# Sampling in the CLSA

Lauren Griffith

Harry Shannon

McMaster University

CLSA Methods Seminar

29 November 2011

# 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

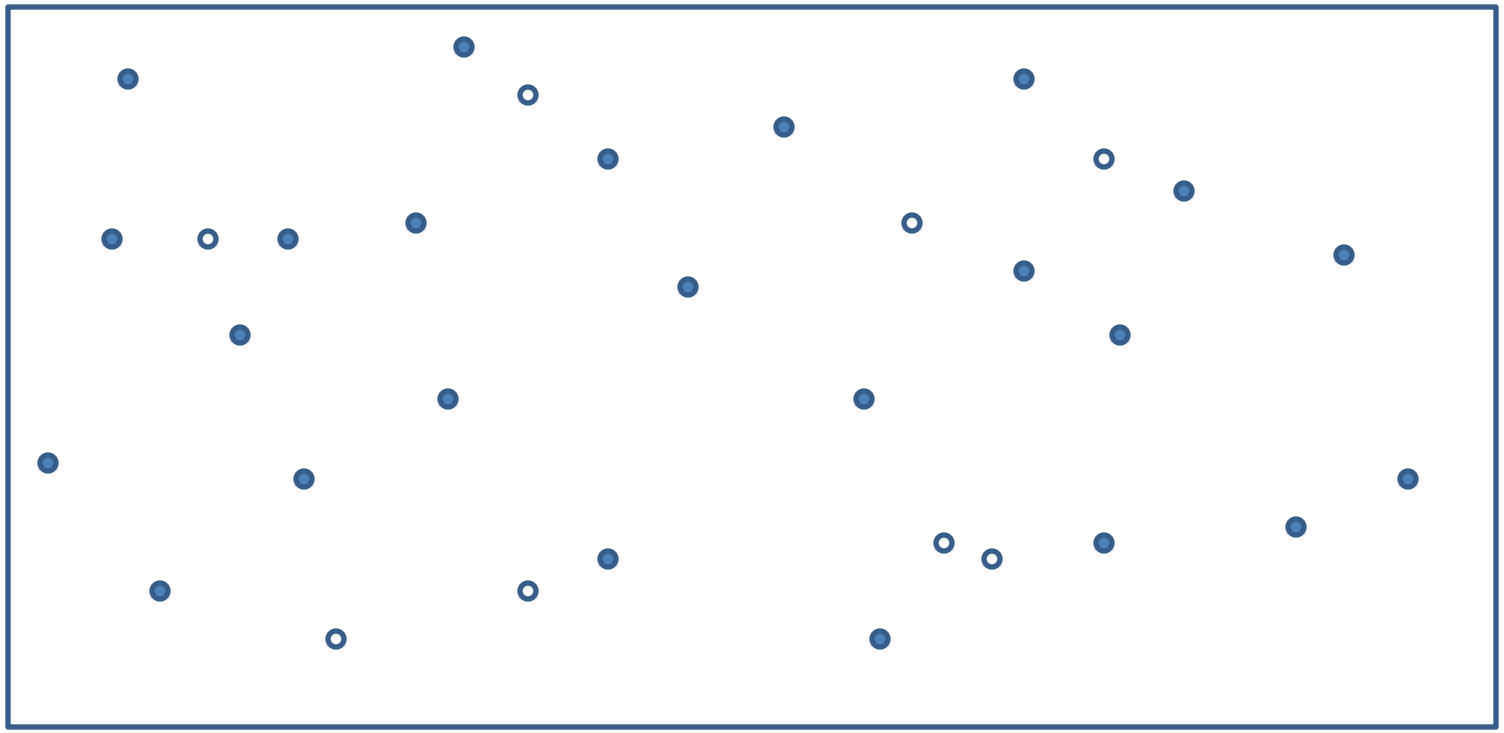
# Principles of sampling

- 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 individual chosen

# Simple random sampling

- All units in target population are known
- Sample is chosen randomly
- Each unit has an equal probability of being chosen

