homelessness.

METHODS AND ANALYSIS

The proposed study is a cohort design based on prospective data from the At Home/Chez Soi (AH/CS) study and the Health and Housing in Transition (HHiT) study linked with administrative health records at ICES (formerly the Institute for Clinical Evaluative Sciences).

AH/CS was a randomised controlled trial conducted from 2009 to 2013 in five Canadian cities: Toronto, Vancouver, Winnipeg, Montreal and Moncton.[20] At enrolment (2009–2011), participants were at least 18 years of age, absolutely homeless or precariously housed and living with a mental illness. *Absolute homelessness* was defined as a 'lack of regular, fixed, or physical shelter' such as sleeping outside in public places or residing in emergency shelters. *Precarious housing* included a primary residence of a single room occupancy, rooming house, a hotel or motel with two or more episodes of absolute homelessness in the past year. Participants were stratified by need level and ethnoracial status and then randomised to receive Housing First or treatment as usual. Interviews were conducted every 6 months over a 2-year period. Data collection included sociodemographic characteristics (eg, age, gender, ethnic identity, marital status), housing history, mental illness, resources and health conditions.[20–22]

HHiT was a longitudinal cohort study of homeless and vulnerably housed single adults (18 years or older) living in Toronto, Ottawa and Vancouver. *Homelessness* was defined as current residence in a shelter, public

space, vehicle, abandoned building or someone else's house. *Vulnerable housing* was current residence in one's own room, apartment or place with an experience of homelessness or at least two moves in the past year. Eligible participants were recruited between January and December 2009 and followed until February 2013.[23 24] Interviews were conducted every 12 months over 4 years to collect sociodemographic, resource and health data. Participants were reimbursed $C20 for each interview.[23]

The AH/CS and HHiT studies aimed to retain 80% of the sample over the study period. To increase the rate of retention, interviewers made efforts to establish rapport with participants and emphasised the importance of their involvement.[20 23]

The Registered Persons Database (RPDB) is a population registry of all Ontario residents who are eligible for health insurance coverage under the Ontario Health Insurance Plan (OHIP). Personal identifiers, such as health card number, are used to assign a unique ICES key number (IKN) for linkage across internal and external data sets. Specifically, data from the AH/CS and HHiT studies can be linked to administrative records if the participants consented to linkage and provided a health card number that corresponds to a valid IKN in the RPDB. This key number also enables internal linkage with other databases including the Canadian Institute for Health Information Discharge Abstract Database (CIHI-DAD), Ontario Mental Health Reporting System (OMHRS), National Ambulatory Care Reporting System (NACRS) and OHIP.

Survey data are linked with administrative health records from 1 year before the date of enrolment (index date) to 5 years after the date of enrolment (end of follow-up). The baseline AH/CS and HHiT survey data and administrative data for 1 year before index were used as independent variables. Administrative data were collected annually to ascertain healthcare utilisation outcomes up to 5 years after index.

Independent variables

The Behavioral Model for Vulnerable Populations is the theoretical basis for the proposed research.[25] *Predisposing factors* include demographics (age, gender, marital status), social structure characteristics (ethnic identity, education, employment), housing history, mental illness, substance use, criminal behaviour and victimisation. *Enabling factors* contain personal and community resources (region of residence, regular source of care, perceived barriers to care). *Need factors* include perceived health status and observed health conditions. Table 1 describes the complete list of predictor variables, with mental health diagnostic codes reported in online supplemental table 1.[26–29]

Outcome variables

Healthcare utilisation includes the number of healthcare encounters and healthcare costs per person over a specified time period. Healthcare visits will be identified from ICES administrative databases, including OMHRS, NACRS, CIHI-DAD and OHIP (table 2). Total costs per person are estimated using an individual-level costing macro based on all healthcare databases at ICES.

Healthcare encounters will be separated into hospitalisations, ED visits and physician visits. Hospitalisations include psychiatric and non-psychiatric inpatient hospitalisations, identified in OMHRS and CIHI-DAD. For descriptive purposes, the number of hospitalisations will be categorised as non-users (0 visit), moderate users (1–2 visits) or high users (≥3 visits) in a given year.[30] For the main analysis, hospitalisations will be modelled either as a binary or count variable as appropriate.

