Training machine learning models to identify patients who have experienced homelessness from operational data

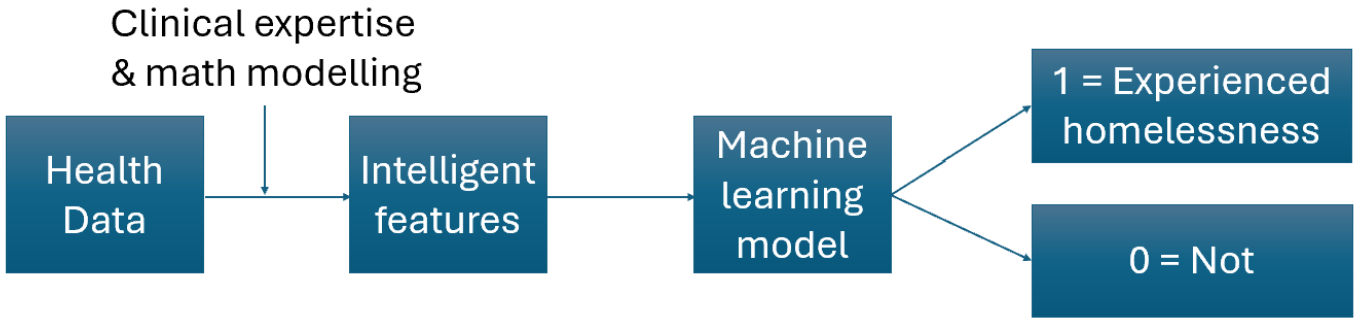
Jeremy Chiu, Langara College and Simon Fraser University

In collaboration with Vancouver Coastal Health

May 12, 2026

BCcupms Articulation 2026

Presentation outline



- Introduction
- Developing *intelligent* predictors of homelessness
  - Bonus: active learning quiz
- Training machine learning models to predict homelessness
  - Supervised vs semi-supervised methods
  - Validation
- Ongoing work and discussion
  - Address features
  - Gender analysis
  - Time series

Vancouver CoastalHealth

Ceinwen Pope
Medical Health Officer,
Public Health

Krisztina Vasarhelyi
Senior Learning
Health Systems,Lead

Ken Hawkins
Data & Analytics,
Manager

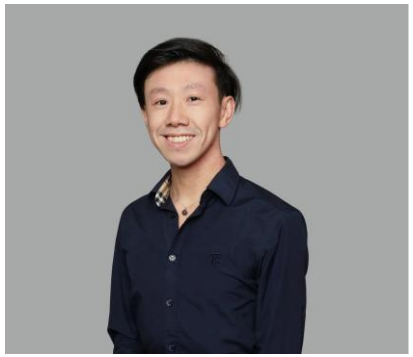
Amit Chalka
Data & Analytics,
Advisor

Yumian Hu
Public Health Surv Unit,
Senior Epidemiologist

Alexander Rutherford
Mathematics,
Adjunct Professor

Jessica Stockdale
Mathematics,
Assistant Professor

JF Williams
Mathematics,
Associate Professor



Jeremy Chiu
Applied Math PhD Candidate,
Langara Mathematics Instructor

SFU SIMON FRASER UNIVERSITY

Langara.
THE COLLEGE OF HIGHER LEARNING.

Background and goals

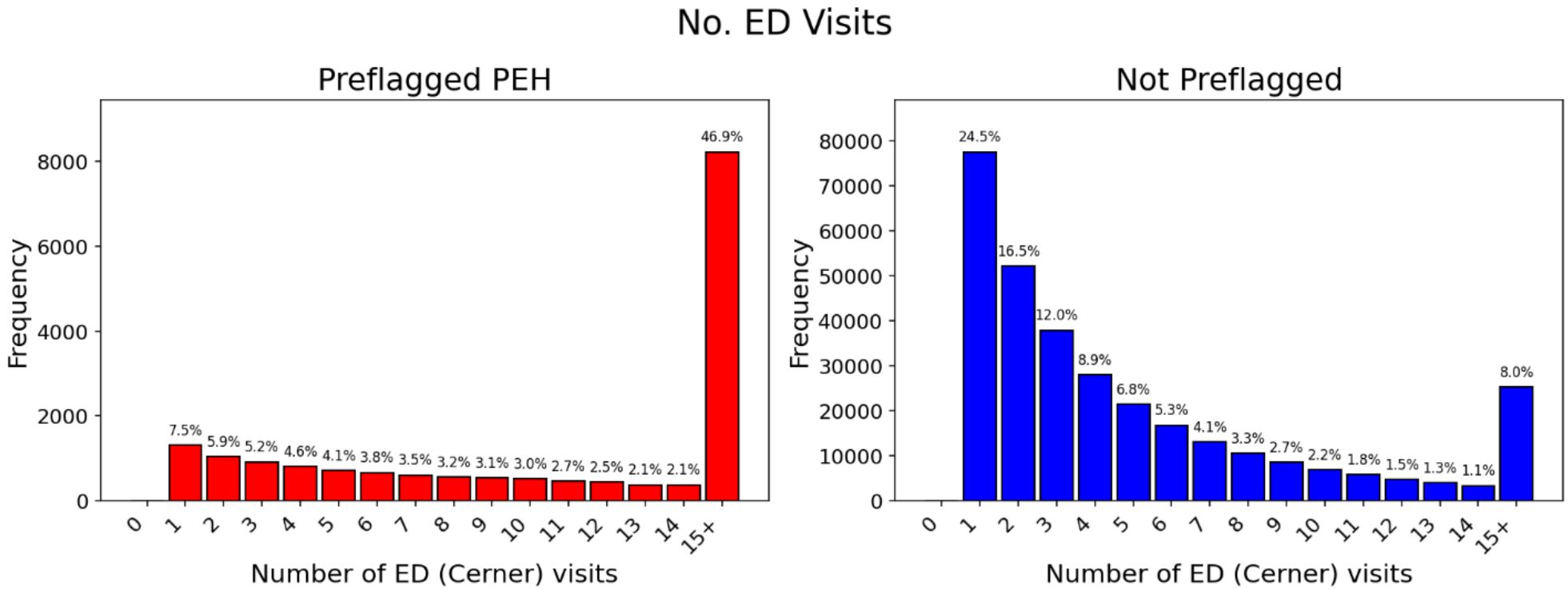
Homelessness is an ongoing and growing public health issue. People experiencing homelessness (**PEH**) face significant inequities and poor health outcomes.

Main goal: Develop a cohort of patients experiencing homelessness (PEH) who have engaged with VCH services. Sustainable internal cohort development using local health system data.

Other goals: Characterize the cohort: sociodemographics, health outcomes, episodes of homelessness (chronic, temporary, cyclical).

Why should we care?

PEH typically access health and emergency services far more than housed-individuals.

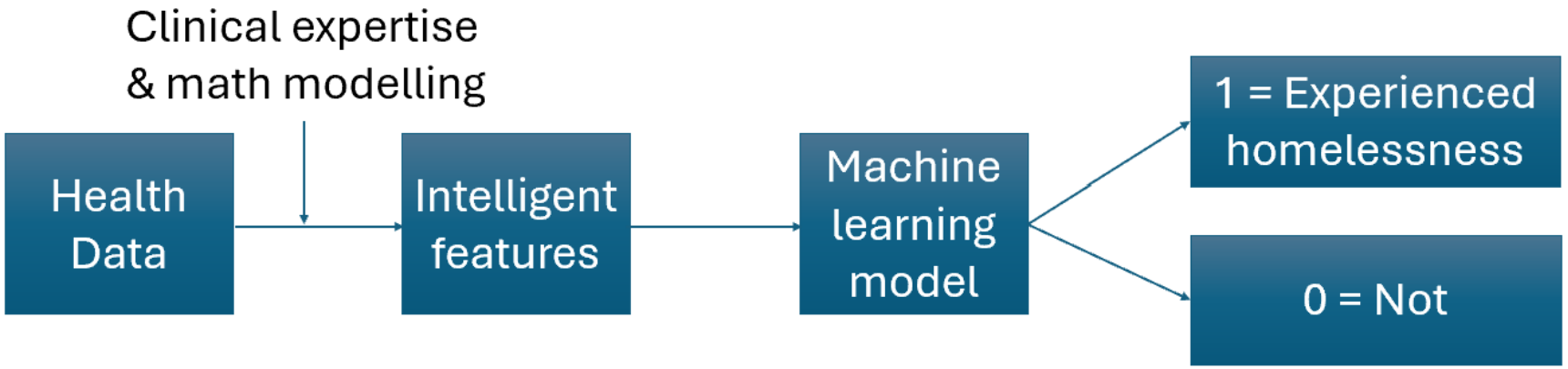


Characterizing the cohort enables 3 **rights**: the right preventative measure, to the right patient, at the right time and place.

	Data set	Data included
[1]	Cerner EMR – latest address	• Latest address (as of April 2025) of clients in Cerner EMR • **1,880,317 patients** with **1,880,586 addresses**
[2]	Paris/Profile – address history	• Client address history in PARIS/Profile EMR • **863,610 patients** with **1,168,774 addresses** • Addresses added to system from **2002-2023**
[3]	Discharge Abstract Database	• Discharge records (**2008-2025**) for clients whose discharge diagnosis included ICD-10 code Z59 • **14,381 patients** with **28,995 discharges** with Z59 code
[4]	Cerner EMR – emergency visit history	• Visits (**2007-2025**) in ED or UPCC for a filtered set of clients (presence of any *is_homeless* flag, at least 2 addresses in row 2, or Z59 discharge in row 3). • **614,763 patients, 3,797,339 visits**
[5]	Patient demographics	• **3,169,996 patients** year of birth and gender on 2025-04

Methodology

1. For each patient, turn their administrative healthcare data into *intelligent features*, clinically-motivated quantities predictive of homelessness.
2. Use patient's features as input to machine learning models that classify patient's as *1 = experienced homelessness* or 0 = *not*.
3. Interpret and validate.



Intelligent features

Recall key data sets:

1. ED visit history. Only postal code at visit (no address)
2. Community health address history.
3. Demographic data.

Case study of patients in all 3 sets.

Feature categories:

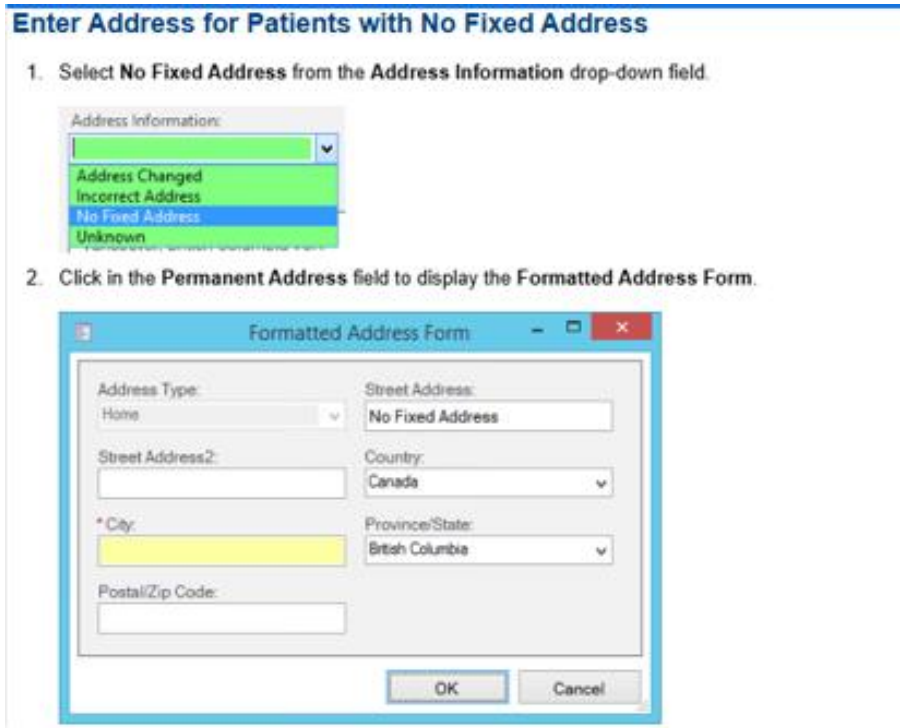
1. Counts and time
2. ED visit postal code features
3. Demographic and housing features

The **preflagged** cohort of PEH

VCH already has methods of indicating homelessness:

- Discharged with Z59.0 ICD10 code
- Address = *No Fixed Address*
- Address of shelter
- ED visit postal code = facility postal code
  [not in preflag]
- + more

These will be model labels.



Define the ***preflagged cohort*** of patients who have ever experienced.

Bonus: active learning quiz

Notation: For a given patient, let

- V be the set of all their ED visits
  - N be the total number of ED visits
  - W be the total number of ED visits where the postal code was well-formed
  - $v_k^{(date)}$ be date at kth visit
  - $v_k^{(PC)}$ be postal code at kth visit
  - $v_j^{(PC_W)}$ be postal code of jth visit with well-formed postal codes
- A be the set of all their addresses from community health
  - $A^{(PC)}$ be the set of all postal codes found in A
- b be their date of birth

1. Match each feature to each equation.

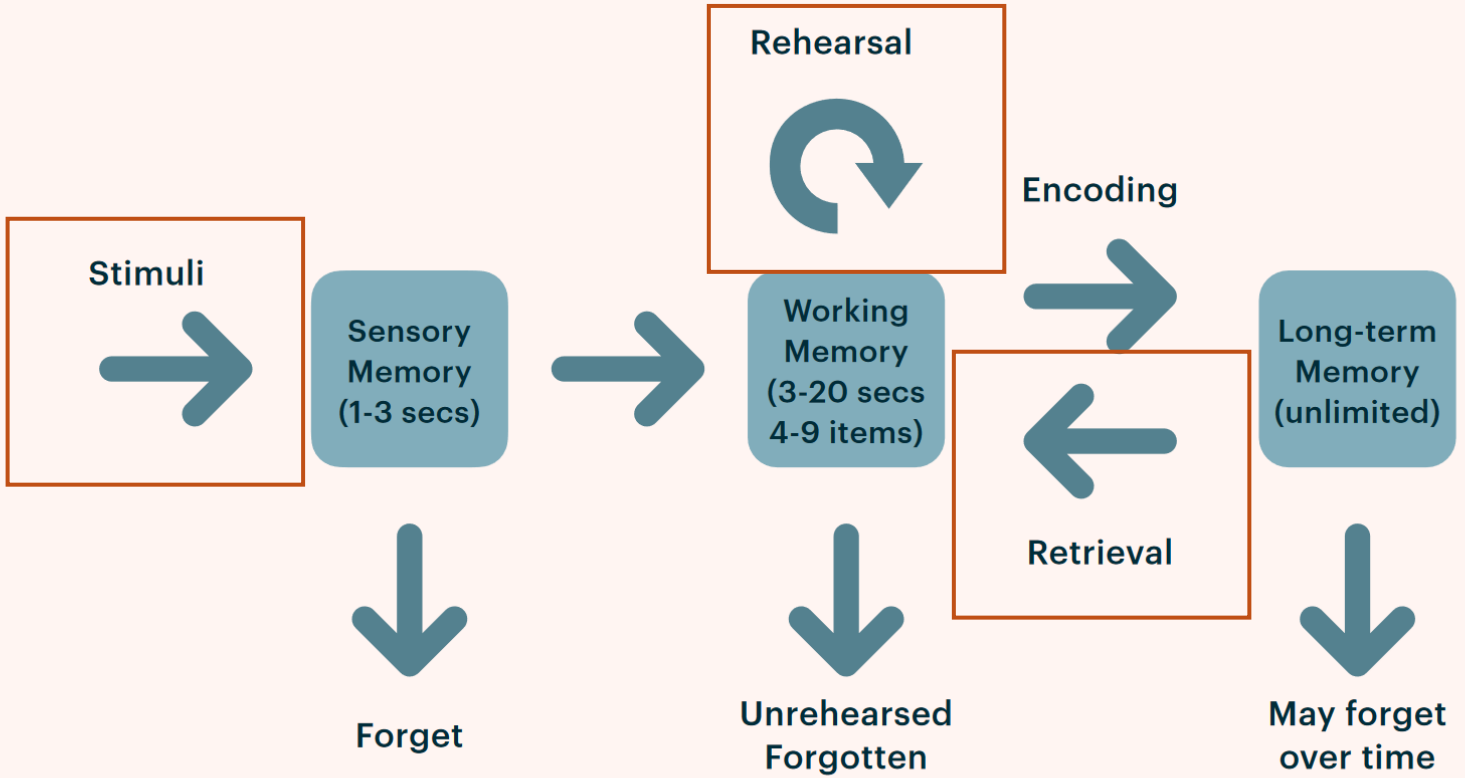
2. Determine top 5 most important features. A for top, B for 2nd, ... E for 5th

