How to take a sample of Canadians; and how we're doing it in the Canadian Longitudinal Study on Aging

Lauren Griffith
Harry Shannon
McMaster University

CEB Rounds
15 February 2012

Outline of presentation

- Background on sampling
- Participants in the CLSA
- Sampling approaches in the CLSA
- CCHS participants
- Sampling from provincial health registries
- Principles of Random Digit Dialing
- Issues with RDD
- Conclusion

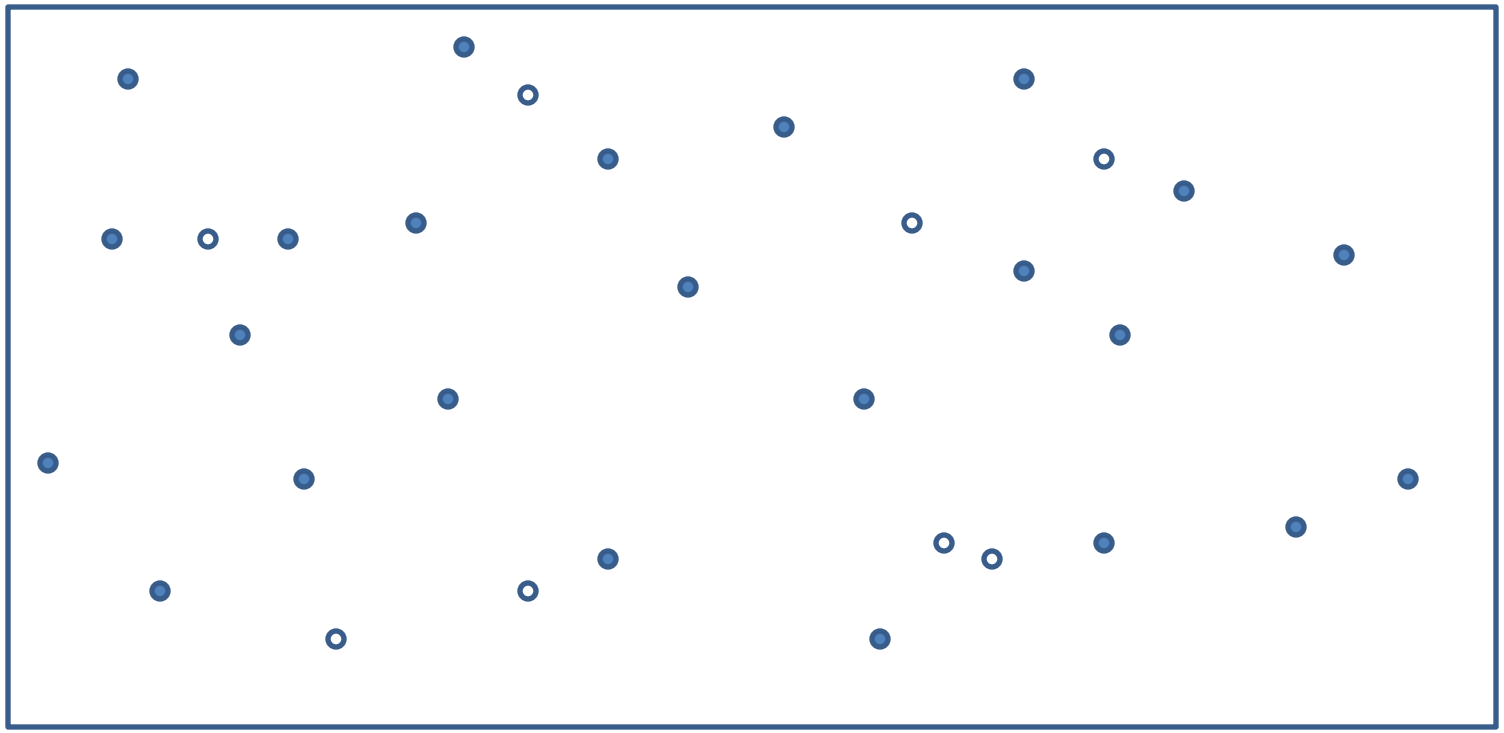
Principles of sampling

- Population vs Sample
- Want representative sample of some target population
- Need every member of the population to have non-zero probability of being sampled
- Must be able to estimate the probability of sampling any unit chosen

Simple random sampling

- All units in target population are known
- Sample is chosen randomly
- Each unit has an equal probability of being chosen
- Units may be individual, households, …

Simple random sampling



Unit sampled

Unit not sampled

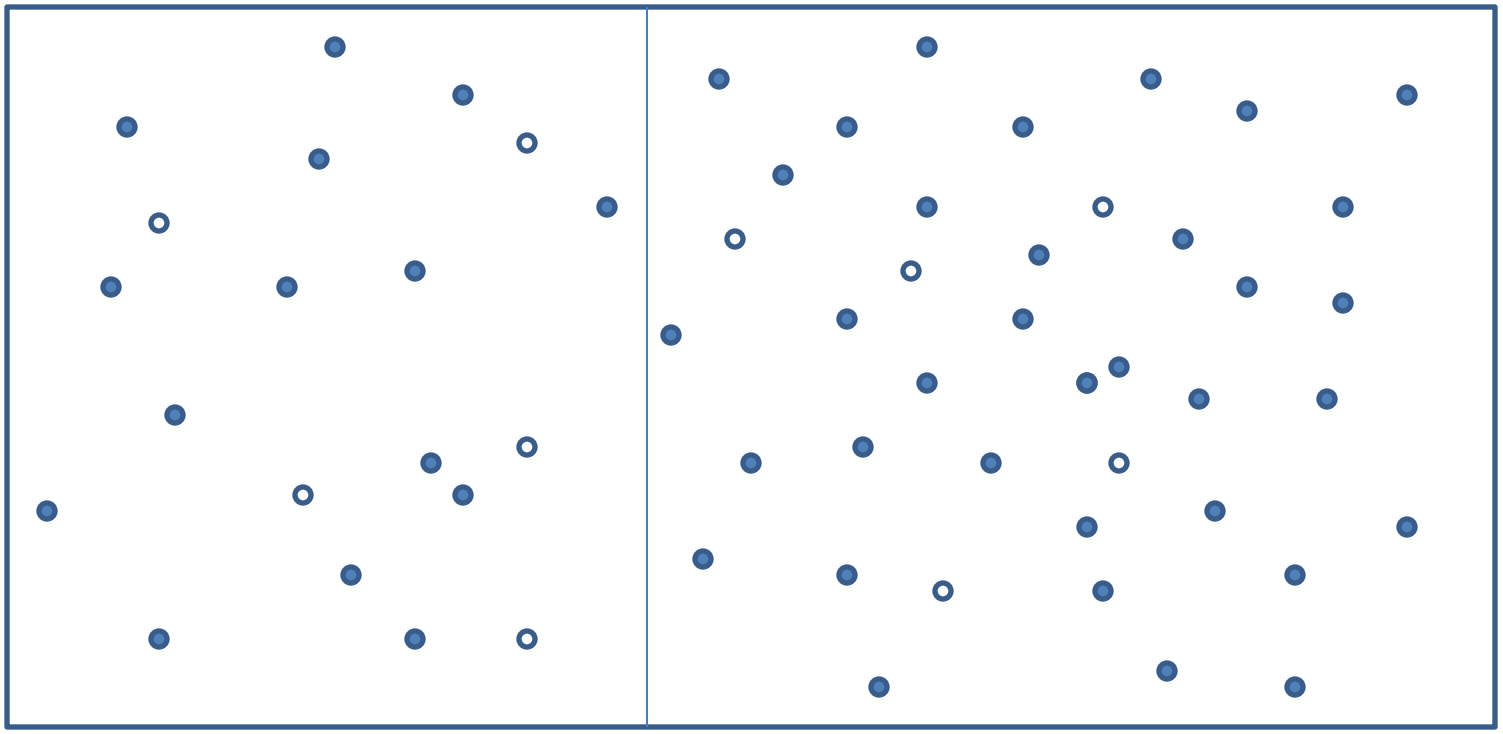
More complex designs

- Stratification
- Clustering
- Multi-stage
- Combinations

Stratified random sampling

- Population of interest is divided into strata (e.g., male and female; young, middle-aged, old)
- Simple random sample is chosen from each stratum
- Probabilities of selection between the strata can vary
- May be more cost-efficient than simple random sampling

Stratified random sampling



Stratum 1

Stratum 2

Unit sampled

Unit not sampled

Calculations in stratified sampling using weights

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04		
2	2000	400	0.2	0.1		
Total	7000	900	0.13	0.06		

Calculation of weights in stratified sampling

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04	200	
2	2000	400	0.2	0.1	200	
Total	7000	900	0.13	0.06	400	

Calculation of weights in stratified sampling

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04	200	16
2	2000	400	0.2	0.1	200	43
Total	7000	900	0.13	0.06	400	59

Calculation of weights

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04	200	16
2	2000	400	0.2	0.1	200	43
Total	7000	900	0.13	0.06	400	59

Weight, w = 1 / P(selected)

Stratum 1: $w_i = 1/ 0.04 = 25$

Stratum 2: $w_i = 1/ 0.1 = 10$

Estimation of number in population with disease

- Label $X_i = 0$ if the disease is absent

  and 1 if it's present for person i
- The our estimate of the number of people with the disease in the population is

  $\Sigma (w_i X_i)$
- And the estimate of the proportion in the population with the disease is

  $\Sigma (w_i X_i) / \Sigma w_i$

Application to our numerical example

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04	200	16
2	2000	400	0.2	0.1	200	43
Total	7000	900	0.13	0.06	400	59

For Stratum 1, there are 16 people with $X_i = 1$
and 184 people with $X_i = 0$
The weight for each person is 25
Do the same for stratum 2.