ED visits are ascertained from hospital records recorded in NACRS. For descriptive analysis, the number of ED visits will be categorised as non-users (0 visit), moderate users (1–4 visits) or high users (≥5 visits) in a given year. While no widely used method for classifying ED use exists, these groups are based on a systematic review of frequent users and other reports in the homeless population.[31] The number of ED visits will be modelled as a count variable for the main analysis.

Physician visits include primary care, psychiatrist and other medical specialist visits reported in OHIP. The number of physician visits in a given year will be classified as non-users (0 visit), moderate users (1–4 visits) or high users (≥5 visits).[30] The frequency of physician visits will also be modelled as a count variable for the main analysis.

As a secondary analysis, ED visits and hospitalisations from ambulatory care sensitive conditions (ACSC) will be explored, as these visits are considered avoidable with adequate primary care. According to the CIHI, ACSCs include grand mal status and other epileptic seizures, chronic obstructive pulmonary disease, asthma, heart failure and pulmonary oedema, hypertension, angina and diabetes.[32] Relevant diagnostic codes are listed in online supplemental table 2.[32] ACSC-related ED visits and hospitalisations will be measured as the total number of visits in a given year.

Total annual healthcare costs for each participant will be calculated with a person-level validated costing algorithm at ICES.[33] This algorithm combines the frequency of healthcare encounters with intensity of resource utilisation based on a weighted per unit cost. Total annual healthcare costs for each participant will be calculated for the 5 years following index date. Each year, the total costs for each participant will be categorised as the bottom 0%–50%, top 11%–50%, top 6%–10%, top 2%–5% and top 1% of healthcare users in Ontario according to predetermined cut-offs from a general population sample.[14 34]

Participants will further be classified as high-cost users in a given year if their total annual healthcare expenditure is above the cut-off for the top 5% of healthcare users in Ontario. If a participant is in the top 5% of users for at least 2 years over the 5-year follow-up, they will be classified as a recurrent high-cost user.

Table 1 Survey questions from the At Home/Chez Soi (AH/CS) study and Health and Housing in Transition (HHiT) study interviews and the predictor definitions for participants linked to administrative data[20 23]

Factors	AH/CS	HHiT	Combined predictor definition
Predisposing			
Age	What is your date of birth?	What is your date of birth?	Age at index in the ICES Registered Persons Database (RPDB) will be used. Administrative date of birth will be checked against survey date of birth.
Gender	What is your gender?	Your gender is…	Survey gender will be used. If data are missing, then sex from the RPDB will be used.
Marital status	Are you currently single… (list options)?	What is your marital status?	Marital status is categorised into single; widowed, separated or divorced; and married or partnered.
Ethnic identity	What is your ethnic or cultural identity?	To which racial or cultural group(s) do you belong?	Ethnic identity is based on self-report of ethnoracial status. In the AH/CS study, individuals who identified as Asian, Black, Latin American, Indian-Caribbean, Middle Eastern or mixed background were classified as part of an ethnoracial minority. The same set of criteria was applied for HHiT.
Education	What is your level of education?	How much school have you completed?	Education is dichotomised as either having graduated from high school or having less than high school education.
Employment	What is your current primary employment status?	In the past 12 months, did you work at a paid job? Are you currently working?	Classified as being employed part-time or full time, and not employed.
Housing status	Homeless or precariously housed at enrolment; Residential Time-Line Follow-Back Inventory (RTLFB)[48]	Homeless or vulnerably housed at enrolment; RTLFB	Classified as homeless or housed based on enrolment criteria; time spent homeless in past 6 months.
Duration of homelessness	In your lifetime, what is the total amount of time you have been homeless (months)?	…if you add up all the times in your life, how many weeks, months or years have you been homeless (years)?	Duration of homelessness is classified as the total amount of time spent homeless in years.
Cigarette smoking	At the present time, do you smoke cigarettes daily, occasionally or not at all?	How often do you smoke?	Cigarette smoking is classified as being a current daily smoker versus an occasional smoker or non-smoker.
Criminal behaviour	In the past 6 months, have you been arrested? … have you been arrested for criminal activity more than once, or been imprisoned at least once, or served probation or other community sanction?	In the past 12 month, were you arrested by the police? Incarcerated, whether in preventive detention, prison or a penitentiary? If yes, how many times?	Criminal behaviour is classified as having been arrested or incarcerated for criminal activity at least once in the past 6–12 months versus no criminal behaviour.
Victimisation	During the past 6 months, did anyone threaten to hit or attack you, or threaten you with a weapon? Has anyone forced you or attempted to force you into any unwanted sexual activity, by threatening you, holding you down or hurting you in some way?	In the past 12 months were you beaten up or physically attacked? Did anyone force you or attempt to force you into any unwanted sexual activity? If yes, how many times?	Victimisation is classified as having experienced physical or sexual victimisation at least once in the past 6–12 months versus no victimisation.