Feature	Equation
1. No. ED visits	i. N
2. No. addresses	ii. $\frac{W}{N}$
3. Time in ED system	iii. $\lvert A \rvert$
4. Average time between visits	iv. $v_N^{(date)} - v_1^{(date)}$
5. Postal code visit-to-visit consistency	v. $\frac{1}{N}\sum_{k=1}^{N}\left(v_k^{(date)} - b\right)$
6. Postal code visit-to-visit consistency, time-weighted	vi. ♀ $\to 0$, ♂ $\to 1$, else 0.5
7. Proportion of ED postal codes matching community health	vii. $\frac{1}{N-1}\sum_{k=2}^{N}\left(v_k^{(date)} - v_{k-1}^{(date)}\right)$
8. Proportion of well-formed postal codes	viii. $\frac{1}{W}\sum_{j=1}^{W}\mathbf{1}\left\{v_j^{(PC_W)} \in A^{(PC)}\right\}$
9. Proportion of postal codes matching ED facility	ix. $\frac{1}{W}\sum_{j=1}^{W}\mathbf{1}\left\{v_j^{(PC_W)} = \text{ED PC}\right\}$
10. Presence of address on *priority and non-market* list	x. $\frac{1}{W-1}\sum_{j=2}^{W}\mathbf{1}\left\{v_j^{(PC_W)} = v_{j-1}^{(PC_W)}\right\}$
11. Visit age, mean	xi. $\mathbf{1}[\exists\, a \in A : a \in \text{Priority non-market list}]$
12. Gender numeric	xii. $\frac{1}{W-1}\sum_{j=2}^{W}\frac{T-(v_j^{(date)}-v_{j-1}^{(date)})}{(W-2)T}\cdot\mathbf{1}\left\{v_j^{(PC_W)} = v_{j-1}^{(PC_W)}\right\}$, $T = v_W^{(date)} - v_1^{(date)}$

Active learning

Atkinson-Shiffrin's Multi-Store Model, image adapted by Anna Stokke

"Learning is a change in long-term memory" – Kirschner, Sweller, Clark



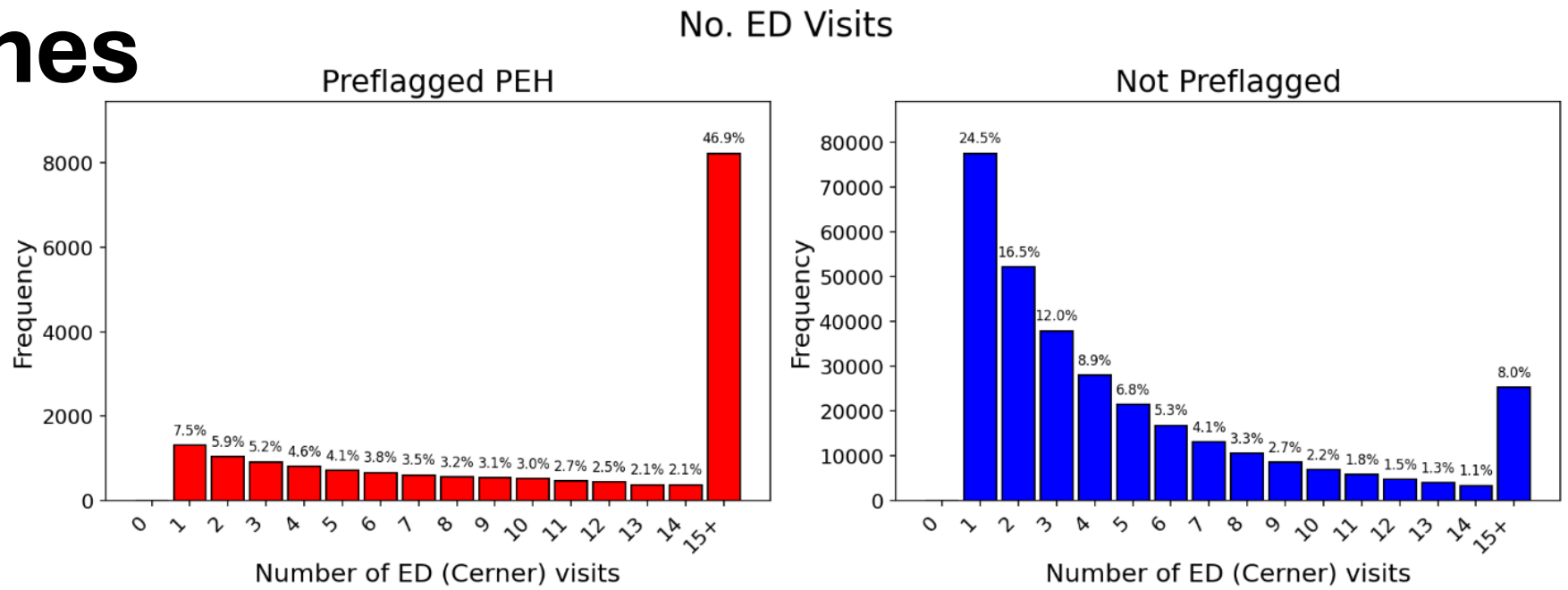
Adapted from Atkinson, R.C. and Shiffrin, R.M. (1968). "Human memory: A Proposed System and its Control Processes". In Spence, K.W. and Spence, J.T. *The psychology of learning and motivation,* (Volume 2). New York: Academic Press. pp. 89–195.

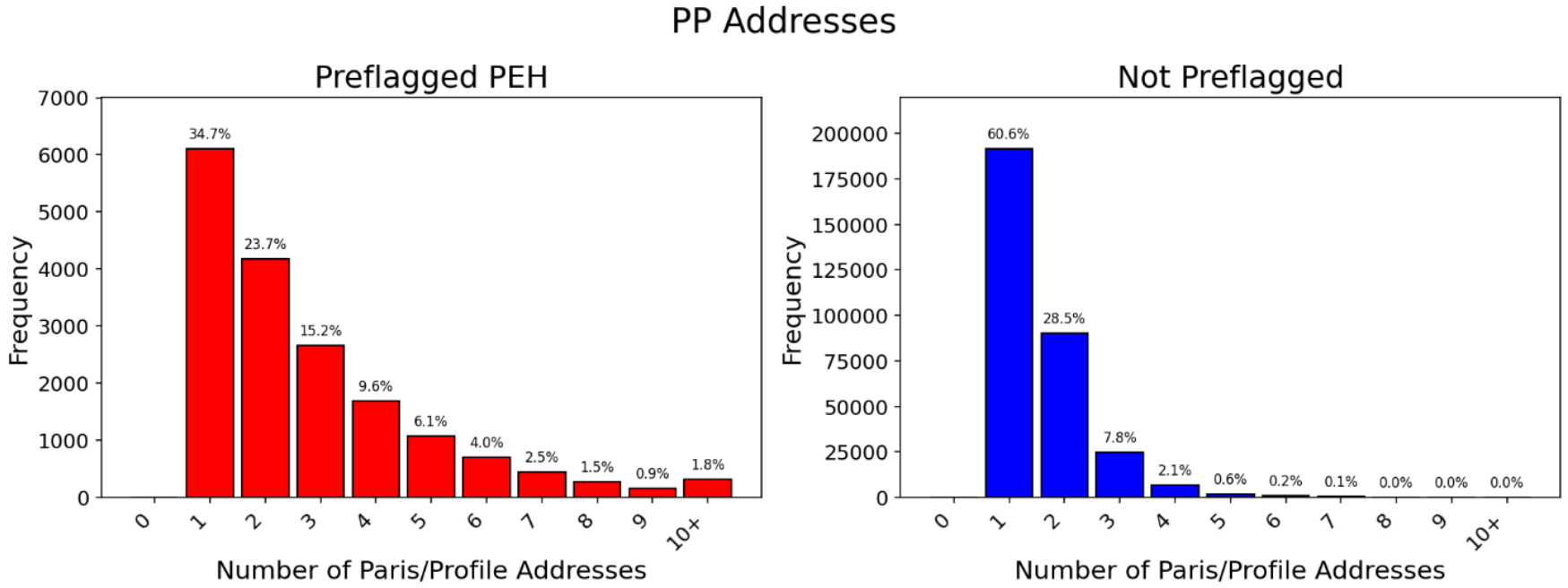
Counts and times

visits and addresses

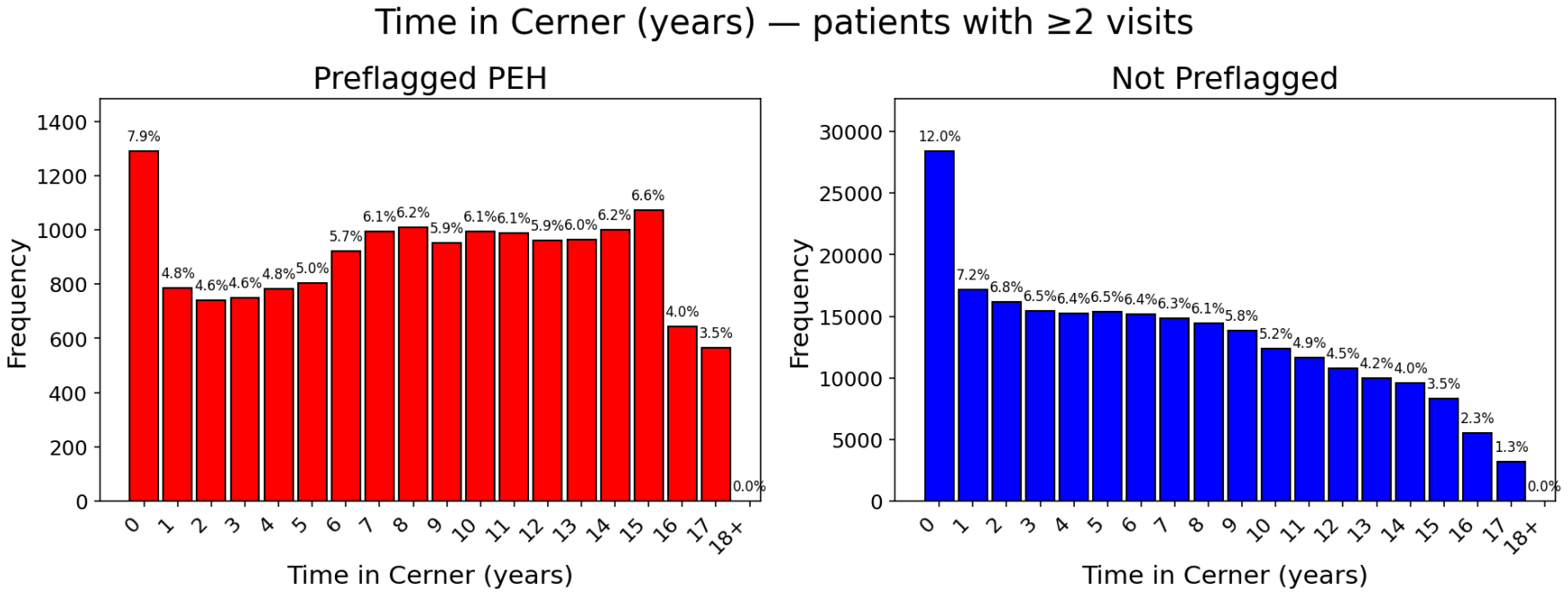
Observations:

- Preflagged cohort have more ED visits and addresses in Paris/Profile than non-preflagged cohort





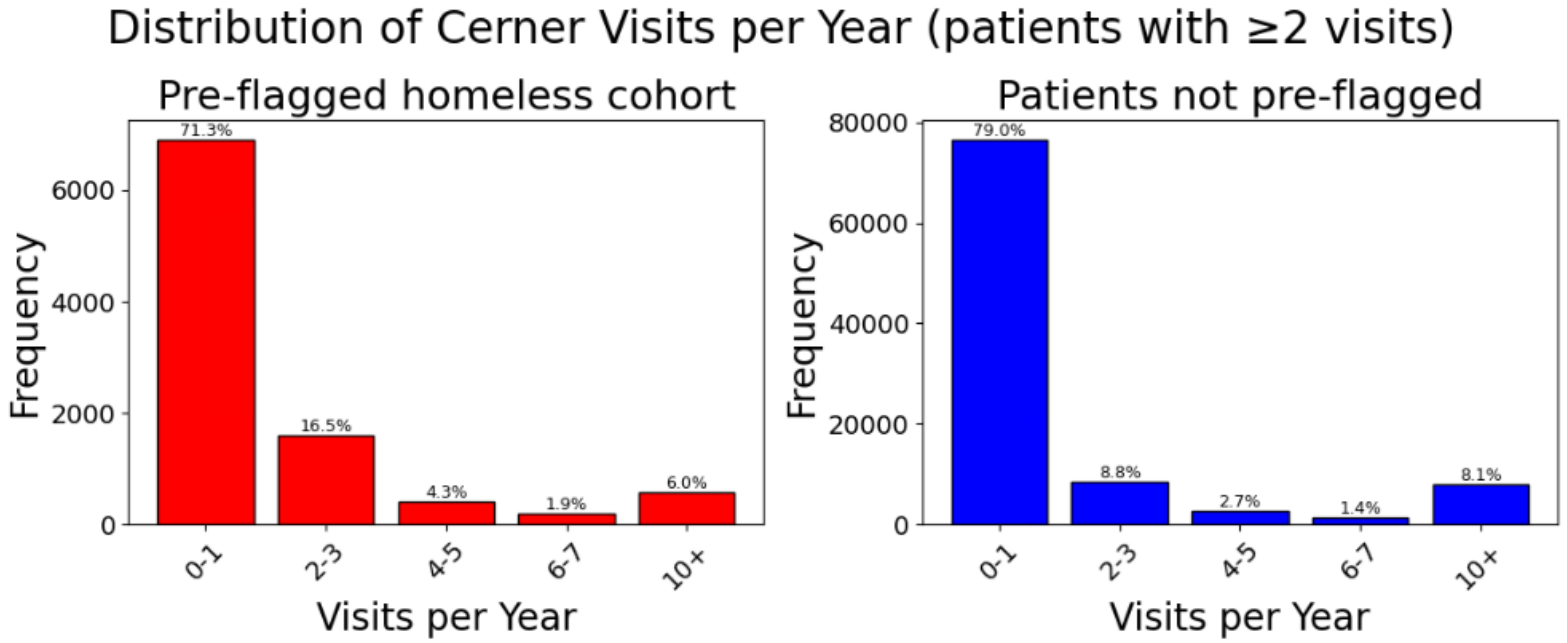
Time in system



- For patients with at least 2 visits:
  x = time in system = time from first ED visit to most recent visit
- Boundary condition: if 1 visit, $x = 0$
- Hypothesis: in preflagged, there should be more people in < 7-years buckets
- Lucie Richard (https://www.cmajopen.ca/content/11/6/E1188) commented on their homelessness definition using period of 180 days: ~"longer period of time gives higher chance patient interacts with healthcare system: short time means patient might not have interacted with healthcare system"