Then $\Sigma (w_i X_i) = 830$ is our estimate of the number with the disease

Application to our numerical example

Stratum	Population			Sample		
	N	# with disease	Proportion	Fraction sampled	n	# with disease
1	5000	500	0.1	0.04	200	16
2	2000	400	0.2	0.1	200	43
Total	7000	900	0.13	0.06	400	59

Estimate of the population proportion with disease

$= \Sigma (w_i X_i) / \Sigma w_i$

$= 830 / 7000$

$= 0.12$

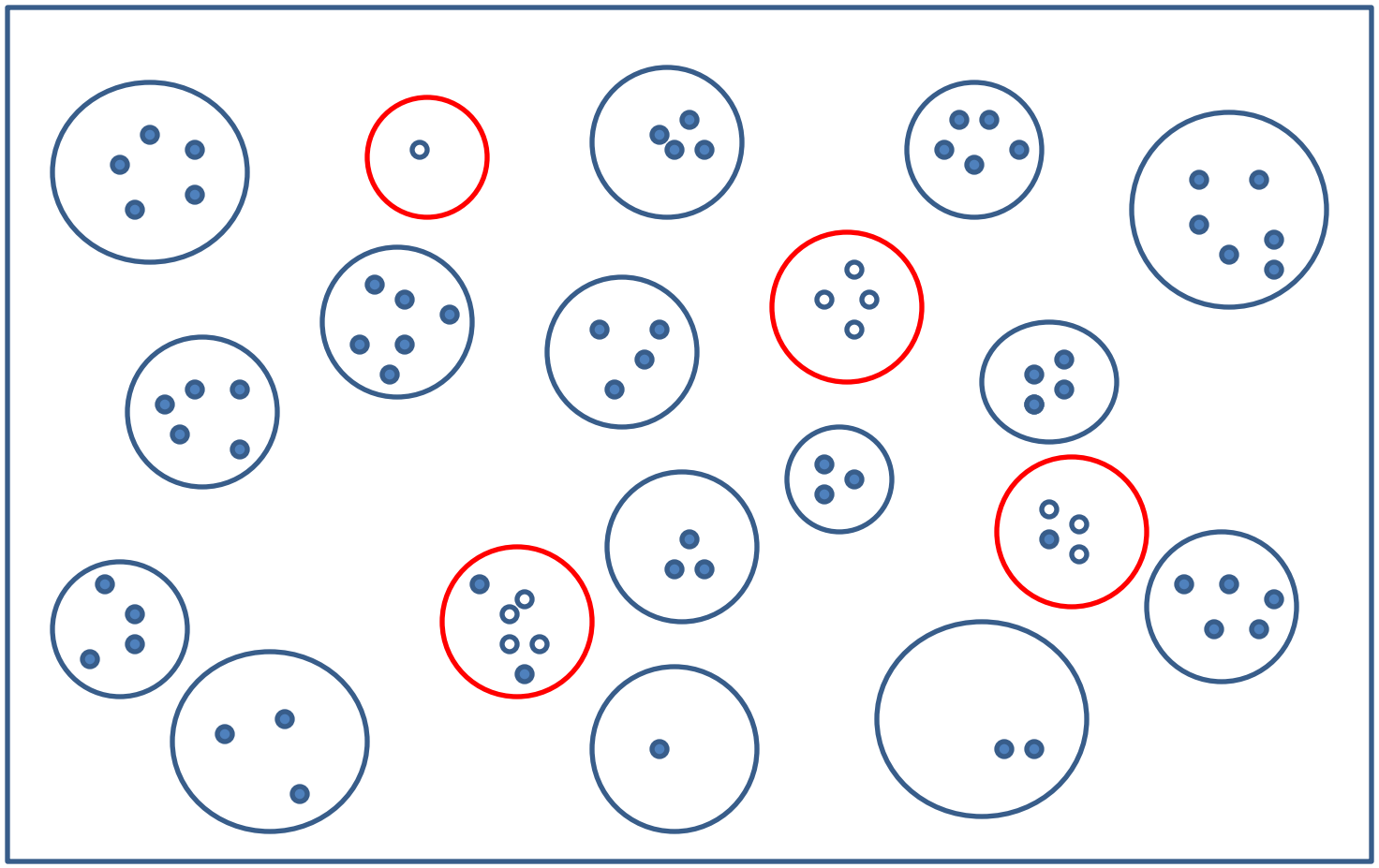
There are formulae to estimate the variance of the proportion.

And we can do this for continuous (interval) data.

Cluster sampling

- For efficiency, one may sample people within certain groups
- Examples:
  - sample towns and then sample people or households within each town
  - Sample households and interview everyone in household

Cluster sampling



Unit (cluster) sampled

Unit not sampled

Individual sampled

Individual not sampled

Cluster sampling

- Must allow for the lack of independence in the sampling – e.g., people in same family have similar diet
- Effective reduction in sample size, related to the ‘intra-cluster correlation’
- Trade-off between cost of sampling at random and need to sample more units (e.g., people) in total

Sampling in difficult situations

- E.g., disaster areas, war zones, Low Income Countries
- Various alternative methods
- E.g., Extended Program on Immunization (EPI)
- Methods typically have some limitations
- May have to balance bias, precision, speed, cost