Continued

Table 1 Continued

Factors	AH/CS	HHiT	Combined predictor definition
Mental illness (baseline)	Mini-International Neuropsychiatric Interview (MINI)[49]	Self-reported mental illness	For consistency, mental illness is defined with mental health diagnostic codes from administrative data in the 1 year before baseline.[26–29] *Any mental illness*: ≥1 mental health diagnostic code. *Type of mental illness*: psychotic disorder (schizophrenia and related disorders), other mental disorder (eg, affective disorders, personality disorders) or no mental disorder (no diagnostic code). Ideally, the past 3 years would be used to identify[29] psychotic disorder; however, data restrictions only enable 1 year of data before baseline data to be linked. The MINI will be used for comparison and to supplement the administrative data for AH/CS participants.
Problematic alcohol use (past year)	MINI Alcohol abuse or dependence	Alcohol Use Disorders Identification Test (AUDIT)[50]	Problematic alcohol use was classified as either a flag for alcohol abuse or dependence for AH/CS participants or an AUDIT score of ≥20 for HHiT participants. Survey data were supplemented with administrative health records using diagnostic codes for alcohol use disorders in the 1 year before baseline.
Problematic drug use (past year)	MINI Drug abuse or dependence	Drug Abuse Screening Test (DAST-10)[51]	Problematic drug use was classified as either a flag for drug abuse or dependence for AH/CS participants or a DAST-10 score of ≥6 for HHiT participants. Survey data were supplemented with administrative health records using diagnostic codes for drug use disorders in the 1 year before baseline.
Enabling			
Region	Toronto	Toronto or Ottawa	Region is classified as the location of interview, either Toronto or Ottawa.
Regular source of care	Do you have a regular medical doctor?	Do you have a regular medical doctor?	Regular source of care is classified as reporting a regular medical doctor versus not reporting a regular medical doctor.
Perceived barrier to care	In the past 6 months, was there ever a time when you felt that you needed healthcare but you didn't receive it? Thinking of the most recent time, why didn't you get care?	(In the past 12 months) Have you needed care but were not able to get help? What were the reasons you were unable to get help?	Perceived barrier to care is classified as having needed care that was not received in the past 6–12 months versus no barrier to care.
Food insecurity	(In the past 30 days) Do you ever eat less than you feel you should because you can't get enough food?	Do you have trouble getting enough to eat?	Food insecurity is measured as an indicator of not having enough to eat.
Need factors			
Perceived general health	In general, would you say your health is poor, fair, good, very good, excellent?	In general, would you say your health is poor, fair, good, very good, excellent?	Perceived general health is categorised into poor, fair, and good or very good or excellent.

Continued

Table 1 Continued

Factors	AH/CS	HHiT	Combined predictor definition
Physical health conditions	(Asks about a series of conditions) Do you have ANY other serious medical condition? If yes, what is it?	Do you have any of the following medical conditions?	Physical health conditions were classified as the number of chronic conditions, including asthma, hypertension, myocardial infarction or heart disease, chronic obstructive pulmonary disorder, bowel disorders, HIV, diabetes and cancer. Two variables will be derived from self-reported conditions from survey data and diagnosed conditions from administrative data.

ICES, formerly Institute for Clinical Evaluative Sciences.

Statistical analysis

Objective 1

Baseline demographic characteristics will be described for the AH/CS and HHiT cohorts. Individuals who were successfully linked to ICES data will be compared with those who were not. Predisposing, enabling and need factors will be examined across levels of hospitalisations, ED visits or physician visits (none, moderate, high) using χ^2 tests or analysis of variance. Bivariate and multivariable count models will identify factors associated with higher rates of ED visits, hospitalisations and physician visits in the following year.

Poisson regression assumes the mean and variance of the model are equal.[35] Common violations of this assumption are overdispersion, or when the data exhibit more zeros than is to be expected under a Poisson distribution. If either violation is evident, alternative models such as the negative binomial distribution or zero-inflated models will be considered.[35] An offset term will be incorporated to account for any difference in observation time between participants. Time at risk for each period will be defined as 365 days, or the number of days between each data collection and end of coverage date (eg, death). Continuous variables will be examined for deviations from linearity.