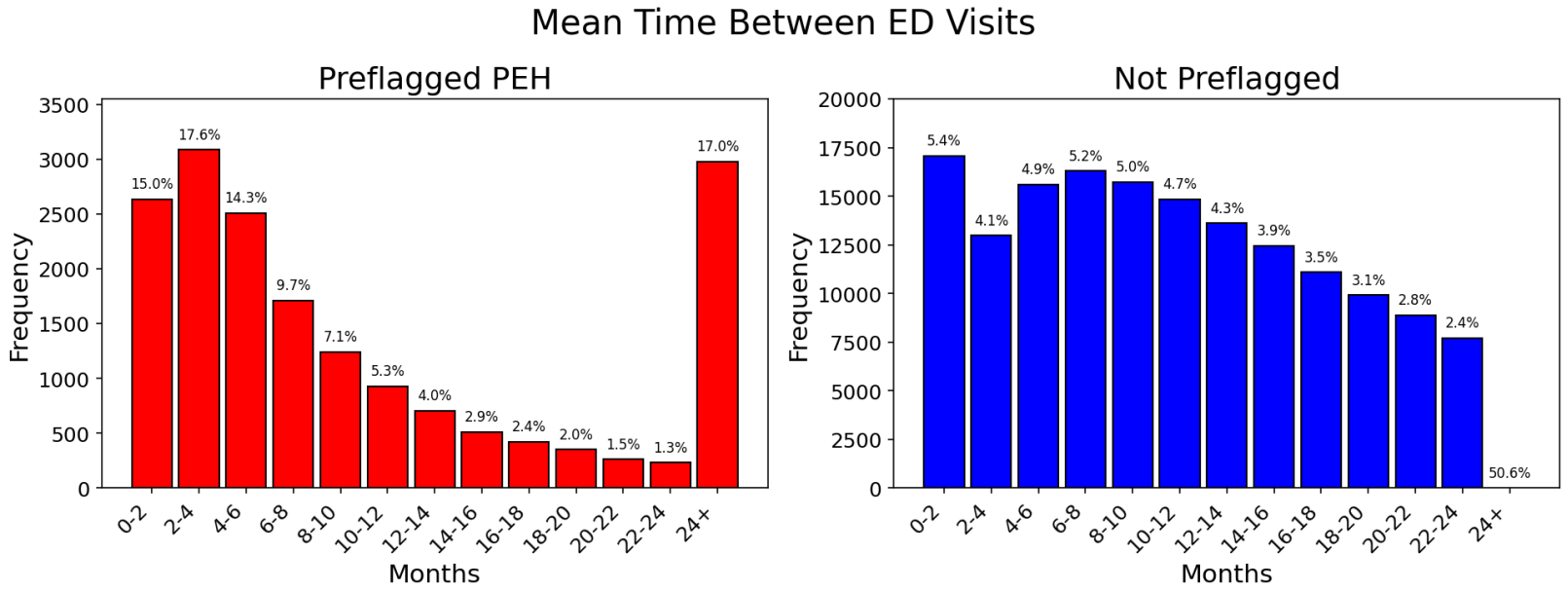
Visits / year

- This feature not predictive of homelessness



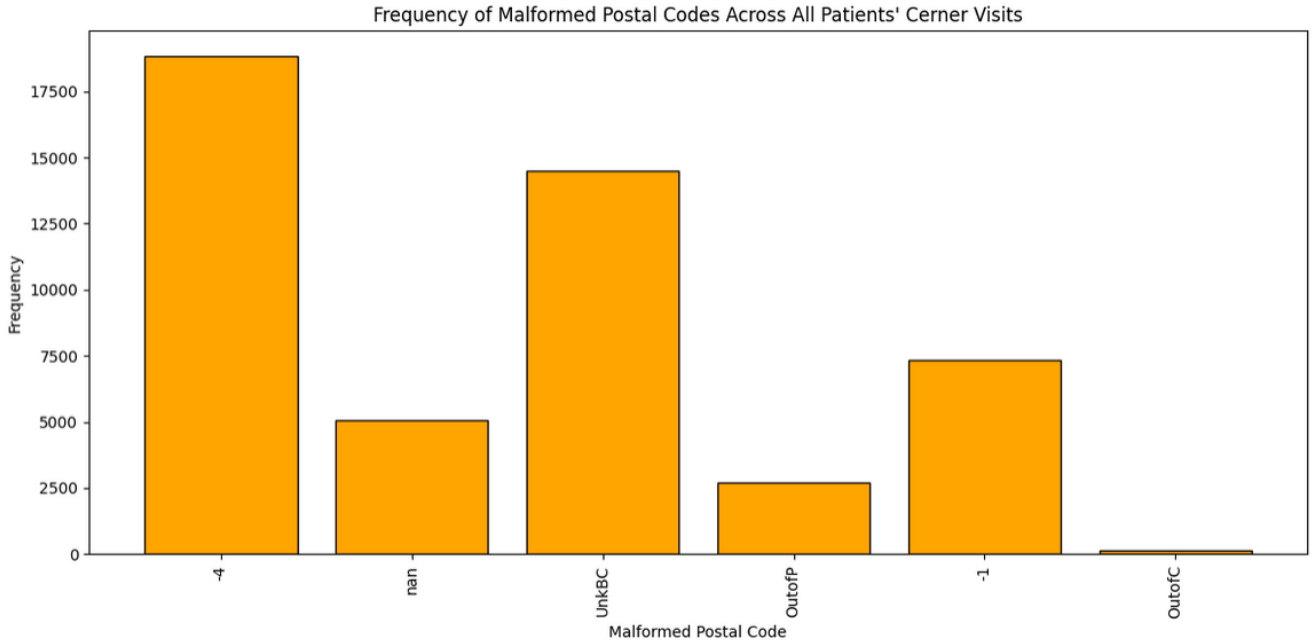
Mean time between visits = *x*

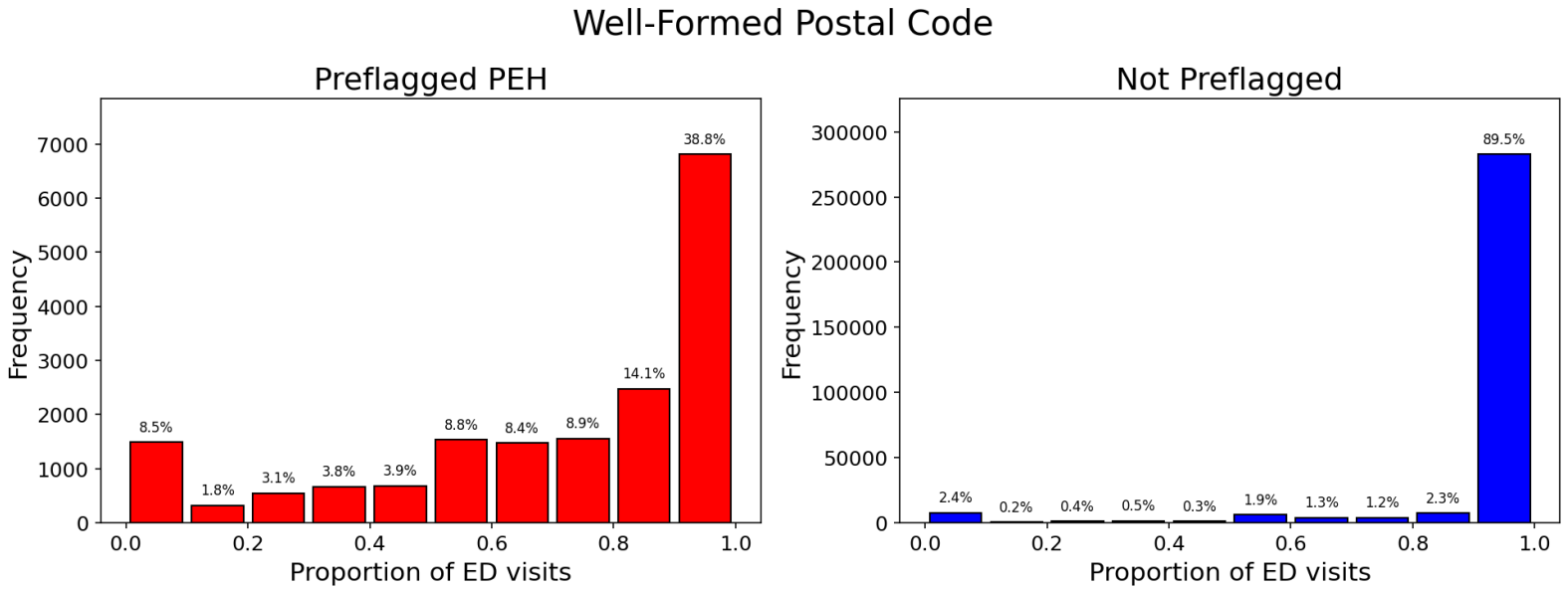
- Boundary condition: if 1 visit, $x = \infty \approx 18$ years



ED postal code proportions

Well-formed postal codes: V#L-#L#





- -4: invalid
- -1: not provided
- nan: not a number

Any thoughts how to handle various malformed codes?

Visit-to-visit consistency

- Proportion of well-formed postal codes during a subsequent visit that is consistent with previous visit.
- Eg:
  - [A, A, B, C, B, B].
    Consistent: #1-2, #5-6.
    Inconsistent: #2-3, #3-4, #4-5
    Consistency: 2/5

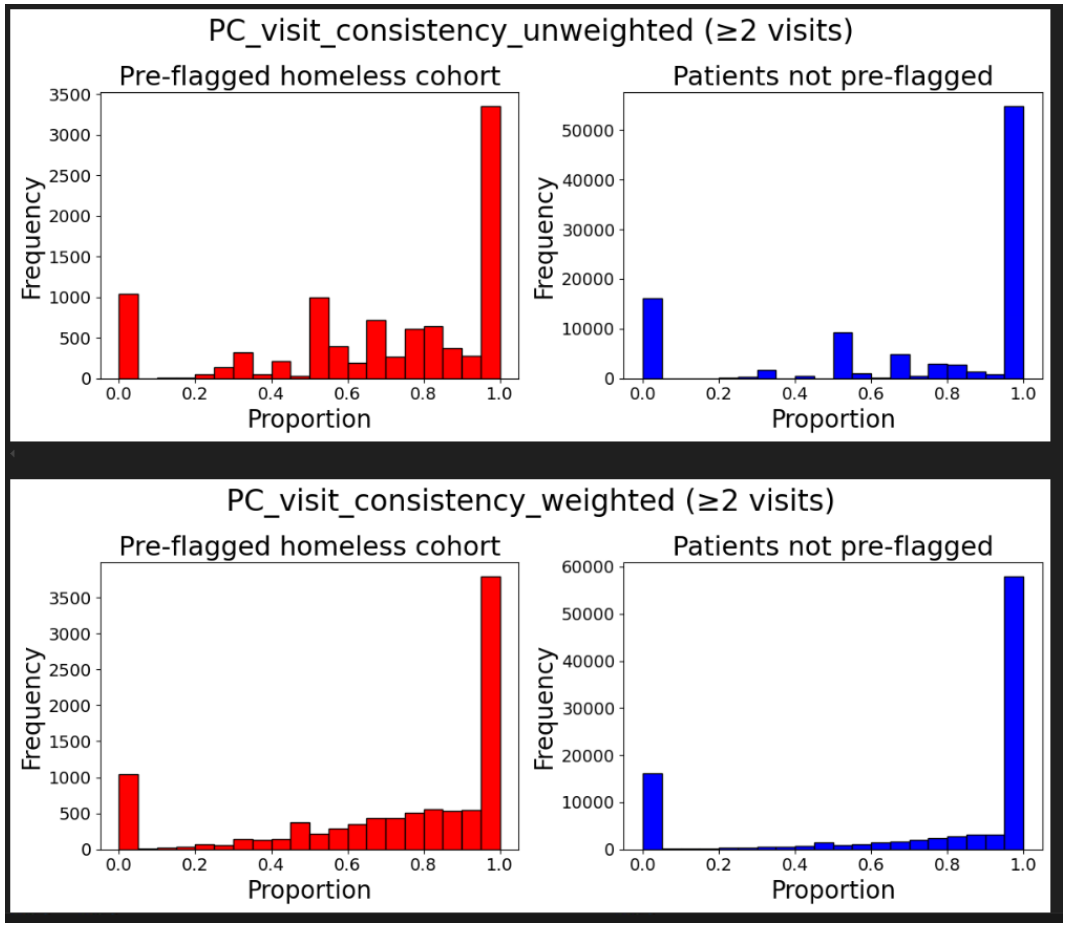
$$v_j^{PC} \text{ is postal code of } j\text{th visit}$$

$$\text{Consis(patient)} = \frac{1}{W-1}\sum_{j=2}^{W} \mathbf{1}\{v_j^{PC} = v_{j-1}^{PC}\}$$