Back to CLSA …

Aims of sampling in CLSA

- Choose representative sample of eligible Canadian residents

CLSA Participant Recruitment

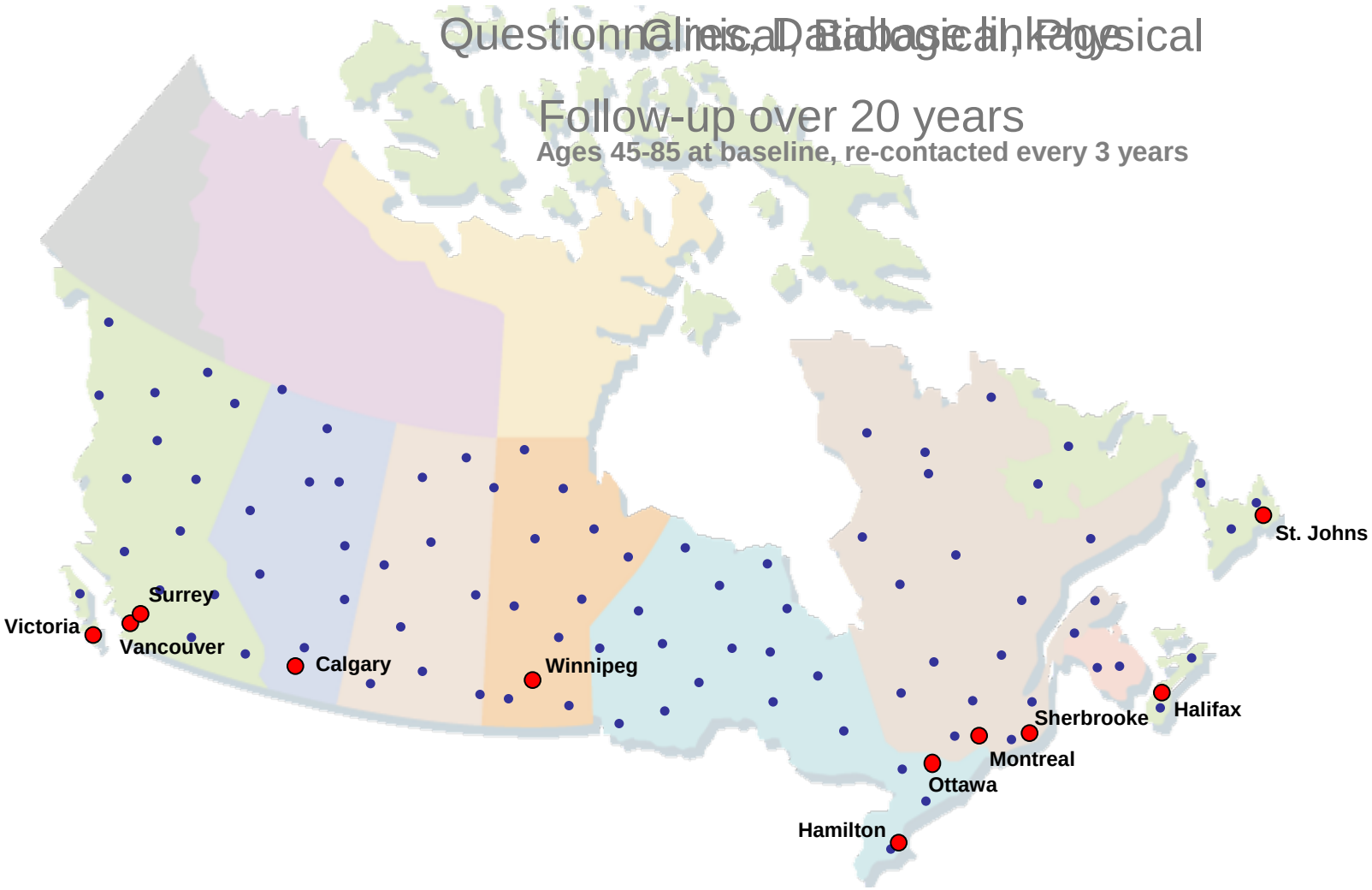
Inception Cohort: 50,000

In-depth data collection on 30,000 (at 11 sites)

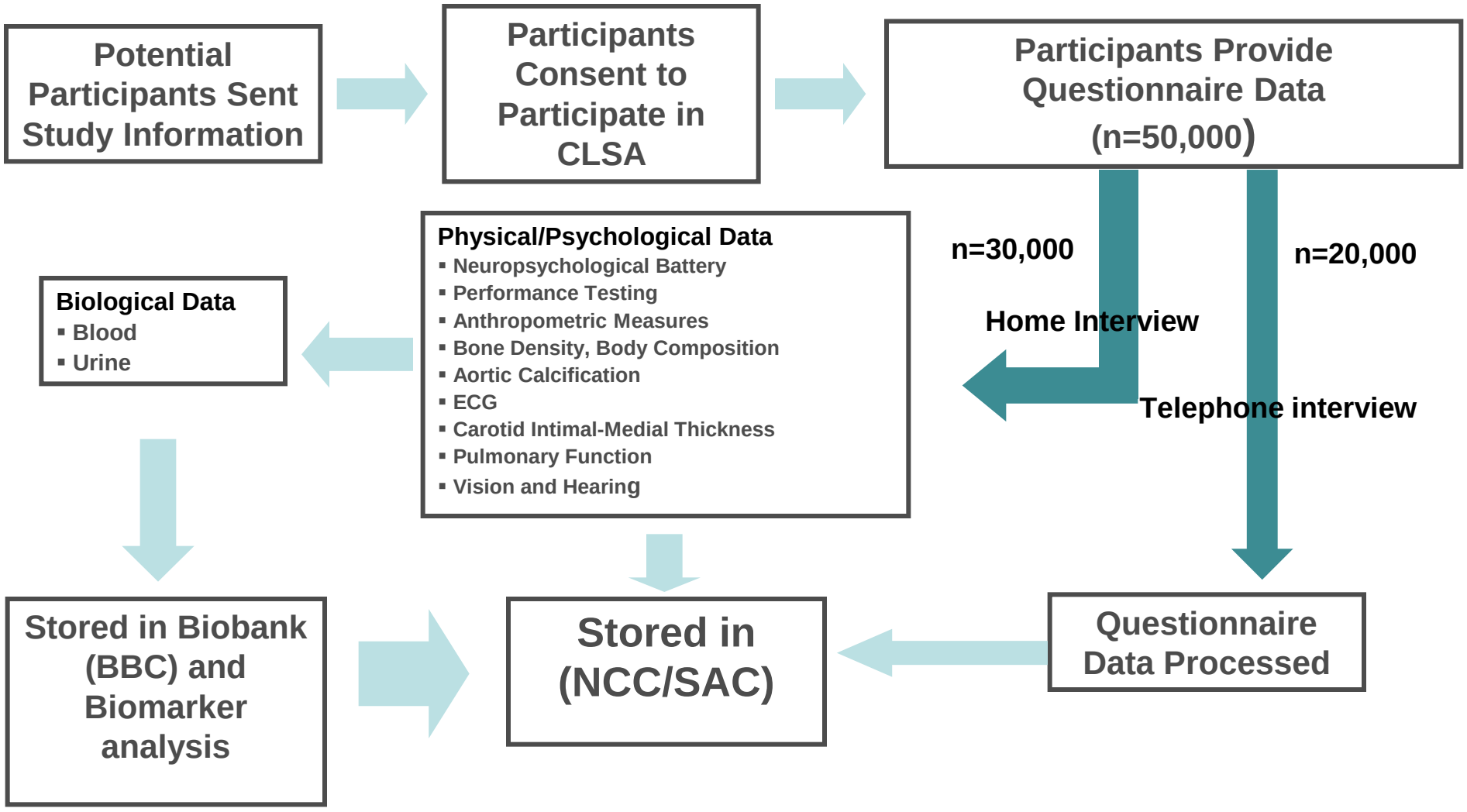
Questionnaire, Basic Physical, Clinical, Biological, Psychosocial

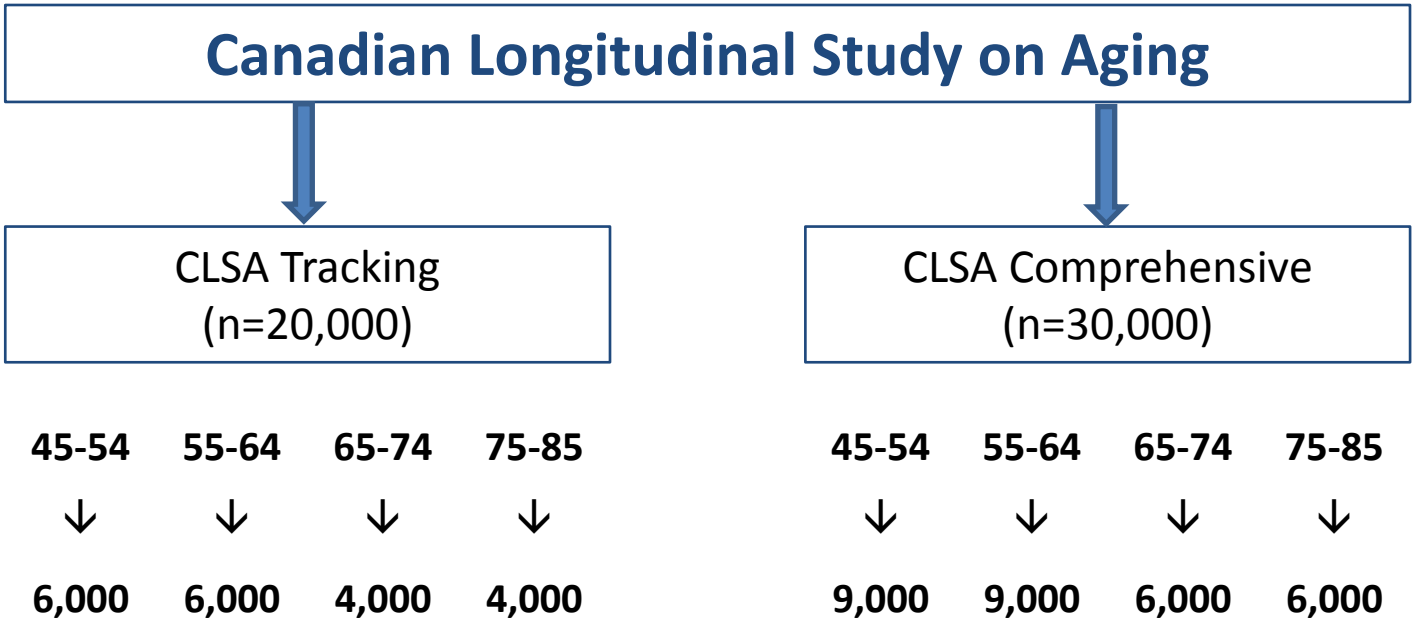
Follow-up over 20 years

Ages 45-85 at baseline, re-contacted every 3 years



CLSA Data Collection




Canadian Longitudinal Study on Aging
CLSA Tracking
(n=20,000)
CLSA Comprehensive
(n=30,000)
45-54
55-64
65-74
75-85
6,000
6,000
4,000
4,000
45-54
55-64
65-74
75-85
9,000
9,000
6,000
6,000