The primary analyses will report unadjusted associations in order to depict real-world circumstances where characteristics are interconnected, instead of adjusting away important effects.[36 37] For comparison, the fully adjusted and imputed models will also be reported. In these multivariable models, all prespecified variables will be retained despite significance.

Objective 2

Participants will be categorised as incurring costs within the top 1%, 2%–5%, 6%–10%, 11%–50% and bottom 50% in the 1 year after baseline, based on the approach by Rosella *et al*[14]. Individual characteristics and costs by type of encounter (eg, inpatient, ED, outpatient, physician, prescription medications, laboratory, other services) will be compared across cost categories. If necessary, the top 1% and top 2%–5% groups will be combined. Ordinal logistic regression will identify factors associated with higher levels of healthcare expenditure using 'bottom 0%–50% of health

Table 2 Definitions for healthcare utilisation variables and ICES administrative databases

Healthcare utilisation	ICES administrative database	Outcome definition
Psychiatric hospitalisations	Ontario Mental Health Reporting System (OMHRS); Canadian Institute for Health Information Discharge Abstract Database (CIHI-DAD)	Inpatient admissions to designated adult mental health beds, as identified in OMHRS, and other psychiatric inpatient admissions, as identified by the most responsible diagnosis in CIHI-DAD. The frequency of psychiatric hospitalisations each year was ascertained as a count variable.
Non-psychiatric hospitalisations	CIHI-DAD	Inpatient admissions for medical conditions or surgical procedures as identified by the most responsible diagnosis in CIHI-DAD. The frequency of non-psychiatric hospitalisations each year was ascertained as a count variable.
Emergency department visits	National Ambulatory Care Reporting System (NACRS)	The frequency of emergency department visits each year was ascertained as a count variable, as identified from the hospital ambulatory care records in NACRS.
Physician visits	Ontario Health Insurance Plan (OHIP)	Physician fee-for-service billings and shadow billings reported in OHIP including primary care physicians, psychiatrists and other medical specialists. The frequency of physician visits each year was ascertained as a count variable.

ICES, formerly Institute for Clinical Evaluative Sciences.

care users' as the reference group. If the proportional odds assumption is violated (eg, Brant test, α=0.05), a less restrictive multinomial model will be used.[38]

Objective 3

Risk prediction modelling will be used to develop a model that predicts high-cost users of the health system over 1-year and 5-year periods. A third model will be developed to predict individuals who are high-cost users for at least 2 years over the 5-year period. This is to separate recurrent high-cost users from individuals who may experience regression to the mean, a phenomenon where high healthcare users exhibit more normalised use over time.[39]

Logistic regression will be used to develop these prediction models. Ideally, the models would be developed in one sample and externally validated in another, however this is not feasible due to barriers accessing out-of-province data (eg, Vancouver site of HHiT or AH/CS studies). Rather, a bootstrapping technique will be applied to derive unbiased estimates of model performance.[40] In this method, the prediction model will be developed in the original sample and validated in a series of bootstrapped samples with replacement. The following model building strategy, based on the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement[41] and work by Harrell and Steyerberg,[42 43] will be carried out for 200 bootstrapped samples.

The initial model will be prespecified based on predictors identified from relevant literature.[3–11] To avoid overfitting the model, data reduction techniques will be explored to limit the df used in developing the model.[42 43] To test the assumption of linearity between the predictors and log odds of the outcome, the functional form of continuous variables will be assessed first with a smoothed residual plot, and then with and without expanded predictor terms (eg, quadratic, cubic splines) to test for inclusion in the model.[40 42] The interactions that are expected to violate the additivity assumption, such as mental illness and substance use,[10] will be examined by a single likelihood ratio test. All interactions will be tested at once to reduce type 1 error.[40] Overly influential observations (outliers) will be assessed by leverage measures, such as dfbetas and dfit statistics.[43] The above steps will be repeated for each of the bootstrapped samples to derive a measure of optimism (bootstrap performance—test performance).[42]