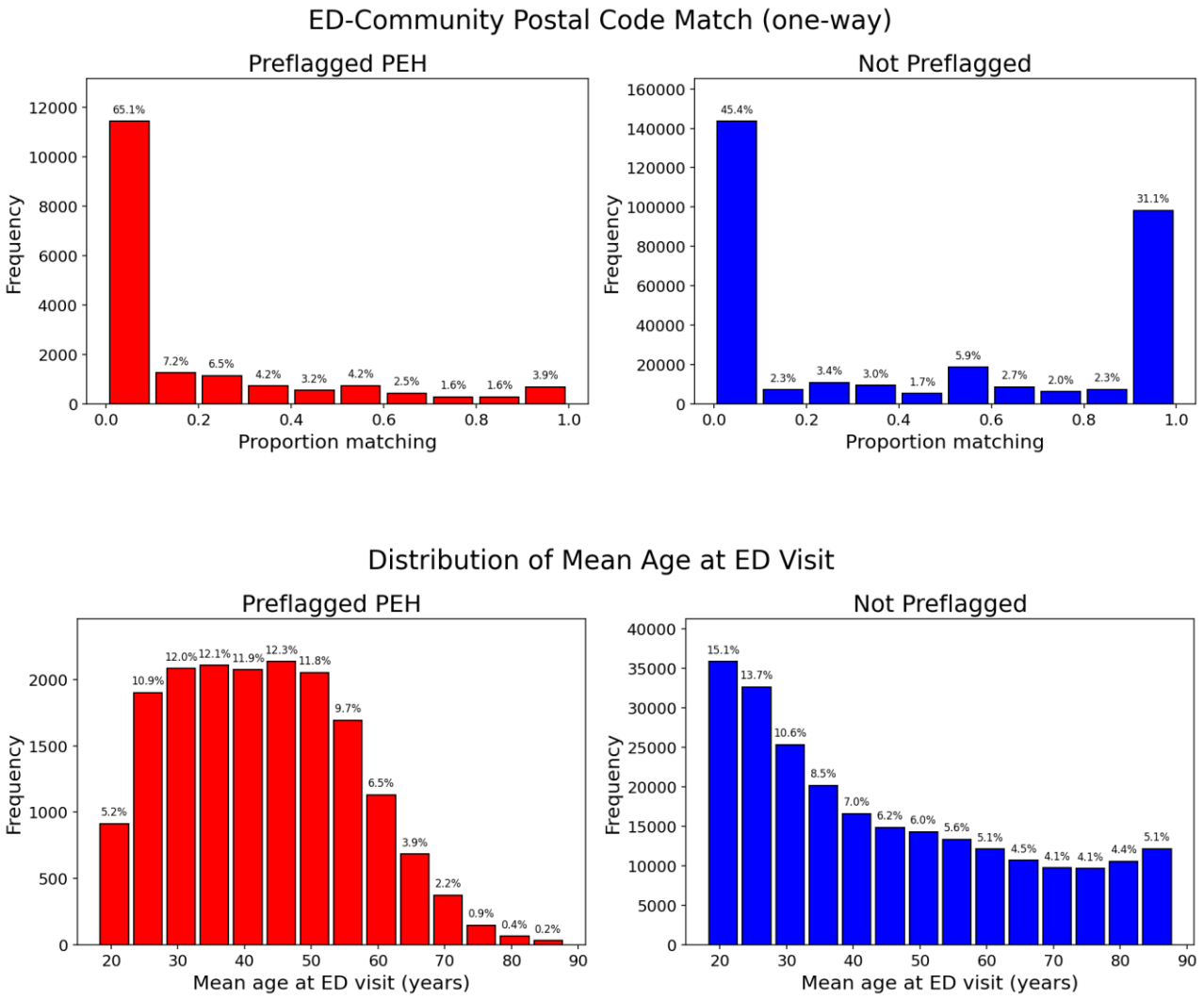
- Time-weighted: change after long time is "less unstable" than change after short time

$$T = \text{time in system}$$

$$\text{Weighted Consis(patient } i) = \frac{1}{W-1}\sum_{v=2}^{W} \frac{T-(t_v - t_{v-1})}{(W-2)\cdot T} \cdot \mathbf{1}\{v_j^{PC} = v_{j-1}^{PC}\}$$



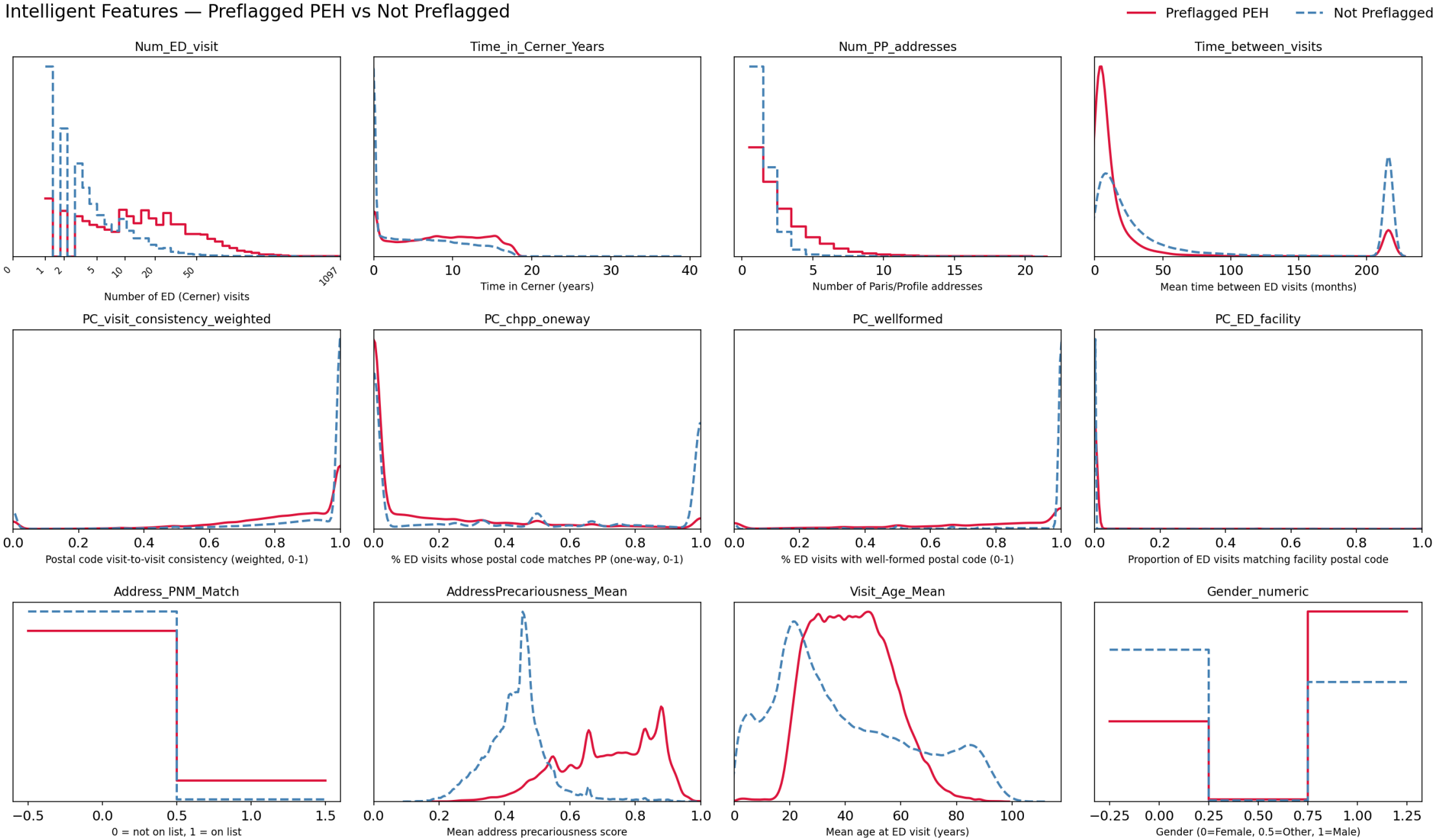
Additional intelligent features



- Consistency between postal codes from ED visits with postal codes from community health records
- Patient has an address that is on VCH's *Priority and non-market housing list*
- *Gender numeric:*

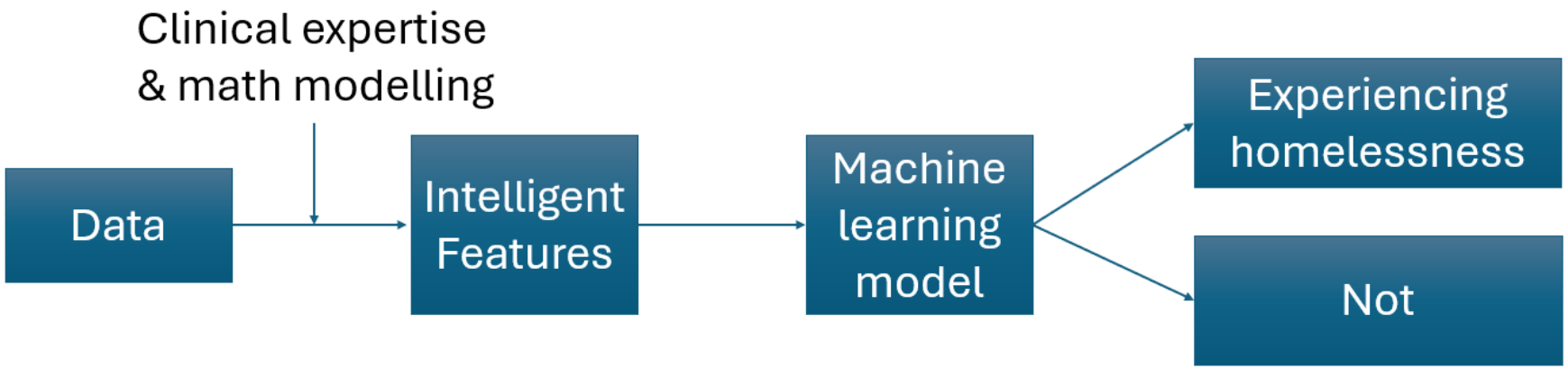
$$\text{♀} \rightarrow 0, \text{♂} \rightarrow 1, \text{else} \rightarrow 0.5$$

- Patient's mean age at ED visit


Intelligent Features — Preflagged PEH vs Not Preflagged
Preflagged PEH
Not Preflagged
Num_ED_visit
0
1
2
5
10
20
50
1097
Number of ED (Cerner) visits
Time_in_Cerner_Years
0
10
20
30
40
Time in Cerner (years)
Num_PP_addresses
0
5
10
15
20
Number of Paris/Profile addresses
Time_between_visits
0
50
100
150
200
Mean time between ED visits (months)
PC_visit_consistency_weighted
0.0
0.2
0.4
0.6
0.8
1.0
Postal code visit-to-visit consistency (weighted, 0-1)
PC_chpp_oneway
0.0
0.2
0.4
0.6
0.8
1.0
% ED visits whose postal code matches PP (one-way, 0-1)
PC_wellformed
0.0
0.2
0.4
0.6
0.8
1.0
% ED visits with well-formed postal code (0-1)
PC_ED_facility
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of ED visits matching facility postal code
Address_PNM_Match
−0.5
0.0
0.5
1.0
1.5
0 = not on list, 1 = on list
AddressPrecariousness_Mean
0.0
0.2
0.4
0.6
0.8
1.0
Mean address precariousness score
Visit_Age_Mean
0
20
40
60
80
100
Mean age at ED visit (years)
Gender_numeric
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender (0=Female, 0.5=Other, 1=Male)

Training machine learning (ML) models

- Supervised ML model
- Semi-supervised ML model
- Validation / results / interpretation

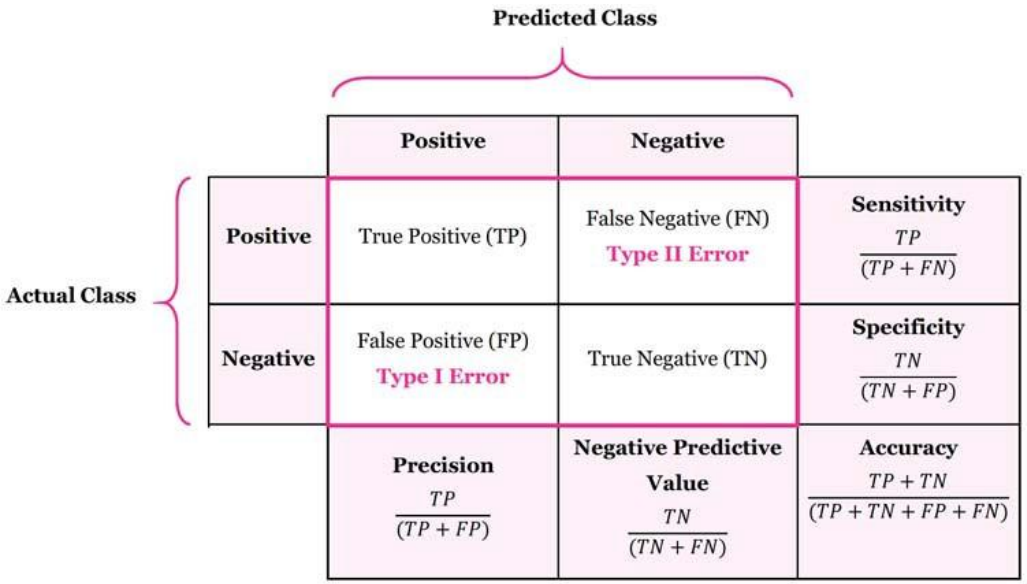


Supervised ML problem steps

- **Patient subset:** at least 1 address, 1 ED visit, demographics: **333,548 patients**
- Use data to **compute** patient's intelligent features. **Normalize** with log transform.
- **Train** ML models: Logistic Regression, Naïve Baye's, Random Forest
  - 90% of data used for train/test (5-fold cross-validation); 10% for validation.
  - **Labels: 19,690 preflagged PEH** (~6%)
    Supervised: assume non-preflagged have never experienced homelessness.
  - **Positive-unlabelled:** preflagged are true, but non-preflagged are assumed unknown

Metric	LR	NB	RF	PU LR	PU RF
Sensitivity TP/TP+FN	**0.828**	**0.516**	**0.799**	**0.876**	**0.867**
Sensitivity test st. dev.	**0.006**	**0.014**	**0.004**	**0.005**	**0.006**
Sensitivity validation	**0.821**	**0.505**	**0.810**	**0.871**	**0.862**
Specificity TN/TN+FP	0.859	0.967	0.927	0.797	0.880
Precision TP/TP+FP	0.246	0.465	0.379	0.194	0.287
F1 Score	0.379	0.489	0.514	0.317	0.431
AUROC	0.919	0.911	0.944	0.910	0.938

- **Results:** average across cross-validation



Data normalization

Normalize the data. There are two types of intelligent features: *counts / time* and *proportions / others*.

- Counts / time: **log transform.** we add 1, then take the logarithm (adding 1 ensures our features with“0” don’t become negative infinity when applying the logarithm); then we divide by the largest value so features are in the range 0 to 1.
- Scikit helper functions:
  - *Log1p*(x) = log(x+1)
  - *MinMaxScaler()*
- *Reference: https://pmc.ncbi.nlm.nih.gov/articles/PMC9036143/*
- Proportions/others: not normalized at all.