Potential Sampling Frames

- Canadian Community Health Survey Participants
- Provincial Health Registration Databases
- Random Digit Dialling

ALL OF THE ABOVE

- CCHS provided first part of sample
- Options for methods of selection of remaining participants:
  - Using provincial health registries - *preferred*
  - Random digit dialing
- In several provinces, we cannot use registries, so need to do RDD

Recruitment from the CCHS

- CLSA collaborated with Statistics Canada to develop the CCHS Healthy Aging Questionnaire
- Target population: People aged 45 and over living in private occupied dwellings in the ten provinces
- Excluded:
  - Residents of the three territories
  - Persons living on Indian reserves or Crown lands
  - Persons living in institutions
  - Full-time members of the Canadian Forces
  - Residents of some remote regions

Recruitment from the CCHS, *ctd.*

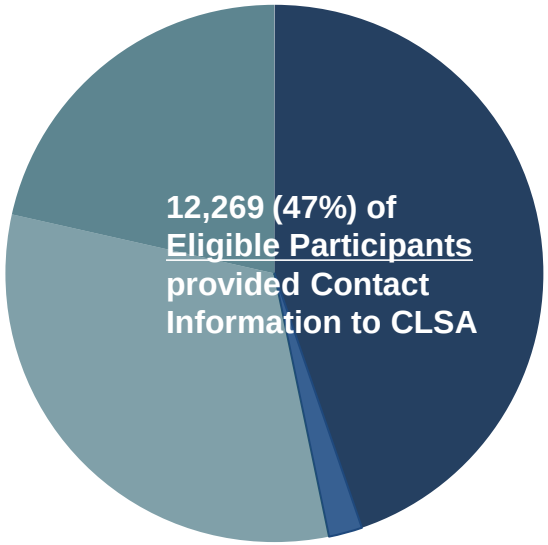
Multi-stage sampling

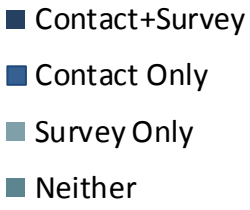
- Sampling frame 2006 Census
- Selection
  - Clusters based on Census dissemination area blocks
  - Dwellings within cluster
  - Person within dwelling
- Response Rate
  - Household-level 80.8%
  - Person-level 92.1%
  - Overall 74.4%

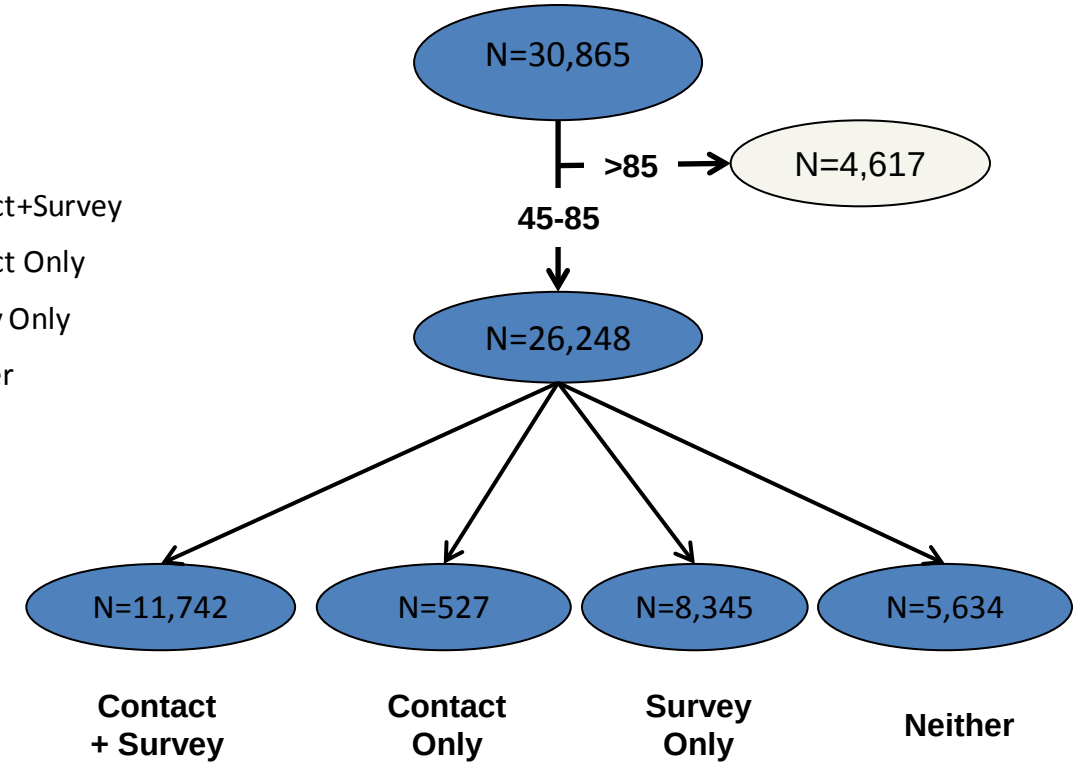
Recruitment from the CCHS, *ctd.*

Participants were asked to share:

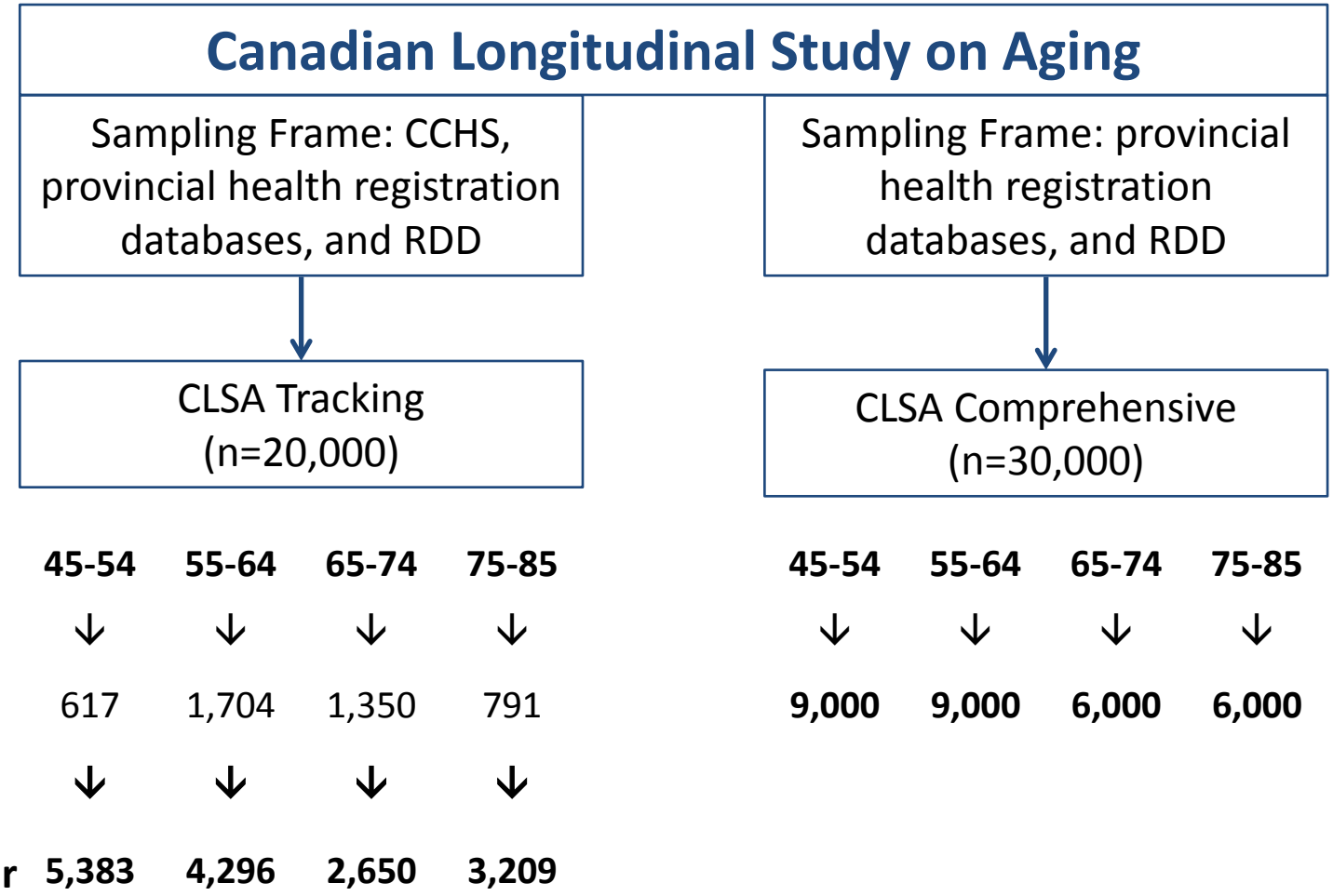
- **Their contact information with the CLSA (for recruitment)**
- **Their survey responses with the CLSA (for analysis)**