# Simple random sampling



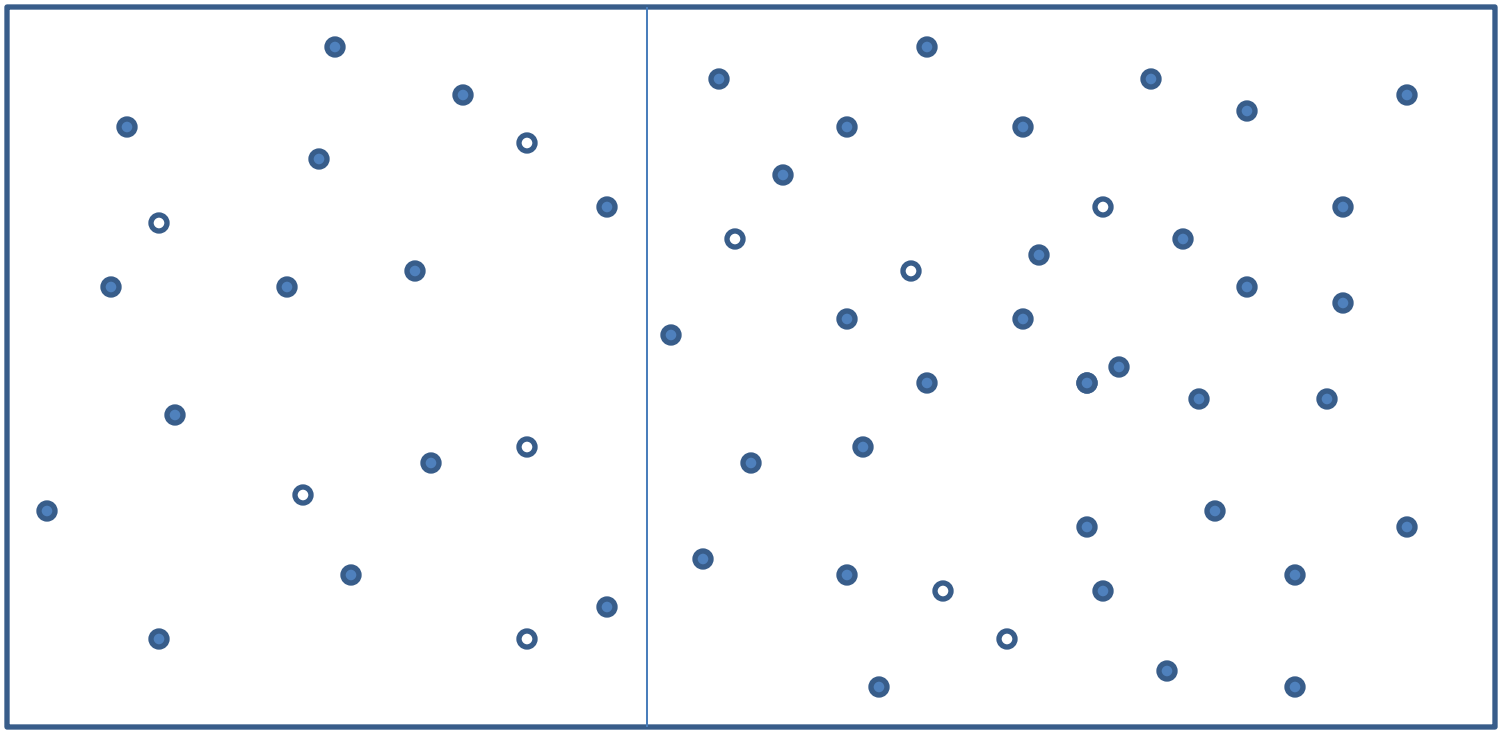
Unit sampled

Unit not sampled

# 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

# Stratified random sampling



Stratum 1

Stratum 2

Unit sampled

Unit not sampled

# More complex designs

- Stratification
- Clustering
- Multi-stage
- Combinations

# Sampling in difficult situations

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

Back to CLSA …

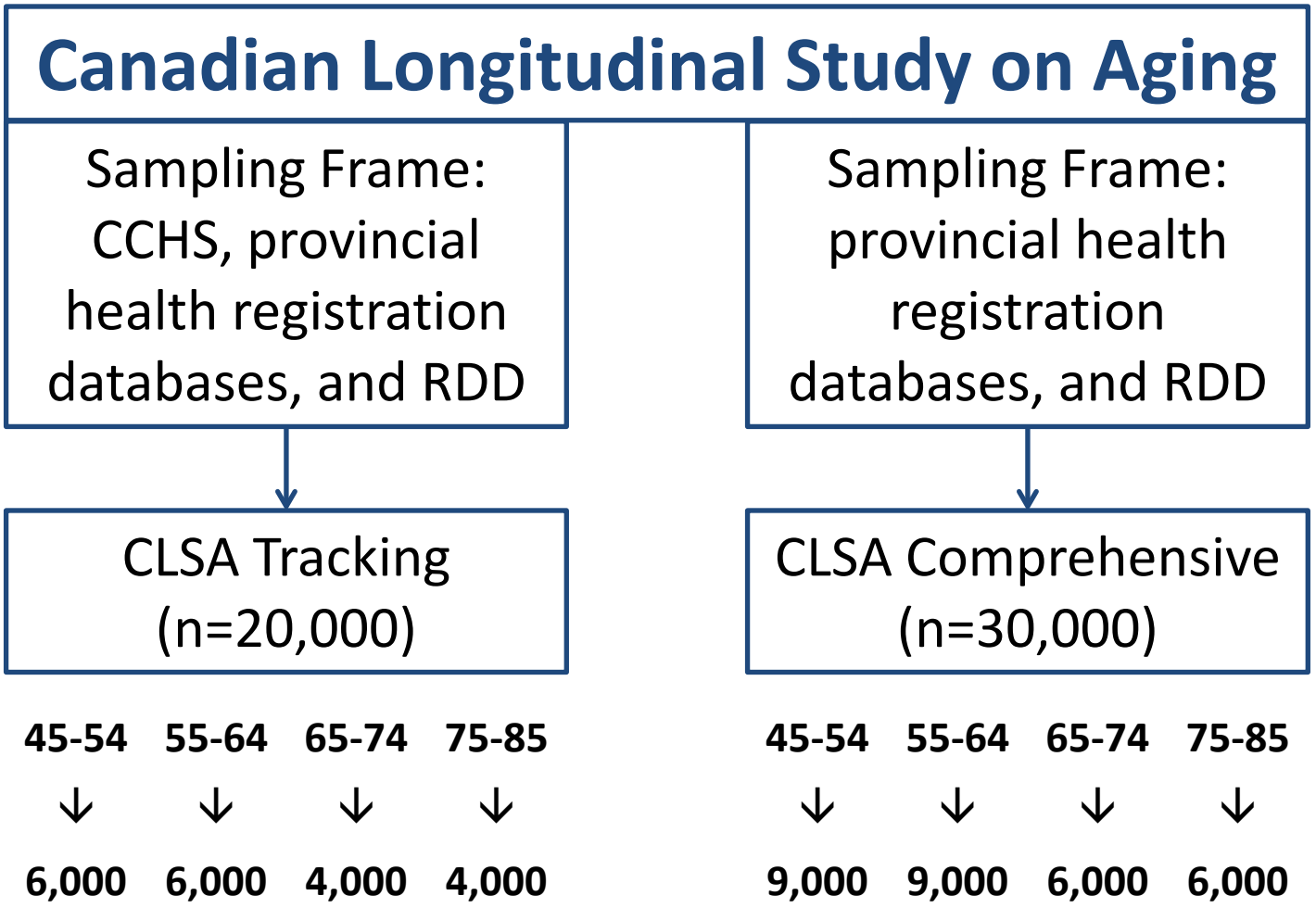
# Aims of sampling in CLSA

- Choose representative sample of eligible Canadians
  - 20K Tracking cohort; 30K Comprehensive cohort
  - Specified numbers in age-sex groups by province