Overall model performance will be estimated using the R^2 statistic, a measure of the variation explained by the model, and the Brier score, a quantity of the squared difference between observed outcomes and predictions.[44] Model performance will further be described by calibration, the agreement between predicted and observed outcomes, and discrimination, the ability for predictions to separate individuals with and without the outcome. An estimate of calibration will be derived from the Hosmer-Lemeshow χ^2 statistic. A visual plot of predicted versus observed outcomes demonstrates the level of calibration, with perfect calibration following a 45° line. Discrimination will be estimated by the concordance (*c*) statistic, which is equal to the area under the receiver operating characteristic curve.[44] Concurrently, a box plot will visually compare overlap in predictions for individuals who are high-cost users (or recurrent high-cost users) and those who are not.[45] Model performance estimates for the R^2 statistic, Brier score, Hosmer-Lemeshow χ^2 statistic and c statistic will be corrected for with bootstrapped optimism (apparent model performance—optimism) to ensure the model performance is not overestimated.[42]

Missing data considerations

Missing data will be examined to identify which variables are simultaneously missing and to describe the distribution of the outcome for individuals with missing predictor values. Multiple imputation methods will be applied as appropriate.[43] According to the survey data, the variable with the most missing data is the duration of time spent homeless for AH/CS (2.3%) and HHiT (3.3%). Demographics (eg, age and gender) and administrative variables have no missing data.

Sample size considerations

Based on independent linkage of the AH/CS and HHiT studies with administrative records, we expect a combined linkage rate of at least 80%, which will result in 1100 participants linked to administrative data. According to a recent study, there are four steps to calculate the sample size, including a calculation to ensure precise estimates (margin of error=0.05), predicted values with a small mean error across individuals (mean absolute prediction error=0.05), a small required shrinkage of predictor effects (shrinkage=0.1) and a small optimism in apparent model fit (optimism=0.05).[46] Table 3 reports the estimated number of predictors that can be modelled with a sample of 1100 participants under various conditions. Assuming the model explains 20% of variance, 12 predictors can be estimated with an outcome proportion of 10% or 16 predictors with an outcome proportion of 20%. For models that include direct or mechanistic measurements (where 50% of variance is explained), 33 or 47 parameters can be estimated with a 10 or 20% outcome proportion. The proportion of high-cost users is expected to be higher than 5% in the homeless samples due to the elevated use of acute services compared with non-homeless groups.[2]

Patient and public involvement

This is a secondary analysis of data from a cohort study and randomised controlled trial which were designed with input from people with lived experience of homelessness. The current analysis does not involve the public in the design, conduct, reporting or dissemination. Findings will be shared in reports and at scientific meetings attended by researchers, policymakers and people with lived experience.

Table 3 The margin of error and number of predictors that can be estimated given a sample size of 1100 participants, based on four sample size requirements outlined by Riley *et al*[46]

Step 1: To produce a precise estimate (margin of error <0.05)						
Outcome proportion†	0.1	0.2				
Margin of error	0.018	0.024				
Step 2: To produce predicted values with a small mean error across all individuals (mean absolute prediction error=0.05)						
Outcome proportion†	0.1	0.2				
Number of parameters	44	31				
Step 3: To produce a small required shrinkage of predictor effects (shrinkage=0.1)						
$R^2_{Nagelkerke}$*	0.1	0.2	0.5	0.1	0.2	0.5
Outcome proportion†	0.1	0.1	0.1	0.2	0.2	0.2
Number of parameters	6	12	33	8	16	47
Step 4: To produce a small optimism in apparent model fit (optimism=0.05)						
$R^2_{Nagelkerke}$*	0.1	0.2	0.5	0.1	0.2	0.5
Outcome proportion†	0.1	0.1	0.1	0.2	0.2	0.2
Number of parameters	27	28	30	36	37	42

*In absence of existing studies for a similar target population, Riley *et al*[46] suggest an $R^2_{Nagelkerke}$ value between 0.1 and 0.2 for prediction models of health-related outcomes. When direct or mechanistic measurements are included, the $R^2_{Nagelkerke}$ may be closer to 0.5.
†Based on existing research that examines healthcare use among homeless adults, we expect the outcome proportion will be greater than in the general population.[2]

Strengths and anticipated challenges

This opportunity to combine data from two large cohorts of adults with a history of homelessness is uniquely achievable due to the overlapping eligibility criteria. The HHiT study follows a cohort of homeless and vulnerably housed individuals, sampled from the general homeless population. Of these participants, nearly 50% report having a mental illness. While the AH/CS study exclusively enrolled individuals with a severe mental illness, these individuals represent a subset of individuals eligible for the HHiT study. In fact, given the overlap in eligibility criteria, it is likely that some of the participants in HHiT were enrolled in AH/CS. It will be important to conduct certain analyses separately, particularly for absolute measures of quantity and distribution of healthcare utilisation; however, most analyses examine relative measures which will be less affected by the elevated proportion of individuals with a mental illness.