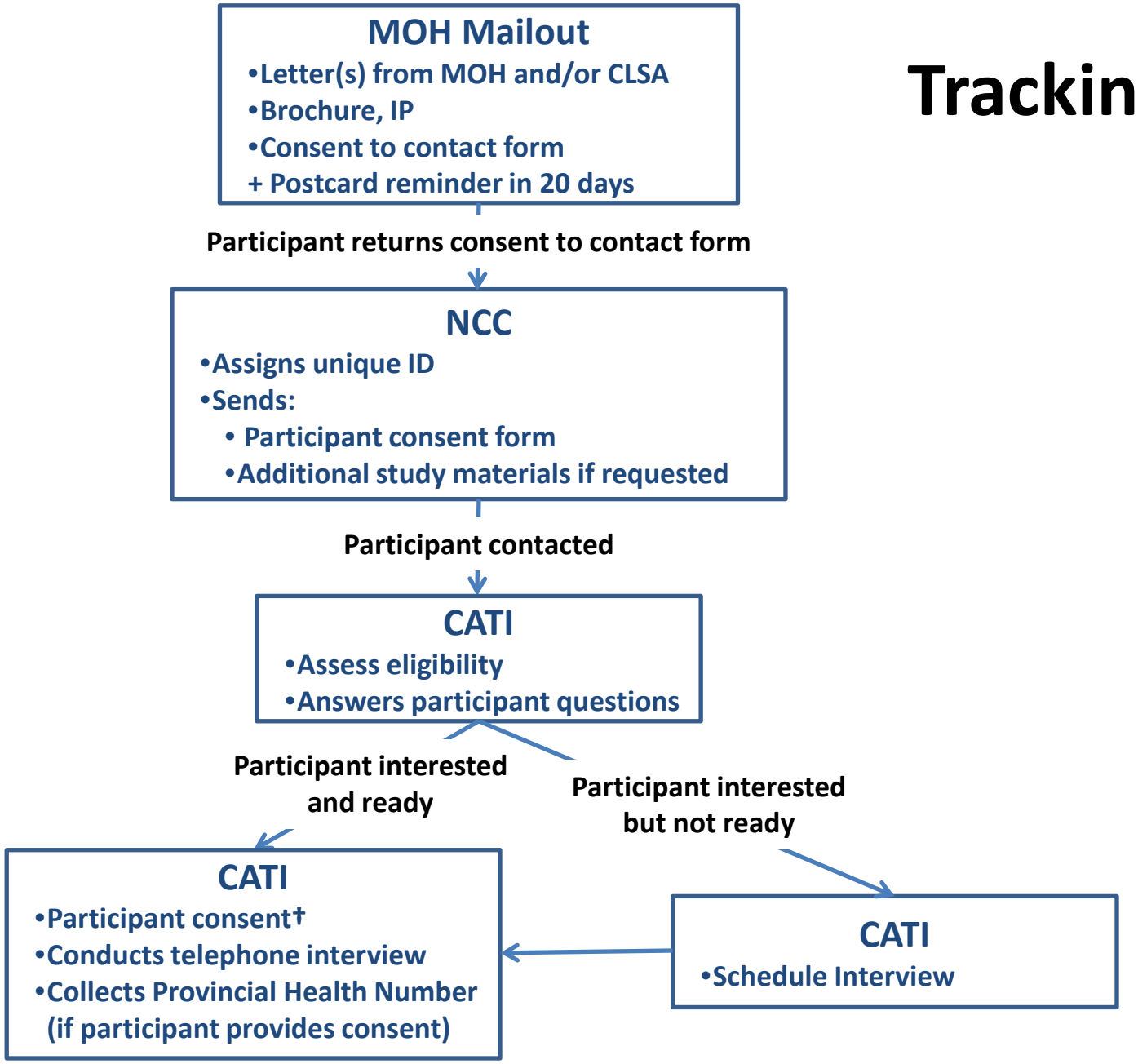
Recruitment from the CCHS, *ctd.*



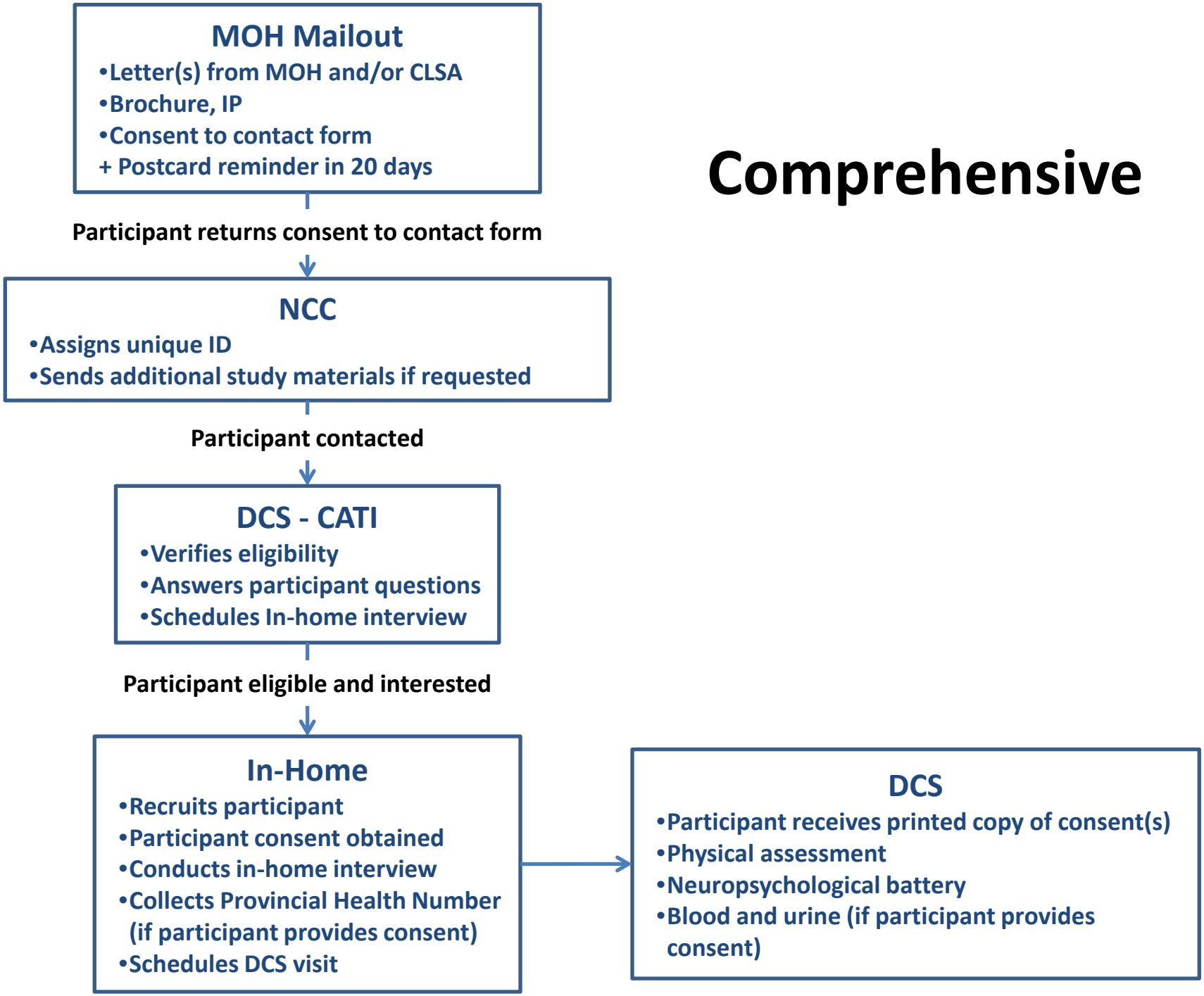
Recruitment from Provincial health registration databases