# Potential Sampling Frames

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

ALL OF THE ABOVE


Canadian Longitudinal Study on Aging
Sampling Frame: CCHS, provincial health registration databases, and RDD
Sampling Frame: provincial health registration databases, and RDD
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

- 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.*

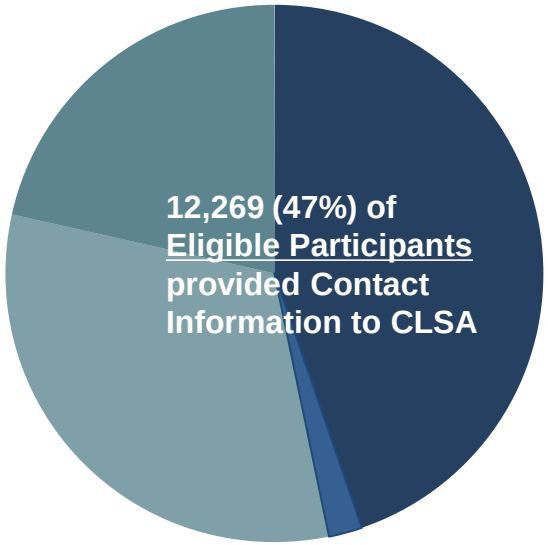
## Multi-stage sampling

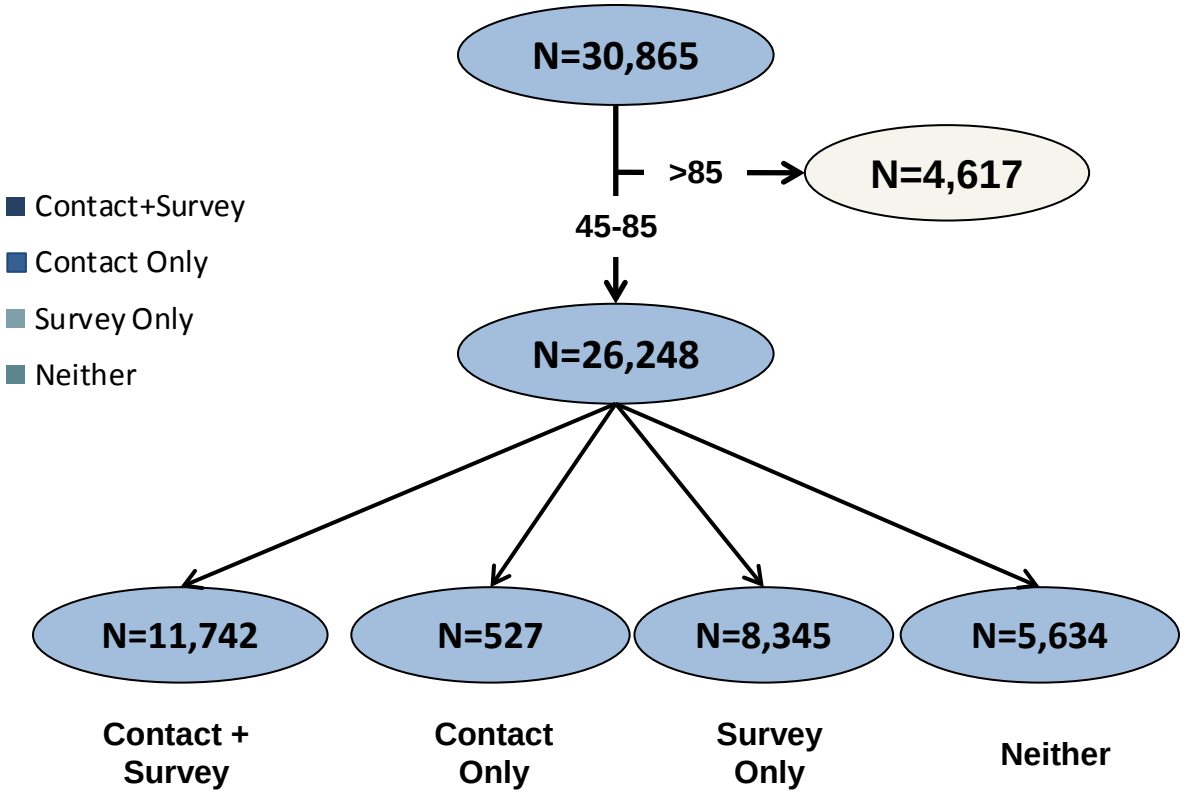
- Sampling frame 2006 Census
- Selection
  - Clusters based on Census dissemination area blocks
  - Dwellings within cluster
  - Person within dwelling
- Response Rate
  - Household-level 80.8%
  - Person-level 92.1%
  - Overall 74.4%