Another consideration is the AH/CS intervention. There are many resources available in large Canadian cities, such as Toronto and Ottawa, that offer similar services to the housing intervention received in the AH/CS trial (eg, housing services, case management).[47] Services are constantly evolving, which makes it challenging to adequately capture the services received by participants in any study, irrespective of whether the study is observational or experimental. We will conduct sensitivity analyses to ensure the presence of an intervention does not alter the results.

The ability to link individuals to administrative data overcomes the common limitation of relying on self-reported healthcare utilisation data and ensures all visits for each linked individual will be captured in the analyses. While past research has dichotomised or categorised healthcare encounters, data linkage permits the use of count data to describe a comprehensive list of predisposing, enabling and need factors associated with higher rates of encounters. Finally, by establishing this large, merged cohort of homeless adults, there are opportunities for future research to examine how time-varying individual-level characteristics affect healthcare utilisation among people experiencing homelessness, who can be a challenging group to follow over time.

There are also some anticipated challenges in combining the AH/CS and HHiT cohorts, as they are two distinct studies with different measurements and timelines. First, it is necessary to restructure the data to align data collection periods and survey questions. The index date will be based on study enrolment and healthcare utilisation will be examined on an annual basis starting at index date. Further, some variables with differing time periods will need to be integrated (eg, criminal behaviour and victimisation in the past 6–12 months).

Second, only participants who consented to linkage and provided a health card can be linked to administrative data. Therefore, participants without a valid health card number (eg, recent immigrants or refugees) or individuals who did not consent to linkage will be excluded. Out-of-province encounters will not be captured either; however, this is likely to be minimal as study participants were contacted annually at minimum from an Ontario location.

Third, mental illness was measured with a diagnostic interview only among AH/CS participants, so to ensure consistency, administrative codes will be used to classify mental illness for both studies. This method cannot identify individuals with a mental illness who did not access health services, and individuals who use services regularly are more likely to be identified as having a mental illness (ie, ascertainment bias). To explore the extent of misclassification among AH/CS participants, the administrative data diagnosis can be compared with the Mini-International Neuropsychiatric Interview diagnostic interview.

Fourth, it is best practice to validate risk prediction models in an external sample; however, this is not feasible

due to barriers accessing out-of-province data. Instead, bootstrapped validation will be used to validate the model as a superior technique to split sample methods.[40 41] There is potential for future external validation using HHiT and AH/CS data from Vancouver, British Colombia.

Implications

The risk prediction models will be useful in conjunction with other general population-based models to improve resource planning. By comprehensively examining the factors associated with healthcare utilisation (objectives 1 and 2) and identifying the individuals who will become high-cost users (objective 3), this collective work can inform targeted healthcare-driven housing and support interventions for people experiencing homelessness which may improve health outcomes and reduce avoidable acute service use downstream.

Ethics and dissemination

This research links deidentified survey data from AH/CS and HHiT to administrative data at ICES, a prescribed entity under the Personal Health Information Act. ICES has rigorous security and privacy practices in place to protect healthcare data. The Unity Health Toronto Research Ethics Board (previously the St Michael's Hospital Research Ethics Board) approved data linkage of the AH/CS and HHiT survey to administrative databases at ICES. Additional approval was received from the Unity Health Toronto Research Ethics Board to combine the AH/CS and HHiT studies linked to administrative health records. The proposed work will be disseminated through publication in peer-reviewed journals, and presentations at conferences that invite researchers, healthcare professionals and people with lived experience. These findings will also be shared in monthly newsletters for healthcare providers and on public platforms to reach the general public.

Author affiliations
[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[2]MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada
[3]Institute for Mental Health Policy Research, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[4]Division of General Internal Medicine, Department of Medicine, University of Toronto, Toronto, Ontario, Canada

Twitter Laura C Rosella @LauraCRosella

Contributors KW and SWH conceptualised the study. LCR and PK advised on the development of the study protocol and methods. KW wrote the manuscript and incorporated coauthor feedback. All authors reviewed the final protocol prior to its submission.