- 2005
  - Feasibility study to explore practical, methodological and ethical aspects of accessing Health Care Utilization data from Provincial databases (published 2009)
- 2009-2011
  - Several meetings with Provincial Data Stewards and Privacy Commissioners to negotiate access to health registration databases for sampling

Tracking


MOH Mailout
•Letter(s) from MOH and/or CLSA
•Brochure, IP
•Consent to contact form
+ Postcard reminder in 20 days
Participant returns consent to contact form
NCC
•Assigns unique ID
•Sends:
• Participant consent form
•Additional study materials if requested
Participant contacted
CATI
•Assess eligibility
•Answers participant questions
Participant interested and ready
Participant interested but not ready
CATI
•Participant consent†
•Conducts telephone interview
•Collects Provincial Health Number (if participant provides consent)
CATI
•Schedule Interview

Comprehensive



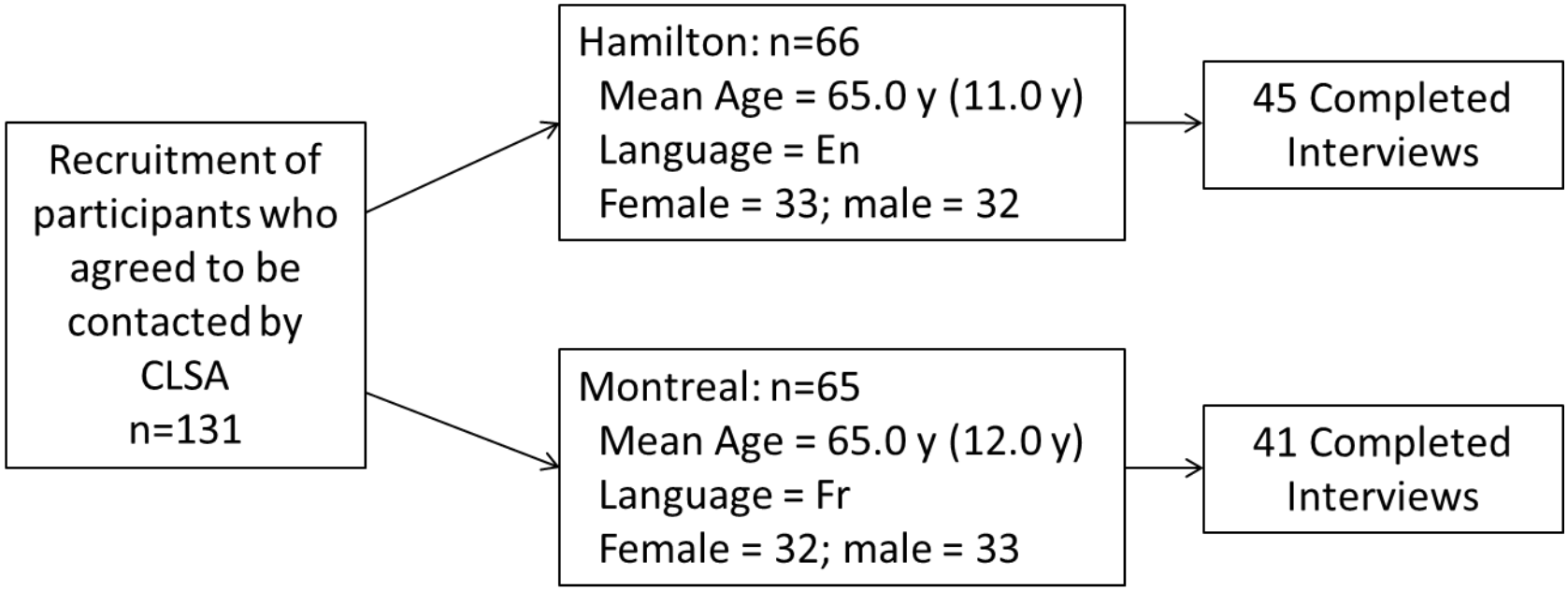
Recruitment from Provincial health registration databases

- Based on previous studies (completed in early 2000's) we anticipated a 15-20% recruitment rate
- Preliminary results from PEI and New Brunswick suggest that the recruitment rate may be lower ~7-10%

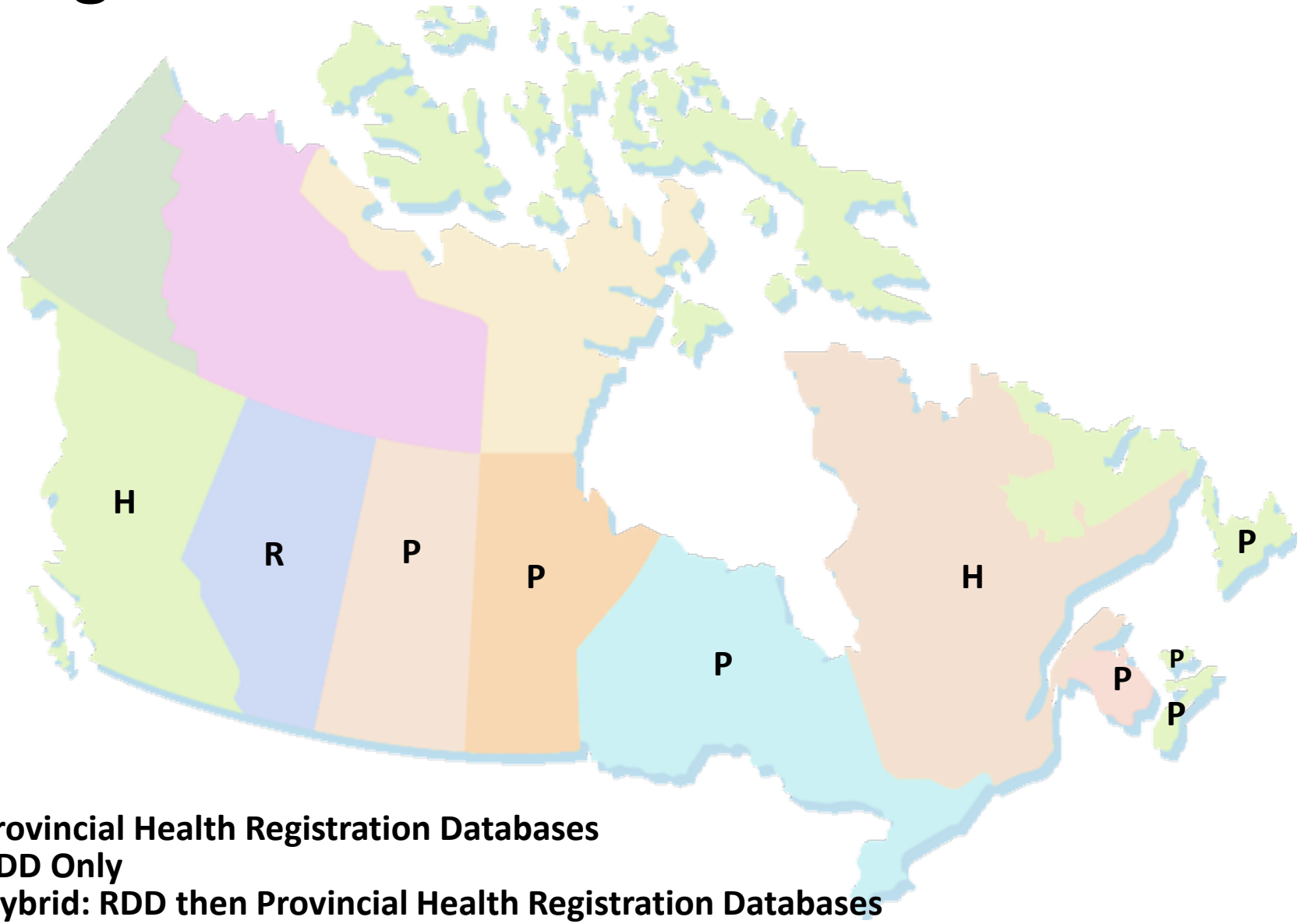
RDD – Tracking + Telephone Administered Questionnaires Pilot