Best practice in statistics: The use of log transformation

Robert M West

Abstract
The log transformation is often used to reduce skewness of a measurement variable. If, after transformation, the distribution is symmetric, then the Welch *t*-test might be used to compare groups. If, also, the distribution becomes close to normal, then a reference interval might be determined.

Features	Iterations	Condition Number	Test Sensitivity	Test Specificity
Raw	7,172	86,100.8	77.5%	84.1%
Transformed	37	57.4	79.4%	83.4%

Semi-supervised machine learning

- Positive-Unlabeled learn allows non-preflagged patients to be PEH
- Requires as input an ML model
- Allows model *confidence* (confident threshold: 0.9 for 1, 0.1 for 0)

	Preflagged PEH = 1	**Not preflagged PEH ≈ 0**
Train 80%	Truth: 15,000 PU learn result: • 70.5% are 1 • 1.5% are 0 • 28.0% unsure	Unlabeled: 252,000 PU learn result: • 7.7% are 1 • 27.4% are 0 • 64.9% unsure
Test 20%	Truth: 4,000 PU learn result: • 69.5% are 1 • 1.7% are 0 • 28.8% unsure	Unlabeled: 63,000 PU learn result: • 7.6% are 1 • 27.7% are 0 • 64.7% unsure
Total	19,000	315,000

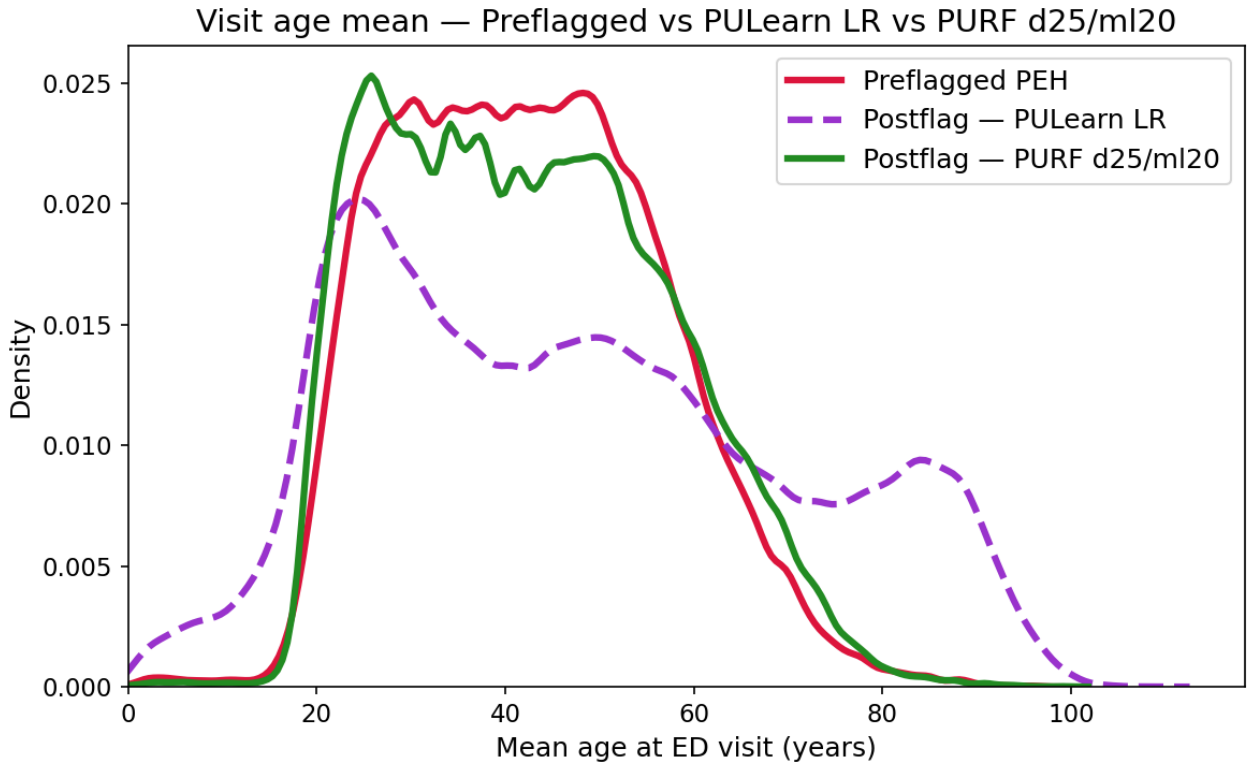
Logistic regression and feature importance

- Magnitude of logistic regression coefficients yield feature importance
- + => predictive of 1
  - => predictive of 0

Feature	Coefficient	AbsCoefficient
Num_ED_visit	7.896346	7.896346
Time_in_Cerner	-1.459783	1.459783
Num_PP_addresses	4.823305	4.823305
Time_between_visits	-1.022961	1.022961
PC_visit_consistency_weighted	-1.017433	1.017433
PC_match_PP	-1.532614	1.532614
PC_wellformed	-2.926557	2.926557
PC_ED_facility	5.049893	5.049893
Address_PNM_Match	2.063569	2.063569
Gender_numeric	0.977114	0.977114

- Pro: highly interpretable, simple
- Con: misses feature interactions

Eg: classifying sick elderly patients as PEH

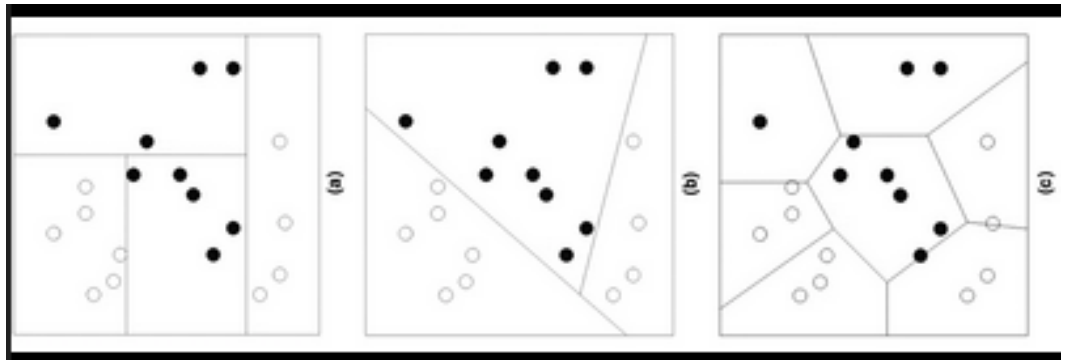


Solution: use Random Forest.

The Random Subspace Method for Constructing Decision Forests

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 8, AUGUST 1998

Tin Kam Ho



- We tune parameters:
  - Depth, max number of splits
  - Minimum sample leaves

One strategy: select parameters to optimize favourite metric[s].

- Accuracy on preflag (sensitivity)
- Small postflag cohort, hopefully high accuracy

— Depth = 10 —

Model	Sens_tst	Spec_tst	ΔSens	PF_all	Sens_val	LTC_post	LTC_post_age
PURF_d10_ml10	88.8%	85.8%	1.9%	60512	88.2%	943	0.879492
PURF_d10_ml20	88.5%	86.0%	1.8%	59949	87.6%	1041	0.884476
PURF_d10_ml30	88.6%	85.7%	1.8%	60931	88.0%	1136	0.889904
PURF_d10_ml50	89.5%	84.6%	1.7%	64603	88.5%	1232	0.892394
PURF_d10_ml100	89.0%	84.2%	1.5%	65733	88.0%	1566	0.904650

— Min_samples_leaf = 10 —

Model	Sens_tst	Spec_tst	ΔSens	PF_all	Sens_val	LTC_post	LTC_post_age
PURF_d10_ml10	88.8%	85.8%	1.9%	60512	88.2%	943	0.879492
PURF_d15_ml10	86.6%	87.9%	6.7%	53571	85.5%	783	0.873052
PURF_d20_ml10	84.3%	89.5%	10.1%	48394	84.2%	759	0.872879
PURF_d25_ml10	84.1%	90.0%	10.6%	46446	83.6%	768	0.873805
PURF_d30_ml10	83.6%	90.1%	10.9%	46244	83.5%	759	0.872928
PURF_d35_ml10	83.9%	90.2%	10.7%	46303	83.8%	758	0.873346
PURF_d40_ml10	83.7%	90.0%	11.0%	46686	83.7%	754	0.873886

— Min_samples_leaf = 20 —

Model	Sens_tst	Spec_tst	ΔSens	PF_all	Sens_val	LTC_post	LTC_post_age
PURF_d10_ml20	88.5%	86.0%	1.8%	59949	87.6%	1041	0.884476
PURF_d15_ml20	87.9%	86.9%	4.6%	57135	87.3%	935	0.881272
PURF_d20_ml20	87.0%	87.8%	6.3%	54298	86.1%	932	0.882217
PURF_d25_ml20	86.5%	88.2%	6.4%	52807	86.2%	925	0.881711
PURF_d30_ml20	86.5%	88.0%	6.7%	53535	86.3%	918	0.880967
PURF_d35_ml20	86.6%	88.1%	6.5%	53219	86.0%	941	0.882388
PURF_d40_ml20	86.3%	88.3%	6.7%	52778	85.7%	921	0.881611

One strategy: select parameters to optimize favourite metric[s].

- Accuracy on preflag **sensitivity**
- Small postflag cohort **PF_all**, hopefully high accuracy

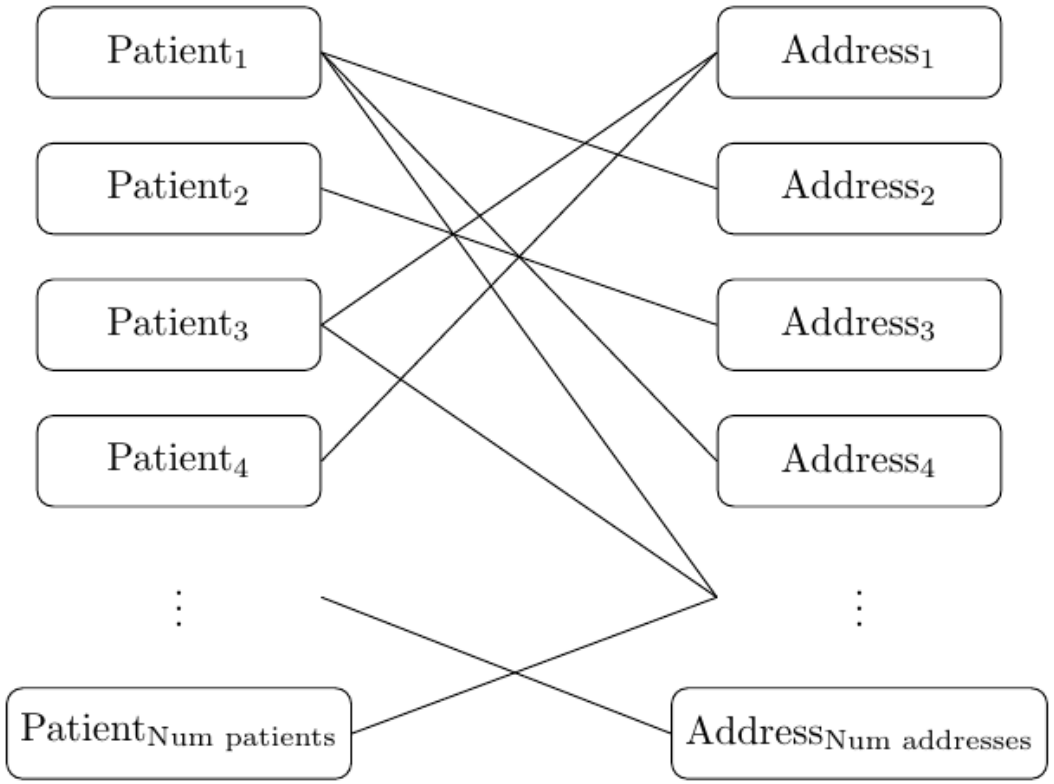
Small postflag cohort:

Depth 10: min leaves 20

Min leaves 10: depth 30

Min leaves 20: depth 25

Address analysis and address features



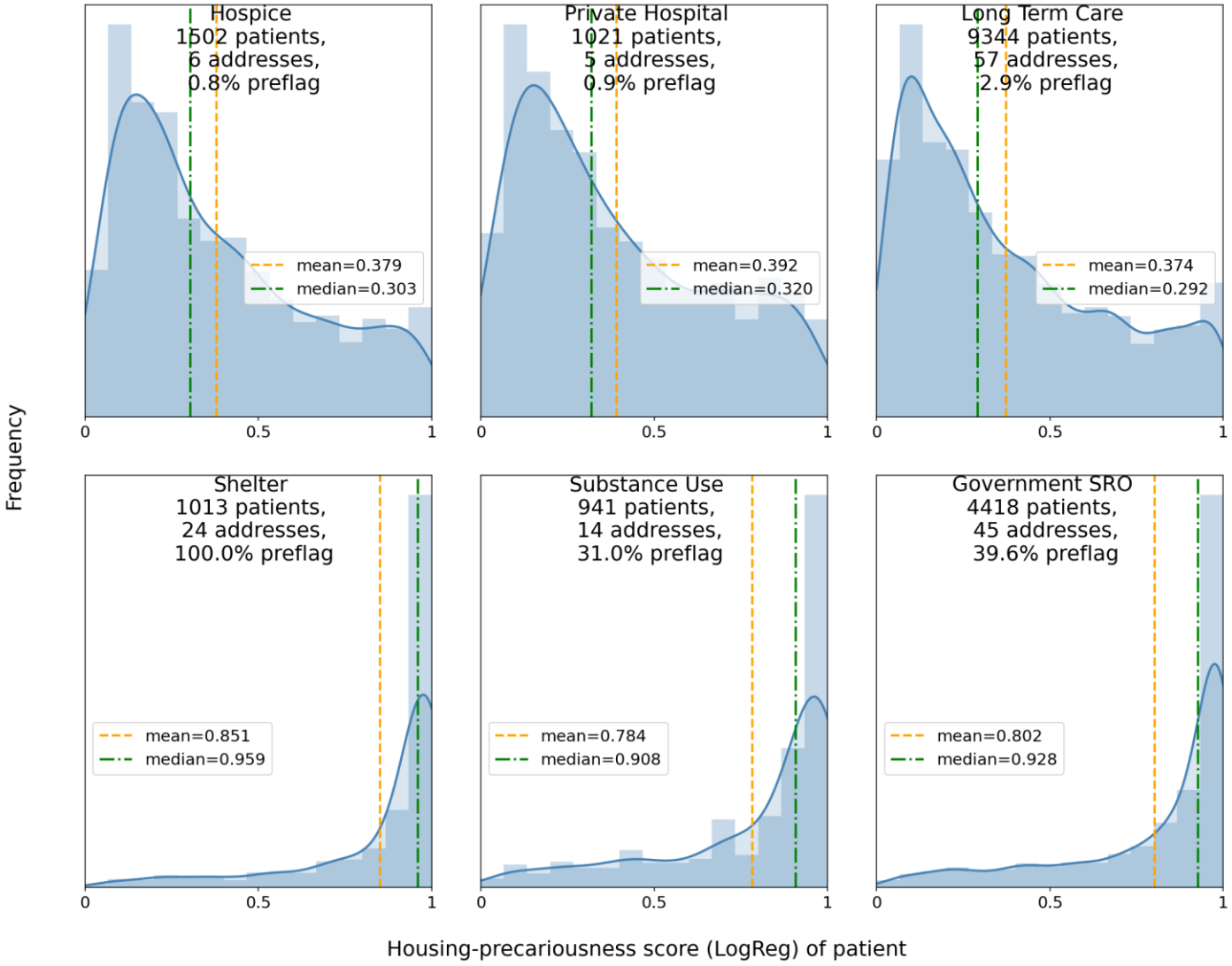
Simple address features:

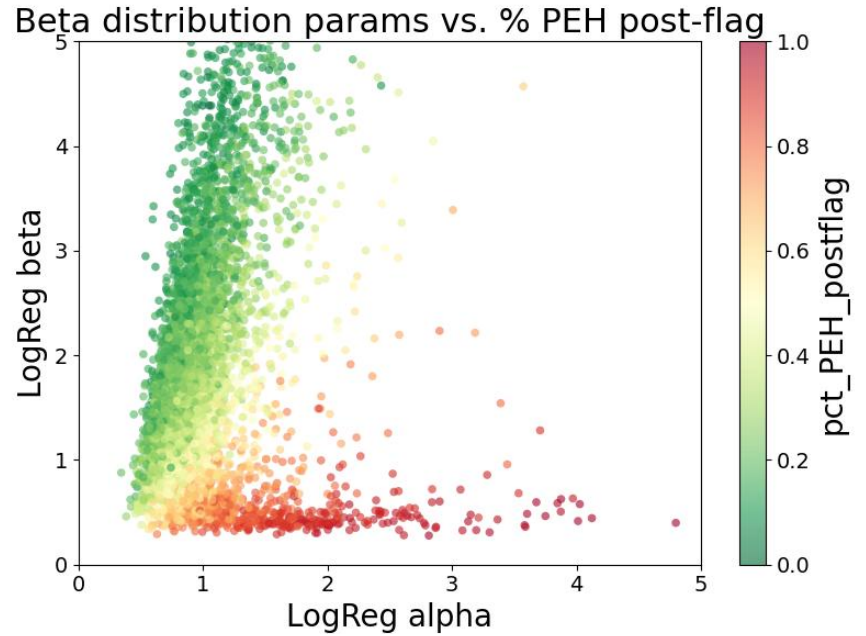
- No. patients with this address
- No. and % of preflagged patients
- No and % of postflagged patients

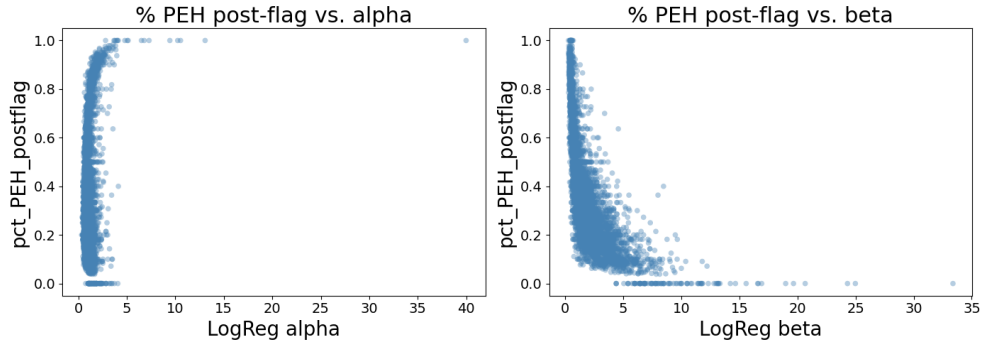
More complex:

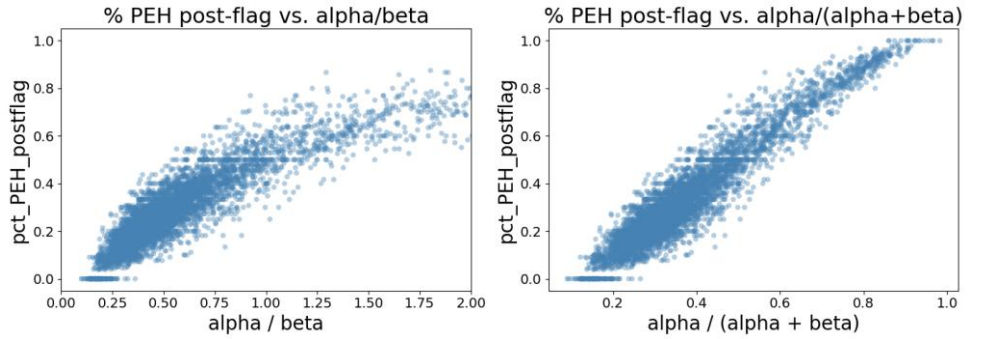
- Score addresses based on it's resident's ML scores
- If few residents, smooth by padding with 0.5

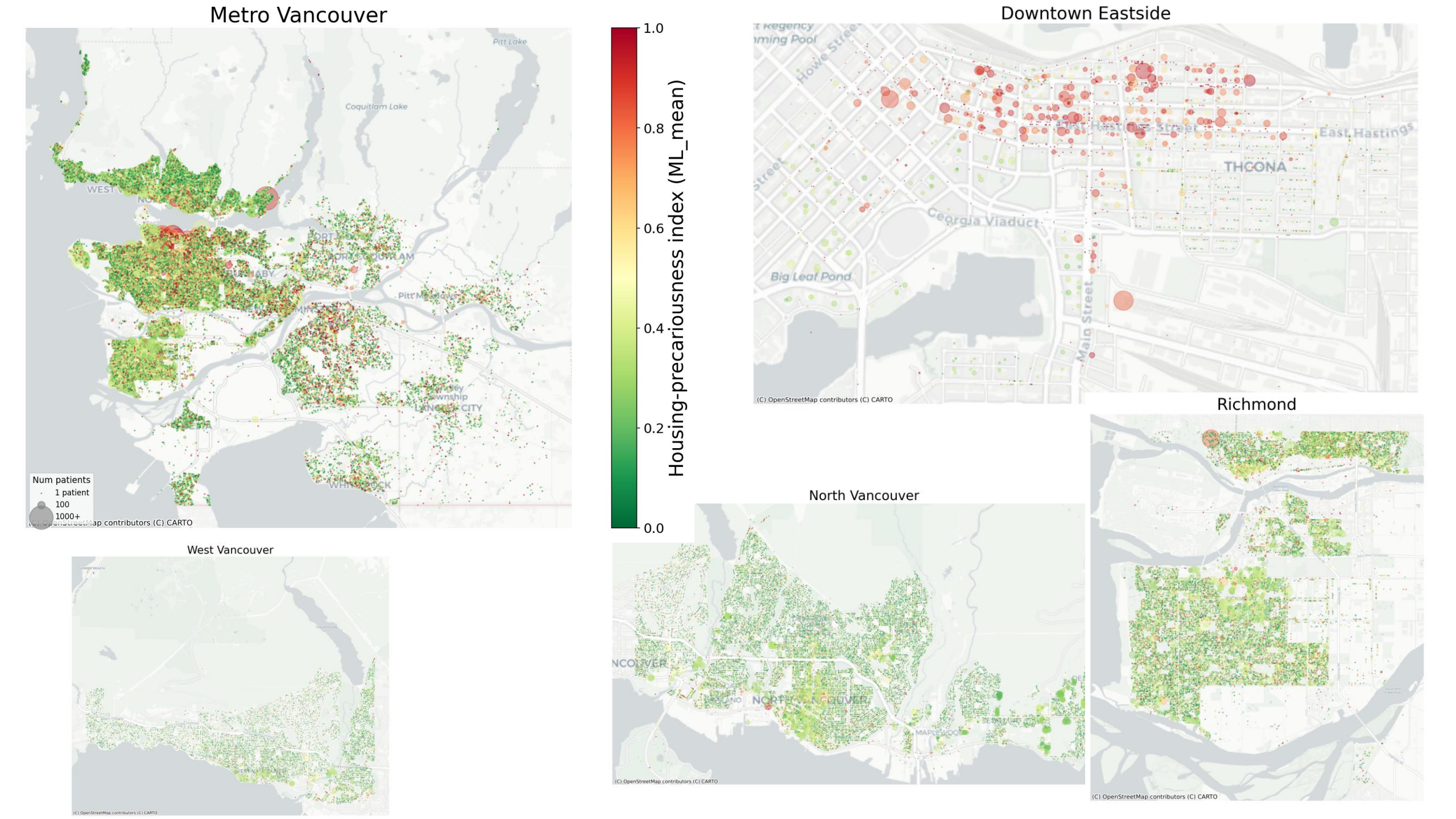
Address housing-precariousness score: from a distribution to a single number



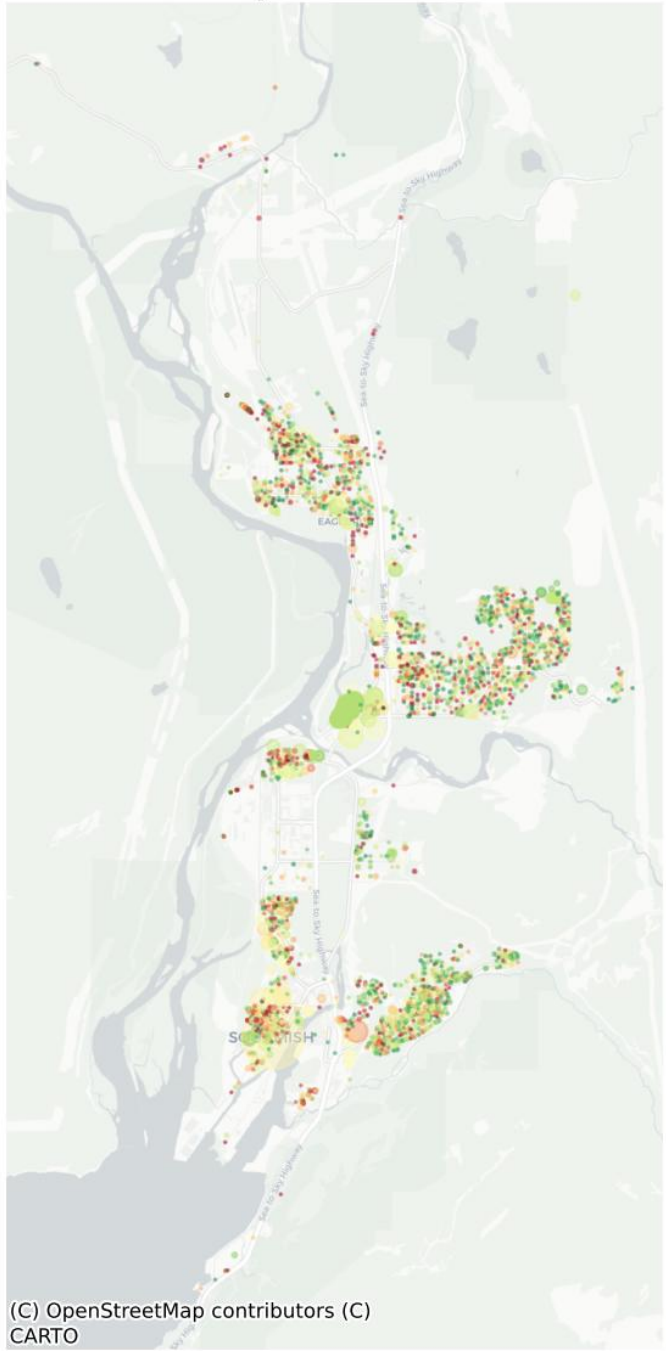




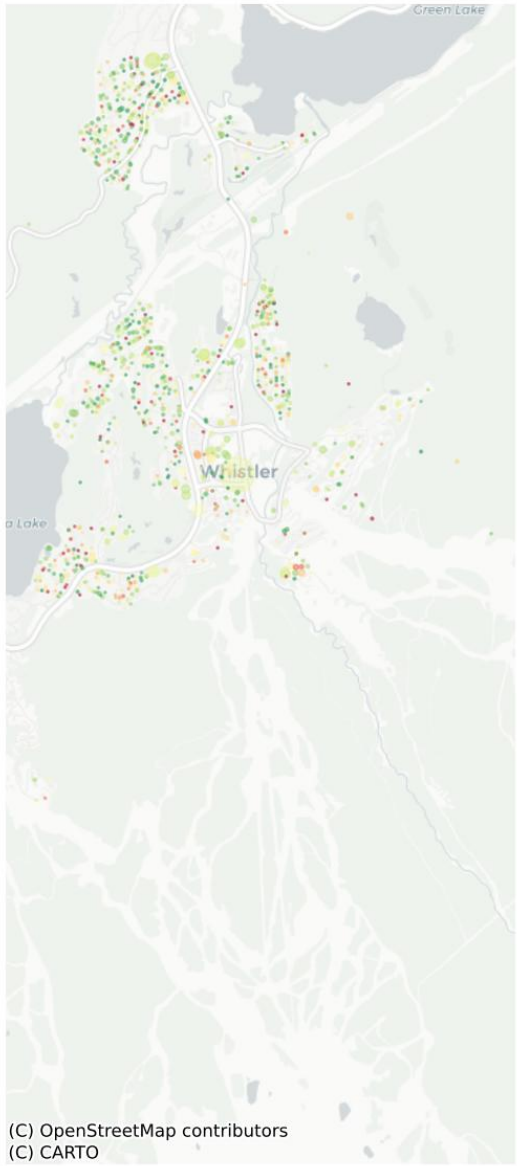



Metro Vancouver
Downtown Eastside
Richmond
North Vancouver
West Vancouver
Housing-precariousness index (ML_mean)
1.0
0.8
0.6
0.4
0.2
0.0
Num patients
1 patient
100
1000+
(C) OpenStreetMap contributors (C) CARTO
Pitt Lake
Coquitlam Lake
Howe Street
Georgia Viaduct
Big Leaf Pond
Main Street
East Hastings
THORNTON