# Recruitment from the CCHS, *ctd.*

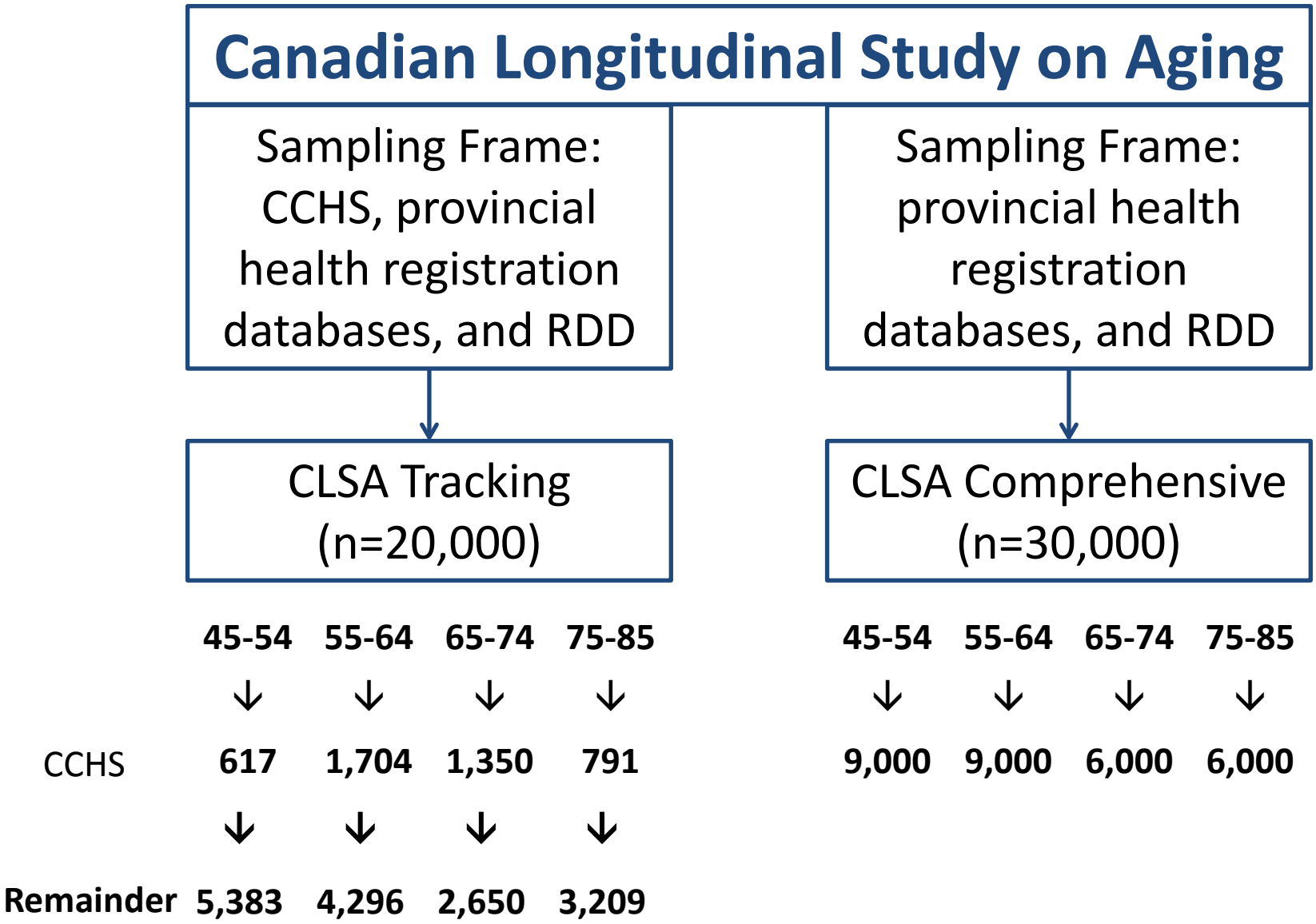
**Participants were asked to share:**

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



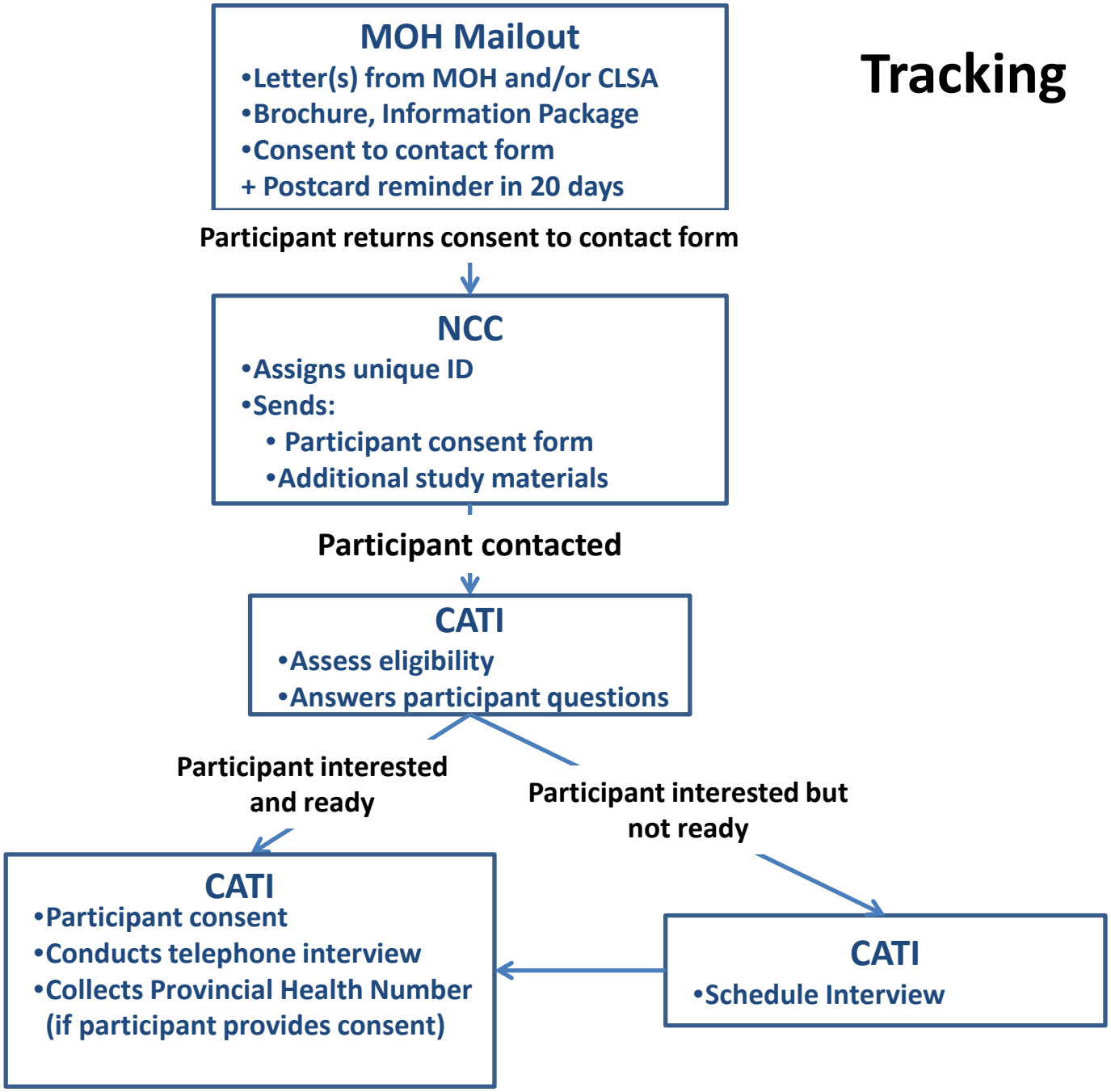
# Recruitment from the CCHS, *ctd.*



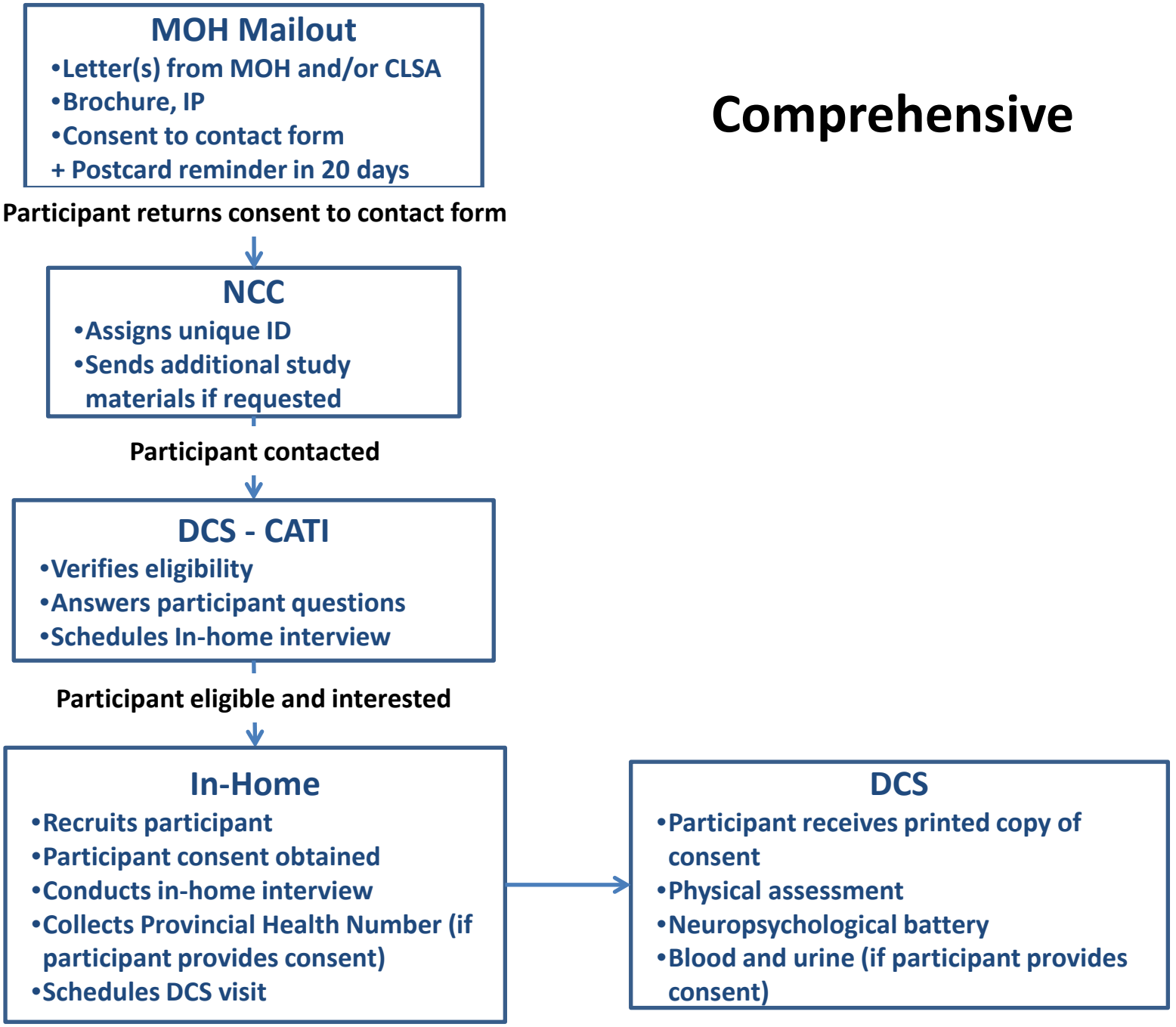
# Recruitment from Provincial health registration databases

- 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, Information Package
•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
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