	Mean Age (SD)	Language	Sex
Injury Module (n=200)	70.5 y (11.2 y)	Fr=100 En=100	F=92 M=108
Tracking Baseline (n=50)	64.3 y (10.6 y)	Fr=23 En=27	F=33 M=17
Maintaining Contact - Comp (n=25)	61.3 y (9.0 y)	Fr=12 En=13	F=12 M=13
Maintaining Contact - Tracking (n=25)	63.1 y (10.0 y)	Fr=15 En=10	F=13 M=12
TOTAL (n=300)	62.7 y (10.8 y)	Fr=150 En=150	F=150 M=150

RDD – Comprehensive Pilot

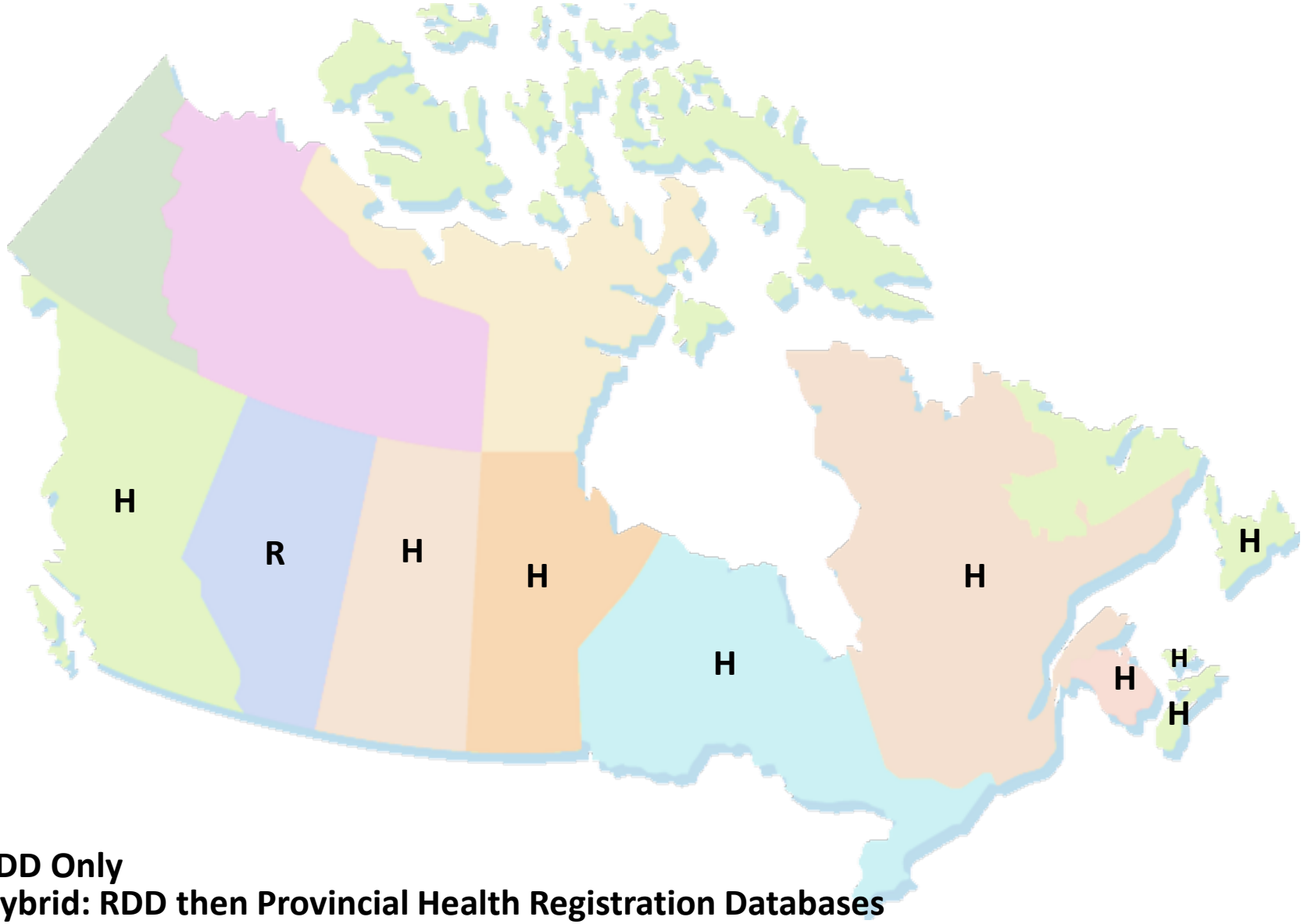


Original Plan for Additional Recruitment



P=Provincial Health Registration Databases
R=RDD Only
H=Hybrid: RDD then Provincial Health Registration Databases

Plan for Additional Recruitment



R=RDD Only
H=Hybrid: RDD then Provincial Health Registration Databases

RDD approach

- In principle, idea is simple
- Randomly sample numbers as far as possible in specified area codes and with next 3 digits in relevant area
- Identify eligible people at each number
- Randomly choose one person
- Recruit willing participants until ‘quota’ filled

Issues in using RDD

- Identifying numbers in specified area
- Having up-to-date list of numbers for target population
- Ability to compute sample weights
- Presence of landlines and/or cellphones
- Eligibility within household – changes over time
- Method of initial contact
- Households without phones
- Numbers may be businesses, out of order, etc.
- People away from home (snowbirds, etc.)

Cell phones and landlines

- Statistics Canada survey December 2010
- Supplement to Labour Force Survey
- Households using cell phones exclusively:
  - Overall: 13%
  - Age 18-34 50%
  - Over 35 8%
  - Over 55 4%
- Increasing over time
- Landlines reach nearly all our eligibles

Combining samples from cell phones and landlines

- Methods have been described
- Need to determine all phones in each household
- Keep logs of unfilled quotas (age-sex numbers)
- Interviewers construct rosters of eligibles within households and randomly choose one

Some issues with cell phones

- Ethical: incoming calls may cost user; privacy; activity when answering (driving, etc); children
- Cost: AAPOR states at least 2x, maybe 3-4x cost of landline survey
- Getting addresses
- Quality of data (may be similar to landlines)

Source: AAPOR

‘Cold calling’ vs prior contact/letters

- Time and expense of mailing letters (only possible when we have name and address)
- May increase willingness to talk to interviewers (call display)
- However, many households will not include any eligible people

Contacting subjects

- On average, anticipate making many calls to recruit a single person
  - Up to 7-10 calls to obtain response
  - Leave message?
  - Willingness to participate
- Working on assumption of 20% ‘recruitment rate’ for health registry data (15% in 75-85 age group)
- Exclude households without a phone

Estimation of sampling weights

- Calculate probability of selecting sampling unit (in CLSA, unit = person)
- Account for different sampling frames
- Allow for non-response
- Use weights to estimate parameters (means, proportions, etc) for the target population
- Various assumptions required

Sources of the CLSA sample

- Tracking cohort:
  - CCHS
  - Health registries
  - RDD
- Comprehensive cohort
  - Health registries
  - RDD