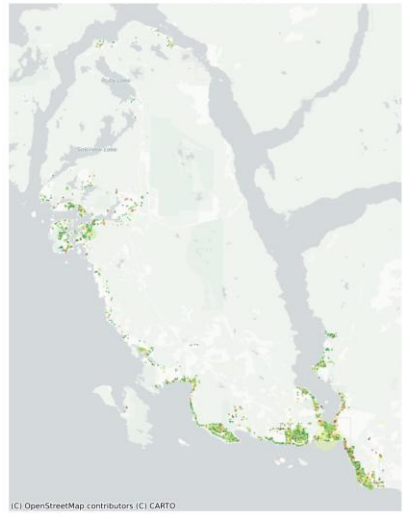
Squamish



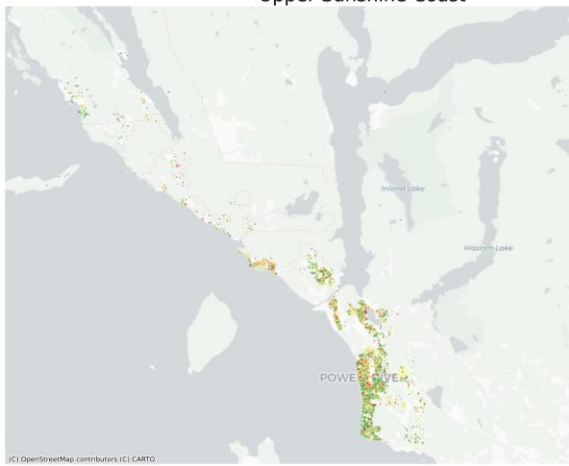
Whistler



Sunshine Coast

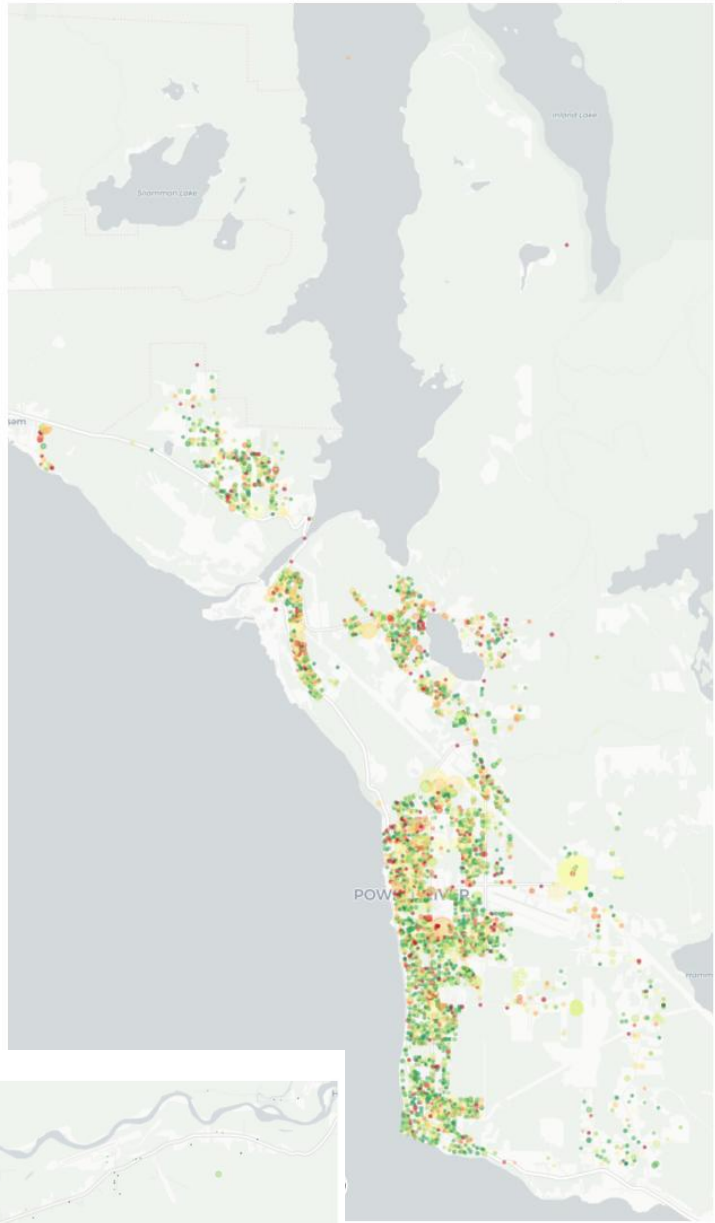


Upper Sunshine Coast

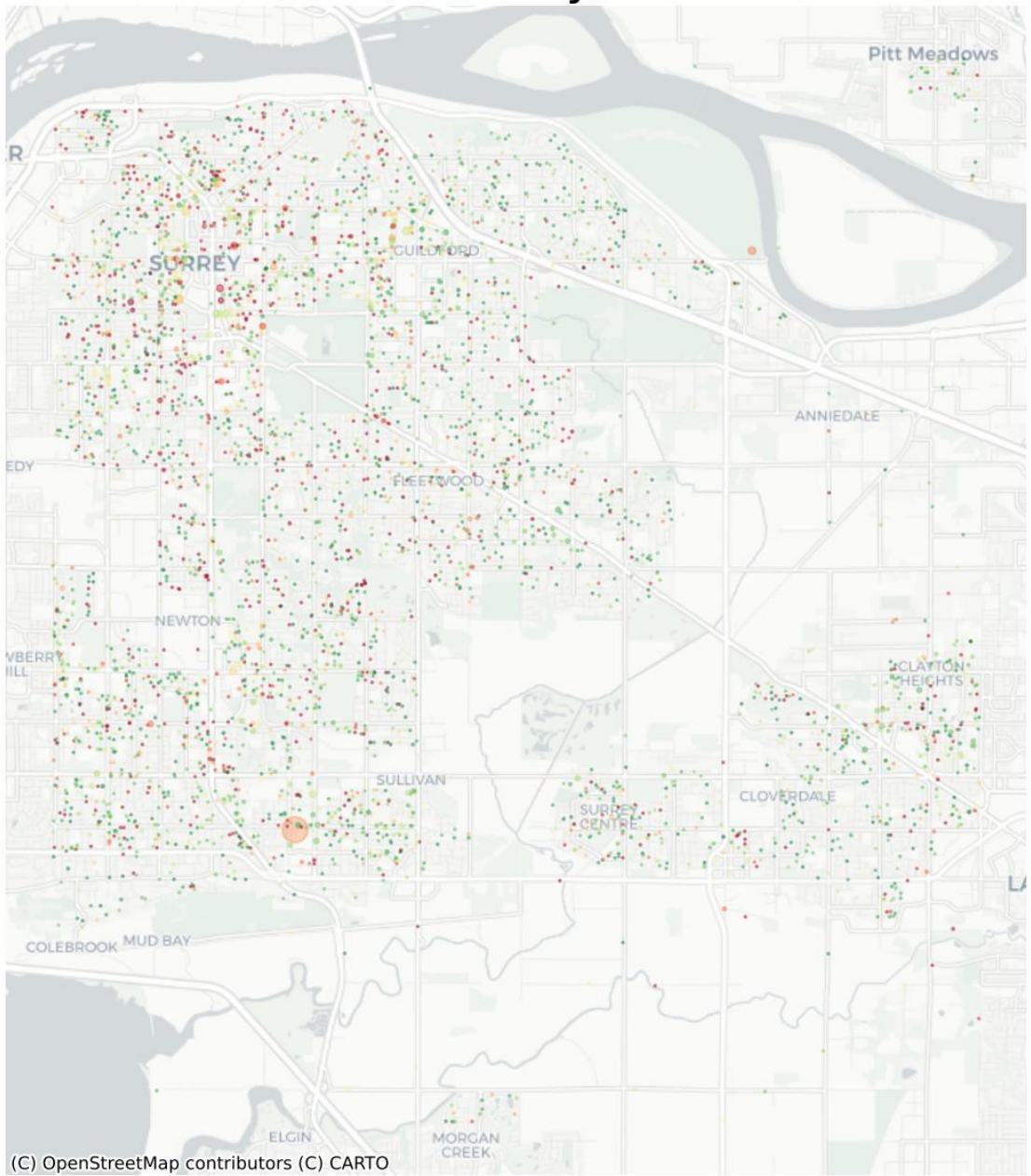


Bella Coola

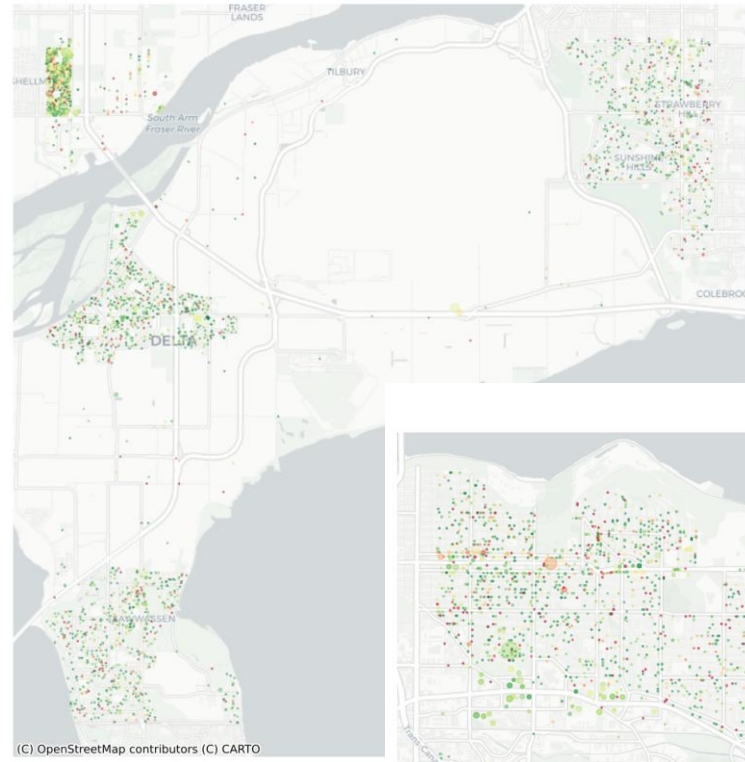
Qathet (Powell River)



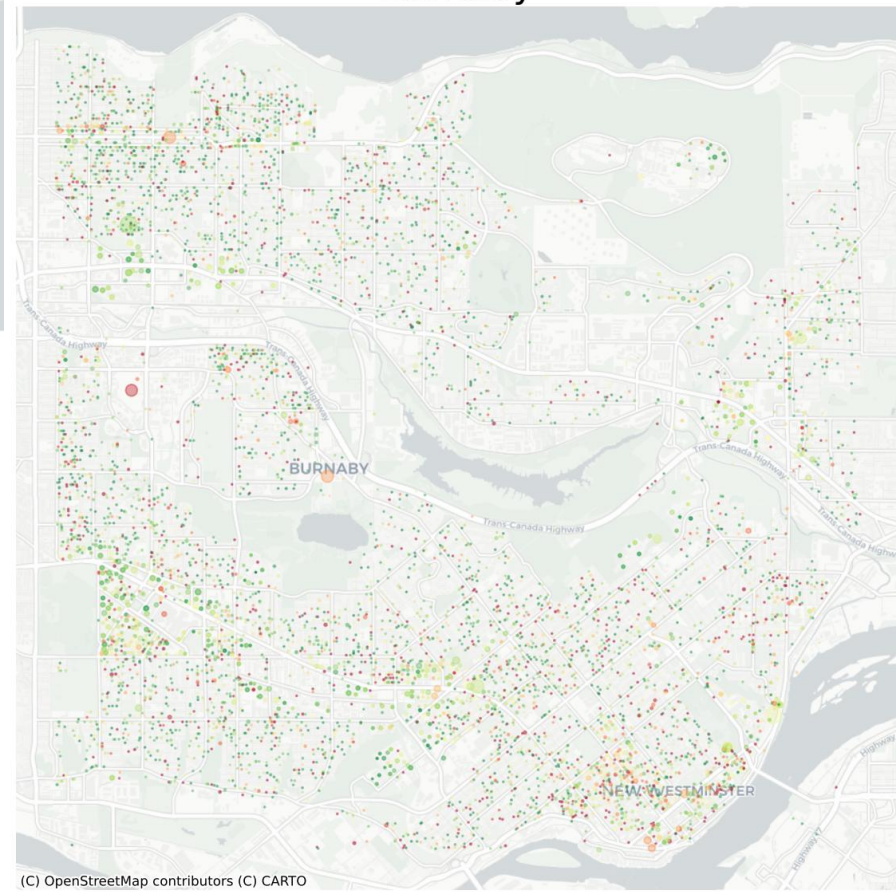
Surrey



Delta



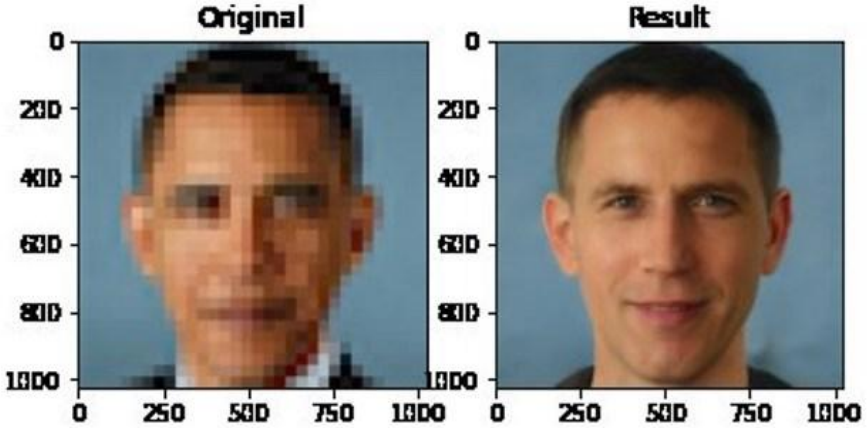
Burnaby

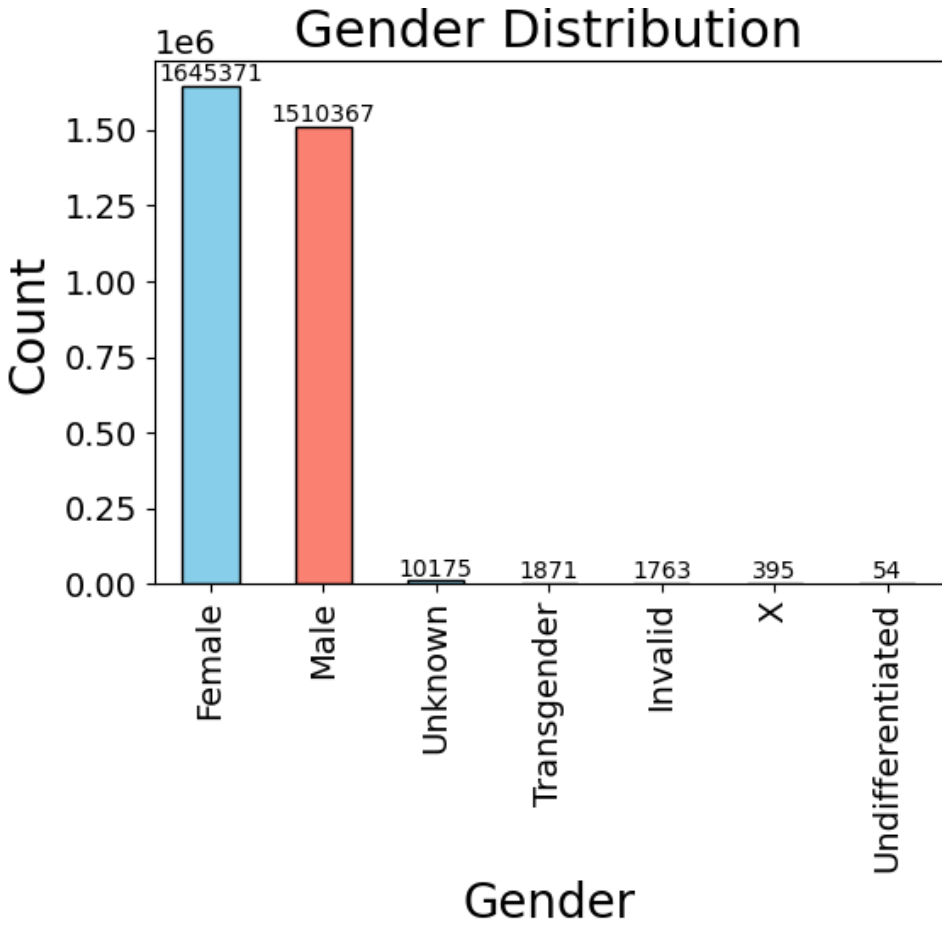


Gender analysis

Demographics data set:

- Female: 51.90%
- Male: 47.65%
- Other: 0.45%





	num_patients	num_preflagged	pct_preflagged
Female	181326	5937	3.27
Invalid	77	2	2.60
Male	151078	13223	8.75
Transgender	675	135	20.00
Unknown	327	20	6.12
X	72	14	19.44
ALL PATIENTS	333831	19349	5.80

Training models to reduce gender bias

Observe:

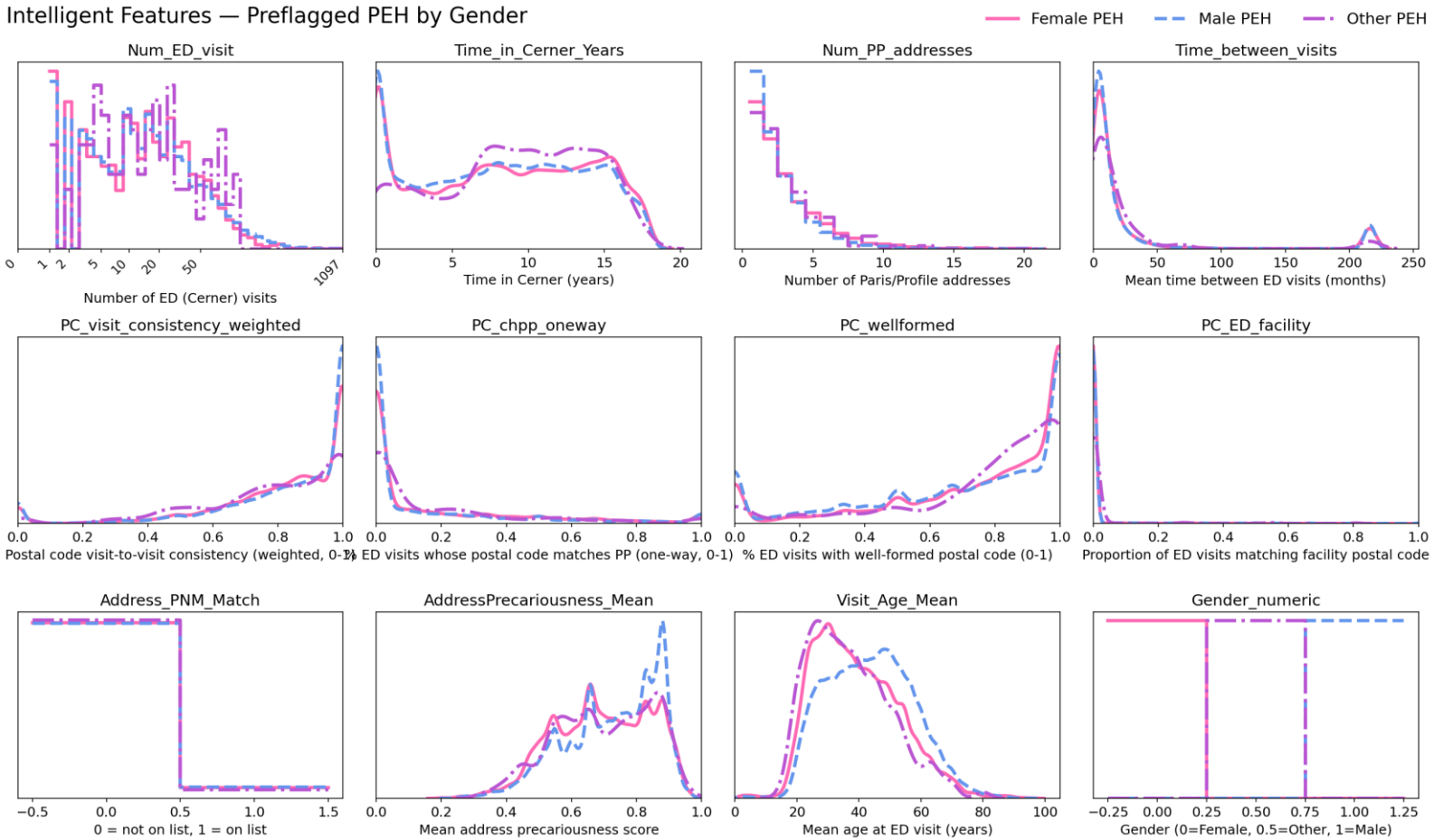
- Models are biased male PEH
  - >2x more male PEH than female PEH

Solution:

- Define *TransNB = X* or *Transgender*
- Train models on **Female**, **Male**, **TransNB**, and **everyone**
- Best: if female, use model trained only on female. Else, use model trained on everyone.

	Everyone	Female	Male	TransNB
No. preflagged	19,709	6,067	13,448	153
LR accuracy	82.2%	70.9%	87.3%	81.7%
PU	88.6%	79.5%	92.7%	88.2%
Best fully-sup. ML	LR	$LR_{♀}$	LR	RF
accuracy	82.2%	79.5%	87.3%	85.6%
Best semi-sup. ML	PU	$PU_{♀}$	PU	PU
accuracy	88.6%	86.7%	92.7%	88.2%

	num_patients	num_preflagged	pct_preflagged
Female	181326	5937	3.27
Invalid	77	2	2.60
Male	151078	13223	8.75
Transgender	675	135	20.00
Unknown	327	20	6.12
X	72	14	19.44
ALL PATIENTS	333831	19349	5.80



Logistic regression coefficients:

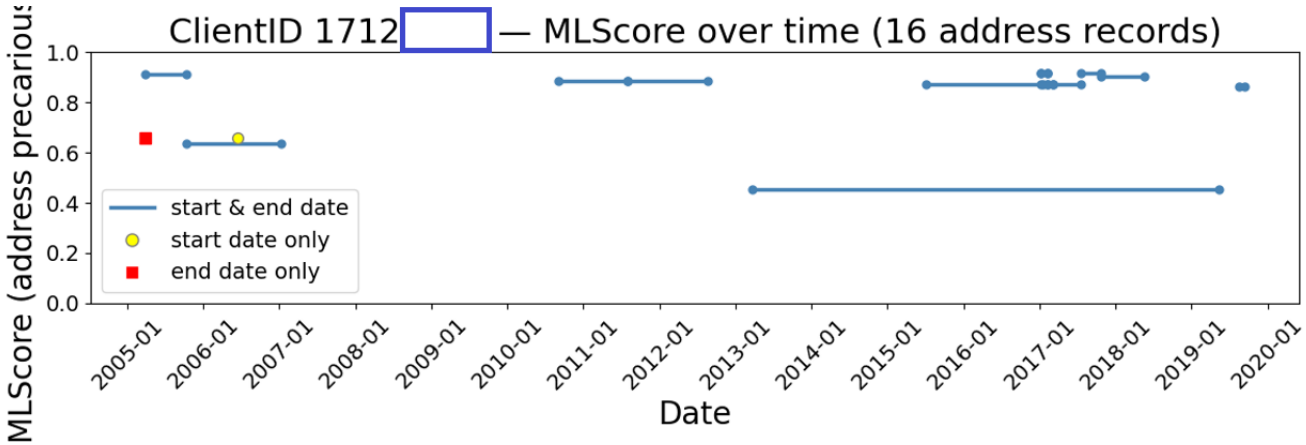
	Feature	Female	Male	Other
0	Intercept	3.0032	1.5333	1.4250
1	Num_ED_visit	8.7644	8.8391	4.4164
2	Time_in_Cerner	-0.9616	-0.8851	2.2375
3	Num_PP_addresses	3.8534	3.4815	2.6990
4	Time_between_visits	-1.5229	-1.4104	-2.4283
5	PC_visit_consistency_weighted	-1.1543	-1.1251	-0.8820
6	PC_match_PP	-1.1542	-1.4808	-1.5877
7	PC_wellformed	-3.6531	-3.7619	-3.1513
8	PC_ED_facility	5.8789	5.9009	0.9305
9	Address_PNM_Match	1.6412	1.3629	1.5770
10	Gender_numeric	0.0000	1.4461	0.0476

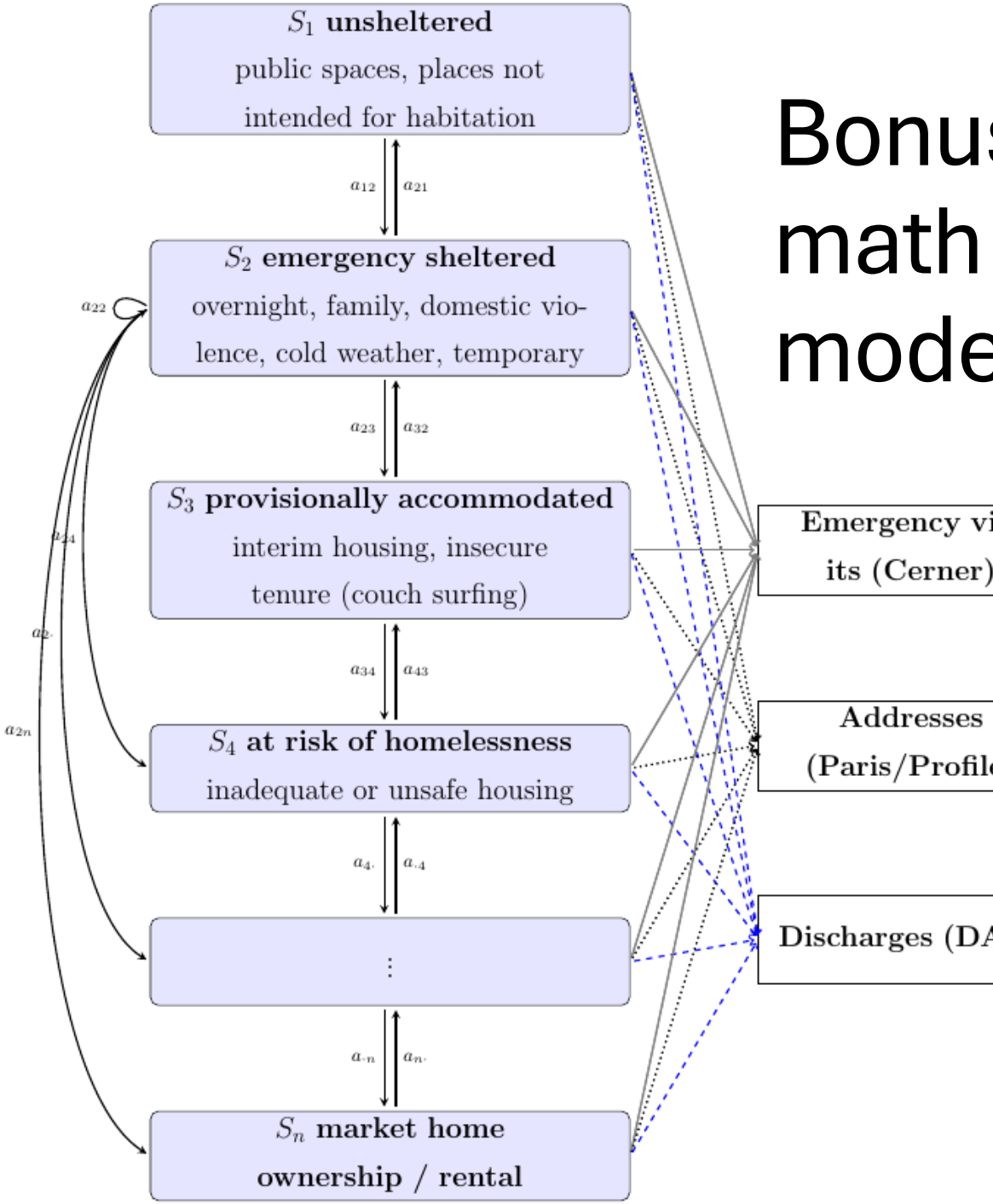
Notice:

- Male PEH often older
- Female PEH typically have more addresses

Time series

- Temporal model of homelessness: rather than predicting if *ever* experienced homelessness, predict at specific points in time





Bonus math mode

Learning health system

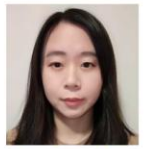
Vancouver CoastalHealth

Ceinwen Pope
Medical Health Officer, Public Health

Krisztina Vasarhelyi
Senior Learning Health Systems,Lead

Ken Hawkins
Data & Analytics, Manager

Amit Chalka
Data & Analytics, Advisor



Yumian Hu
Public Health Surv Unit, Senior Epidemiologist

Alexander Rutherford
Mathematics, Adjunct Professor

SFU SIMON FRASER UNIVERSITY

Jessica Stockdale
Mathematics, Assistant Professor

JF Williams
Mathematics, Associate Professor

Jeremy Chiu
Applied Math PhD Candidate, Langara Mathematics Instructor

Langara.
THE COLLEGE OF HIGHER LEARNING.

I believe that healthcare systems can benefit from people with strong quantitative background

Acknowledgements

- Presentation takes place on the traditional and unceded territories of xʷməθkʷəy̓əm (Musqueam), Sḵwx̱wú7mesh (Squamish) and səlilwətaɬ (Tsleil-Waututh) Nations
- Team received funding from SFU and the BC SUPPORT Unit. I received funding from SFU, Langara Faculty Association, and Langara Applied Research Centre
- Langara Applied Research Centre collaborators: Albert Wong and Nay Zaw Lin
- Various VCH teams:
  - Health Information Management – Coding and Informatics Services: Jason Xie, Scot Kwong, and Aparna Deshpande
  - Extended research team: Meagan Coman, Erin Isnor, Mark Lysyshyn, Althea Hayden, and Saffrin Granby
  - Feedback and suggestions from Rohit Vijh and Public Health Surveillance Unit
  - Communicable Diseases team Audrey Fengler, Jas Sajan, and Jacey Larochelle
  - Clinical Informatics Specialist Christie Hamilton
  - Health Research BC SUPPORT Unit: Larry Mraz and John Ward
- And to you for listening, and for keeping calculus 1 transferable between SFU, Langara, and other universities/colleges across BC