Funding The At Home/Chez Soi study was supported by the Mental Health Commission of Canada; the Ontario Ministry of Health and Long-Term Care (HSRF-259); and the Canadian Institute of Health Research (CIHR MOP-130405 and CIHR FDN-167263) received by SWH and colleagues. The Health and Housing in Transition study was supported by an operating grant (MOP-86765) and an Interdisciplinary Capacity Enhancement Grant on Homelessness, Housing and Health (HOA-80066) from the Canadian Institutes of Health Research.

Disclaimer The funders had no role in the analysis and interpretation of the data or the preparation, review and approval of the manuscript. The views expressed in this publication are the views of the authors and do not necessarily reflect those of the funders.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Kathryn Wiens http://orcid.org/0000-0003-3451-6788
Laura C Rosella http://orcid.org/0000-0003-4867-869X

REFERENCES

1 Fazel S, Geddes JR, Kushel M. The health of homeless people in high-income countries: descriptive epidemiology, health consequences, and clinical and policy recommendations. *Lancet* 2014;384:1529–40.
2 Hwang SW, Chambers C, Chiu S, *et al*. A comprehensive assessment of health care utilization among homeless adults under a system of universal health insurance. *Am J Public Health* 2013;103 Suppl 2:S294–301.
3 Kushel MB, Vittinghoff E, Haas JS. Factors associated with the health care utilization of homeless persons. *JAMA* 2001;285:200–6.
4 Kushel MB, Perry S, Bangsberg D, *et al*. Emergency department use among the homeless and marginally housed: results from a community-based study. *Am J Public Health* 2002;92:778–84.
5 Kushel MB, Gupta R, Gee L, *et al*. Housing instability and food insecurity as barriers to health care among low-income Americans. *J Gen Intern Med* 2006;21:71–7.
6 Weber EJ, Showstack JA, Hunt KA, *et al*. Does lack of a usual source of care or health insurance increase the likelihood of an emergency department visit? results of a national population-based study. *Ann Emerg Med* 2005;45:4–12.
7 Hunt KA, Weber EJ, Showstack JA, *et al*. Characteristics of frequent users of emergency departments. *Ann Emerg Med* 2006;48:1–8.
8 Moore DT, Rosenheck RA. Factors affecting emergency department use by a chronically homeless population. *Psychiatr Serv* 2016;67:1340–7.
9 Small LFF. Determinants of physician utilization, emergency room use, and hospitalizations among populations with multiple health vulnerabilities. *Health* 2011;15:491–516.
10 Zhang L, Norena M, Gadermann A, *et al*. Concurrent disorders and health care utilization among homeless and Vulnerably housed persons in Canada. *J Dual Diagn* 2018;14:21–31.
11 Stergiopoulos V, Gozdzik A, Nisenbaum R, *et al*. Racial-Ethnic differences in health service use in a large sample of homeless adults with mental illness from five Canadian cities. *Psychiatr Serv* 2016;67:1004–11.
12 Brown RT, Kiely DK, Bharel M, *et al*. Use of acute care services among older homeless adults. *JAMA Intern Med* 2013;173:1831–4.
13 Jaworsky D, Gadermann A, Duhoux A, *et al*. Residential stability reduces unmet health care needs and emergency department utilization among a cohort of homeless and Vulnerably housed persons in Canada. *J Urban Health* 2016;93:666–81.
14 Rosella LC, Kornas K, Yao Z, *et al*. Predicting high health care resource utilization in a single-payer public health care system: development and validation of the high resource user population risk tool. *Med Care* 2017;56:e61–9.
15 Chechulin Y, Nazerian A, Rais S, *et al*. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). *Healthc Policy* 2014;9:68–79.