Probabilities for the CCHS

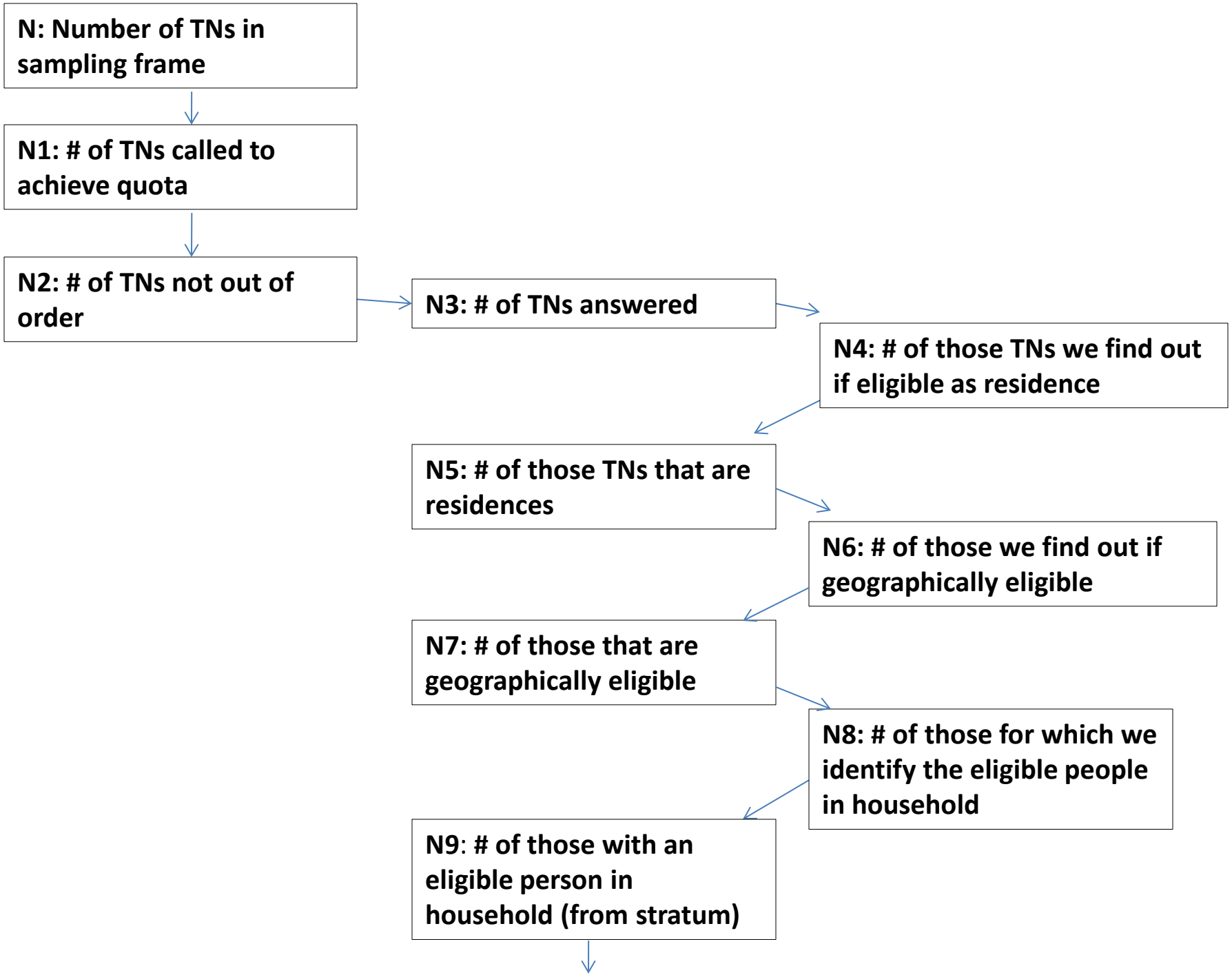
- Provided by StatsCan
- Must allow for non-response in the CLSA
- Some issues on confidentiality – information sharing

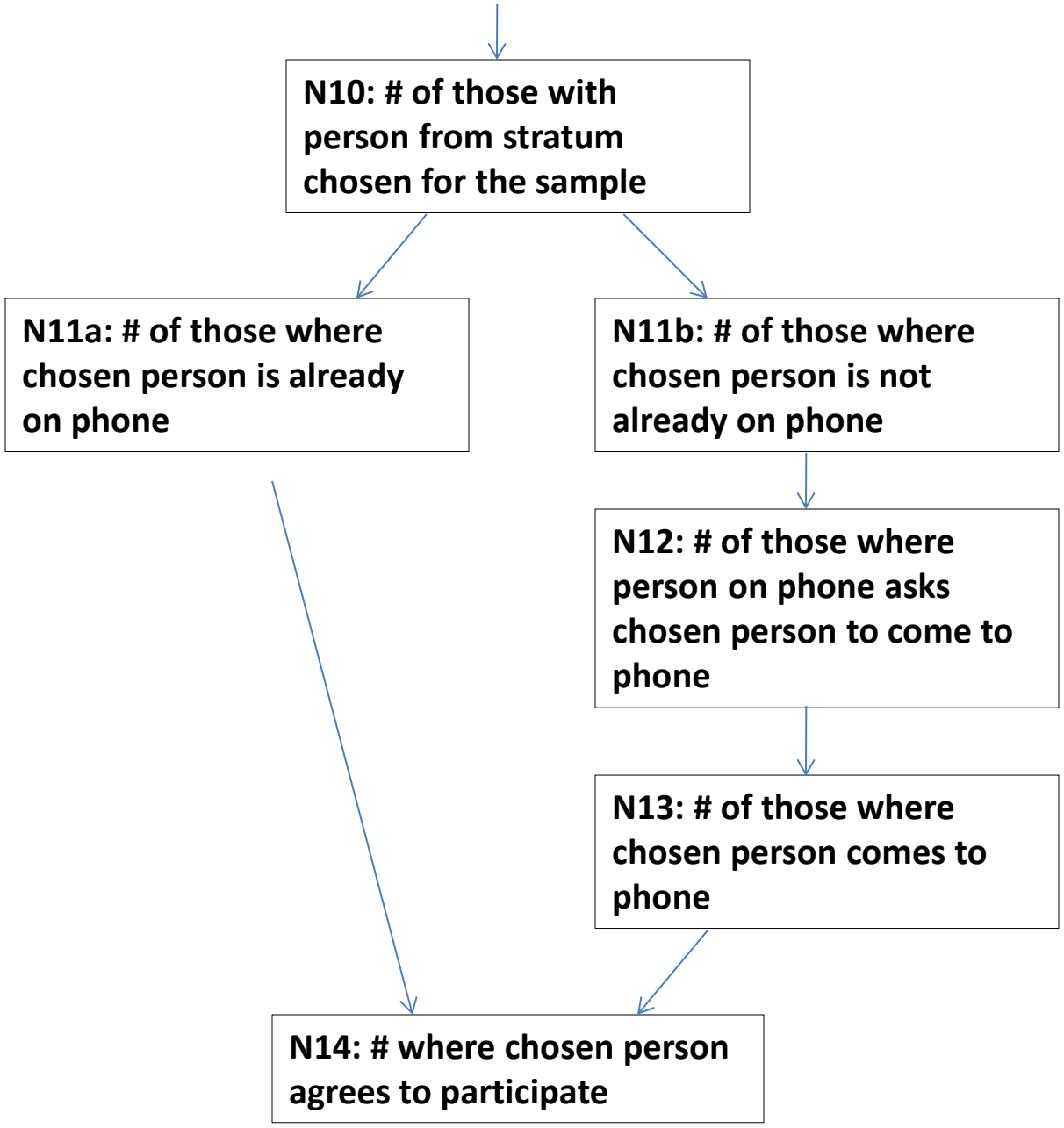
Probabilities for the health registries

- Health registries (HR) have list of (virtually) all target population
- HR can provide numbers of people in each age-sex group for the province (denominators)
- Mail-outs from HRs will lead to estimation of proportion of ineligibles and adjustment of denominators
- Estimate probability of participation

Probabilities for RDD

- Phone numbers in range (population) identified
- For tracking, all numbers in province
- For comprehensive, some eligibility established during contact call
- Eligibility: private residence, geography, age, competent to interview, quota not filled, other
- Probability of selection is product of various probabilities


N: Number of TNs in sampling frame
N1: # of TNs called to achieve quota
N2: # of TNs not out of order
N3: # of TNs answered
N4: # of those TNs we find out if eligible as residence
N5: # of those TNs that are residences
N6: # of those we find out if geographically eligible
N7: # of those that are geographically eligible
N8: # of those for which we identify the eligible people in household
N9: # of those with an eligible person in household (from stratum)


N10: # of those with person from stratum chosen for the sample
N11a: # of those where chosen person is already on phone
N11b: # of those where chosen person is not already on phone
N12: # of those where person on phone asks chosen person to come to phone
N13: # of those where chosen person comes to phone
N14: # where chosen person agrees to participate

Some probabilities estimated

$$P_{noo} = \frac{TNs\ not\ out\ of\ order}{TNs\ called\ to\ achieve\ quota}$$

$$P_{res} = \frac{TNs\ that\ are\ residences}{TNs\ we\ find\ out\ if\ eligible\ as\ residence}$$

$$P_{part} = \frac{number\ \ agreeing\ to\ participate}{number\ selected\ to\ participate}$$

Combining samples from different sources

- Want overall P(Participation)
- Use addition rule of probability
- E.g., for someone chosen via RDD, need P(Selected by RDD) AND P(Selected in CCHS)
- Latter is an average probability, not an individual one
- Similarly for selection through health registries

Additional issues

- When P(Participation) is based on the product of probabilities, have to assume independence of probabilities
- Confidentiality conditions may mean, e.g., we call people in RDD who were in the CCHS and did not want to participate in the CLSA
- In RDD, have to allow for multiple phones in the household
- At some point, likely to fill some age/sex quotas; then only recruit unfilled quotas

Summary

- Various sources of participants for CLSA
- Each has its own strengths and limitations
- Need to estimate sampling probabilities for each source
- Aiming for representativeness – but …
- Various assumptions must be made