16 Chang H-Y, Boyd CM, Leff B, *et al*. Identifying consistent high-cost users in a health plan: comparison of alternative prediction models. *Med Care* 2016;54:852–9.
17 Frost DW, Vembu S, Wang J, *et al*. Using the electronic medical record to identify patients at high risk for frequent emergency department visits and high system costs. *Am J Med* 2017;130:601.e17–601.e22.
18 Lauffenburger JC, Franklin JM, Krumme AA, *et al*. Longitudinal patterns of spending enhance the ability to predict costly patients: a novel approach to identify patients for cost containment. *Med Care* 2017;55:64–73.
19 Hu Z, Hao S, Jin B, *et al*. Online prediction of health care utilization in the next six months based on electronic health record information: a cohort and validation study. *J Med Internet Res* 2015;17:e219.
20 Goering PN, Streiner DL, Adair C, *et al*. The at Home/Chez Soi trial protocol: a pragmatic, multi-site, randomised controlled trial of a housing first intervention for homeless individuals with mental illness in five Canadian cities. *BMJ Open* 2011;1:e000323.
21 Hwang SW, Stergiopoulos V, O'Campo P, *et al*. Ending homelessness among people with mental illness: the at Home/Chez Soi randomized trial of a housing first intervention in Toronto. *BMC Public Health* 2012;12:787.
22 Goering P, Veldhuizen S, Watson A, *et al*. *National at home/chez soi final report*. Calgary, AB: Mental Health Commission of Canada, 2014.
23 Hwang SW, Aubry T, Palepu A, *et al*. The health and housing in transition study: a longitudinal study of the health of homeless and vulnerably housed adults in three Canadian cities. *Int J Public Health* 2011;56:609–23.
24 To MJ, Palepu A, Aubry T, *et al*. Predictors of homelessness among vulnerably housed adults in 3 Canadian cities: a prospective cohort study. *BMC Public Health* 2016;16:1041.
25 Gelberg L, Andersen RM, Leake BD. The behavioral model for vulnerable populations: application to medical care use and outcomes for homeless people. *Health Serv Res* 2000;34:1273–302.
26 World Health Organization. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization, 1992.
27 Canadian Institute for Health Information. *Hospital mental health database data dictionary for fiscal year 2015-2016*. Ottawa: Canadian Institute for Health Information, 2017.
28 MHASEF Research Team. *The mental health of children and youth in Ontario: 2017 Scorecard. Technical appendix*. Toronto, ON: Institute for Clinical Evaluative Sciences, 2017.
29 Kurdyak P, Lin E, Green D, *et al*. Validation of a population-based algorithm to detect chronic psychotic illness. *Can J Psychiatry* 2015;60:362–8.
30 Bharel M, Lin W-C, Zhang J, *et al*. Health care utilization patterns of homeless individuals in Boston: preparing for Medicaid expansion under the Affordable care act. *Am J Public Health* 2013;103 Suppl 2:S311–7.
31 Moe J, Kirkland SW, Rawe E, *et al*. Effectiveness of interventions to decrease emergency department visits by adult frequent users: a systematic review. *Acad Emerg Med* 2017;24:40–52.
32 Canadian Institute for Health Information. Ambulatory care sensitive conditions, 2018. Available: http://indicatorlibrary.cihi.ca/display/HSPIL/Ambulatory+Care+Sensitive+Conditions
33 Wodchis WP, Bushmeneva K, Nikotovic M. *Guidelines on person-level costing using administrative databases in Ontario*. Toronto: Health System Performance Research Network, 2013.
34 Rosella LC, Fitzpatrick T, Wodchis WP, *et al*. High-Cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *BMC Health Serv Res* 2014;14:532.
35 Hilbe JM. *Modeling count data*. Cambridge: Cambridge University Press, 2014.
36 Kaufman JS. Statistics, adjusted statistics, and MALADJUSTED statistics. *Am J Law Med* 2017;43:193–208.
37 Conroy S, Murray EJ. Let the question determine the methods: descriptive epidemiology done right. *Br J Cancer* 2020;123:1351–2.
38 Vittinghoff E, Shiboski SC, Glidden DV, *et al*. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. New York, NY: Springer, 2005.
39 Georghiou T, Steventon A, Billings J. *Predictive risk and health care: an overview. Research summary*. London: Nuffield Trust, 2011.
40 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
41 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
42 Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer Science+Buisiness Media, LL, 2009.
43 Harrell FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ED*. Switzerland: Springer International Publishing, 2015.
44 Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J* 2008;50:457–79.
45 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
46 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
47 Falvo N. Toronto's Housing First programme and implications for leadership. *Housing, Care and Support* 2009;12:16–25.
48 Tsemberis S, McHugo G, Williams V, *et al*. Measuring homelessness and residential stability: the residential time-line follow-back inventory. *J Community Psychol* 2007;35:29–42.
49 Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. The Mini-International neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998;59 Suppl 20:22-33;quiz 34-57.
50 Saunders JB, Aasland OG, Babor TF, *et al*. Development of the alcohol use disorders identification test (audit): who Collaborative project on early detection of persons with harmful alcohol Consumption-II. *Addiction* 1993;88:791–804.
51 Skinner HA. The drug abuse screening test. *Addict Behav* 1982;7:363–71.