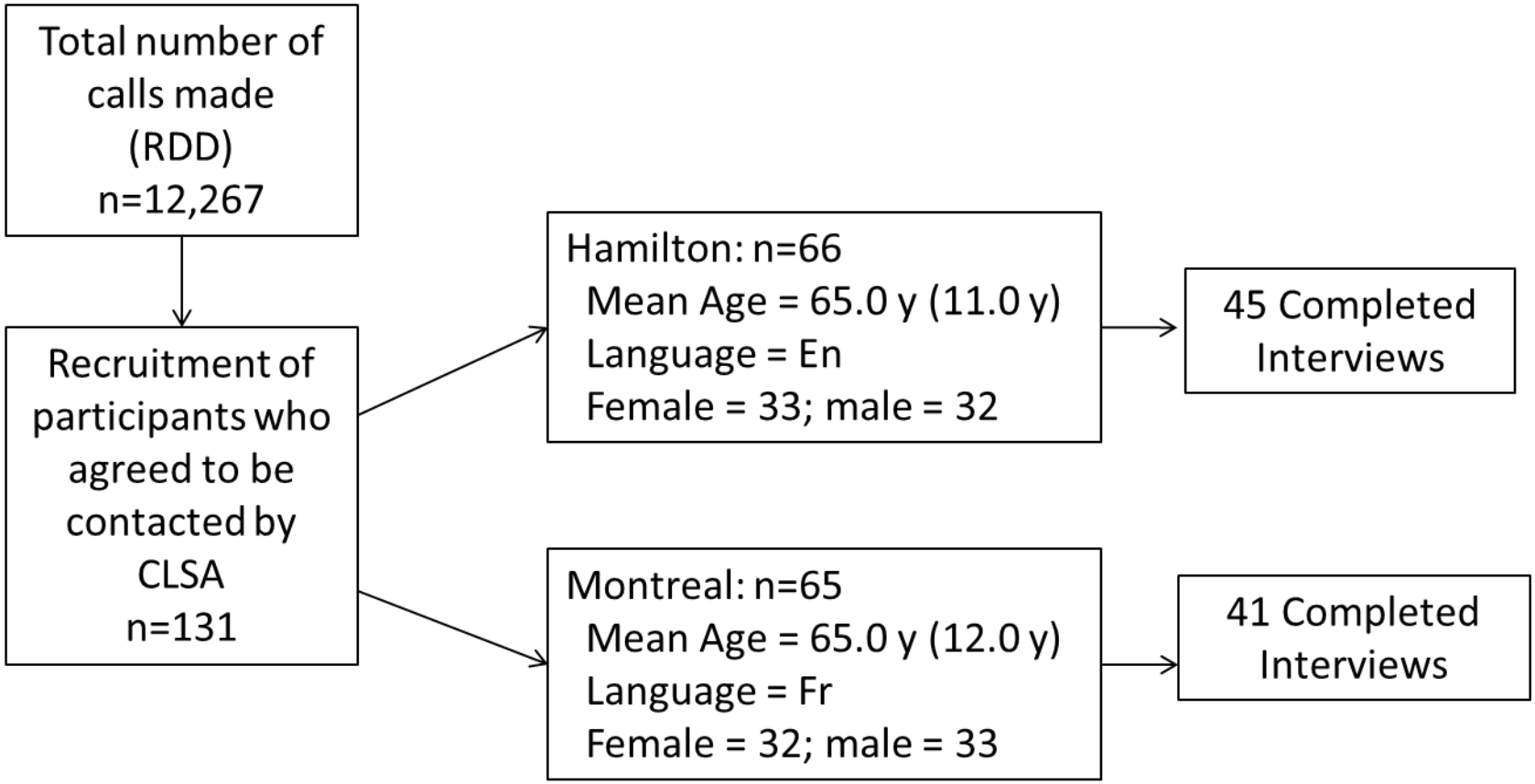


# RDD – Tracking + Telephone Administered Questionnaires Pilot

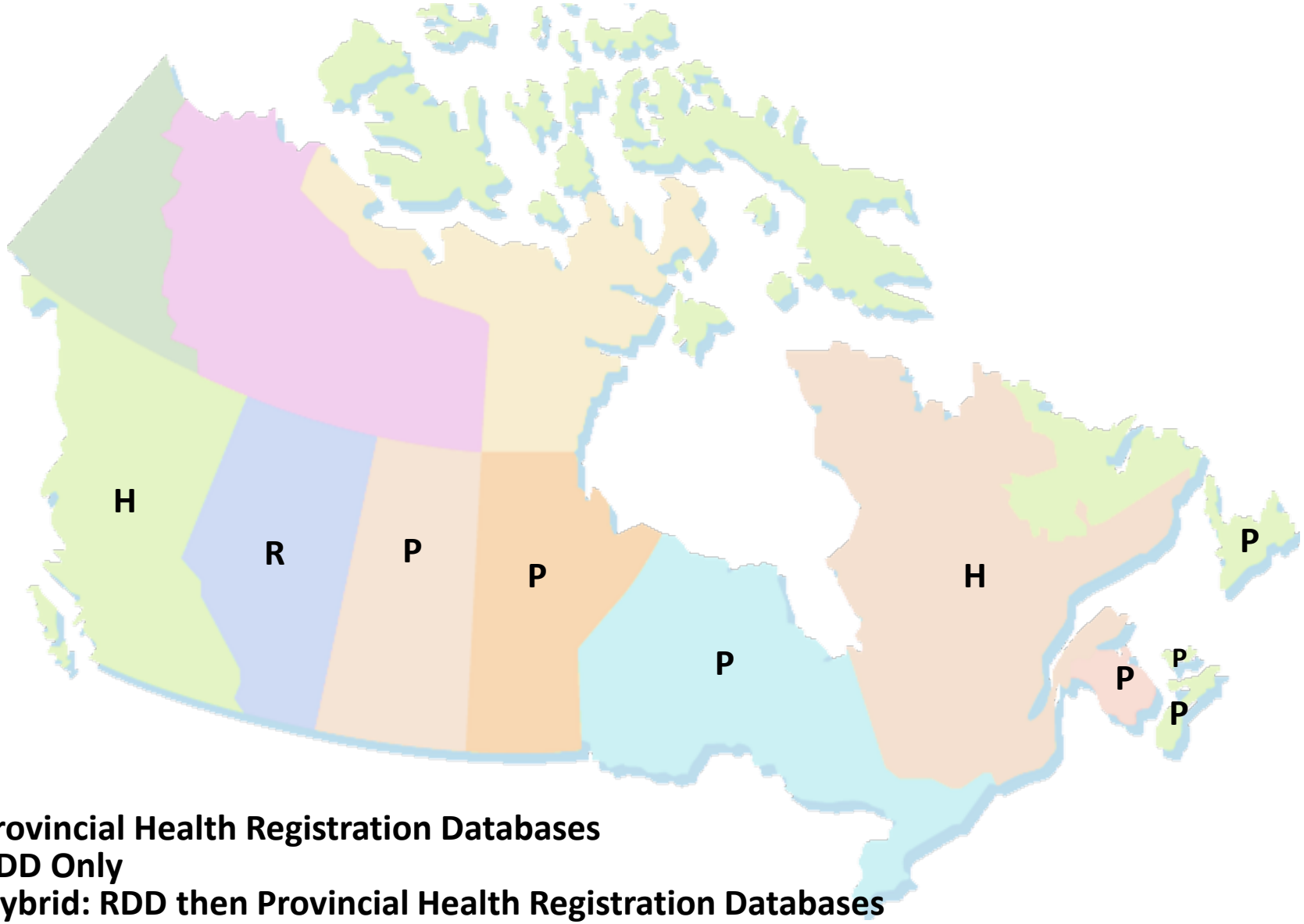
42,860 calls → 300 pilot participants

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

# RDD – Comprehensive Pilot



# Plan for Additional Recruitment




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

# Example of requirement by province
## Tracking cohort

**Alberta**

| | 45-54 | | 55-64 | | 65-74 | | 75-85 | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | |
| # Required | 306 | 306 | 306 | 306 | 204 | 204 | 204 | 204 | 2,040 |
| # Providing Contact Info | 121 | 128 | 153 | 193 | 108 | 138 | 74 | 107 | 1,022 |
| # Anticipated through CCHS | 28 | 35 | 56 | 82 | 53 | 64 | 33 | 25 | 376 |
| # Additional Participants | 278 | 271 | 250 | 224 | 151 | 140 | 171 | 179 | 1,664 |
| # Phone numbers to Sample* | X | X | X | X | X | X | X | X | X |

* This will depend on the recruitment rate per number sampled

# 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 bases
- Allow for non-response
- Weight = 1 / P(selection)
- Use to ‘weight up’ the sample to get estimates of 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

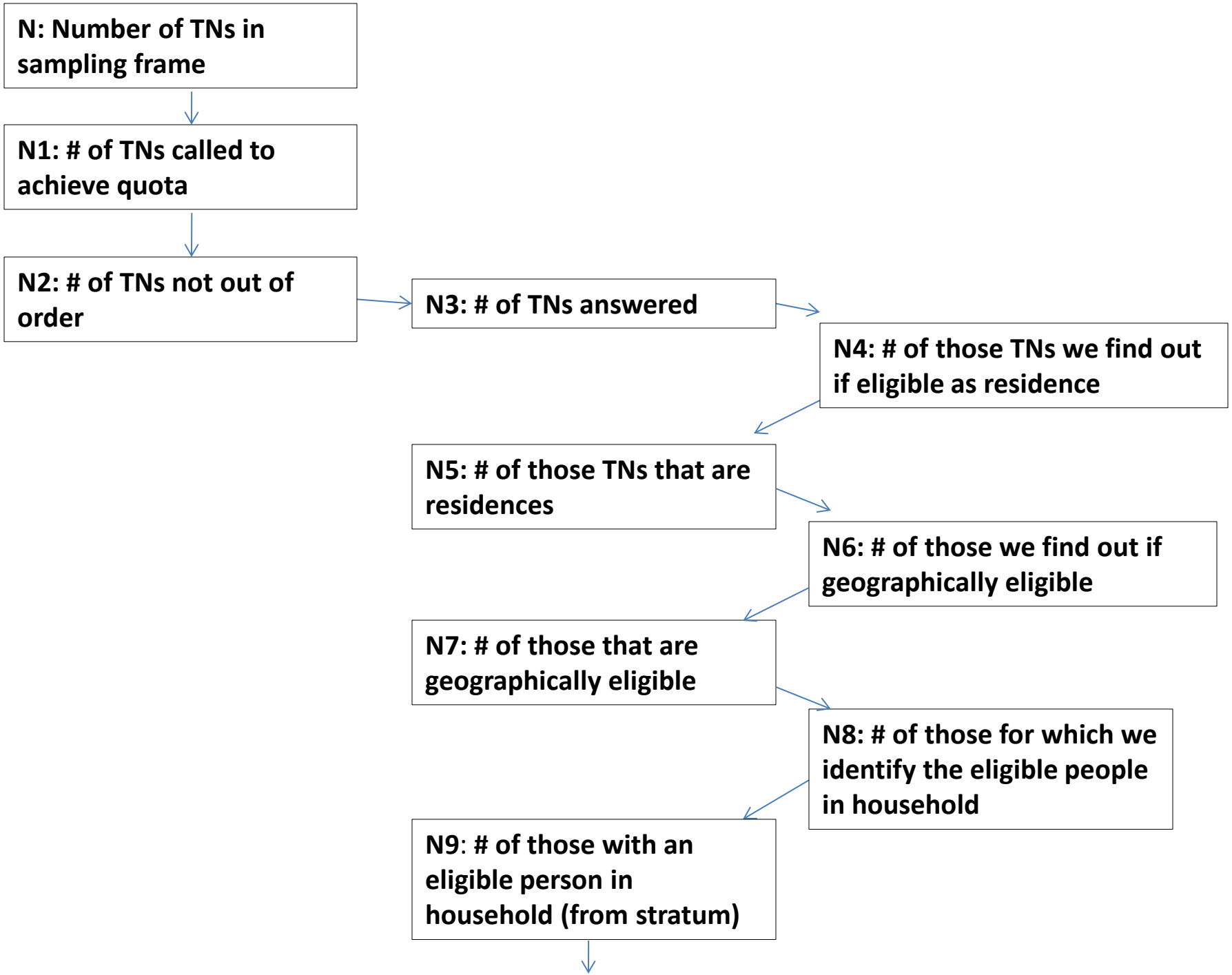
- Provided by StatsCan
- Must allow for non-response in the CLSA
- Some issues on confidentiality – information sharing

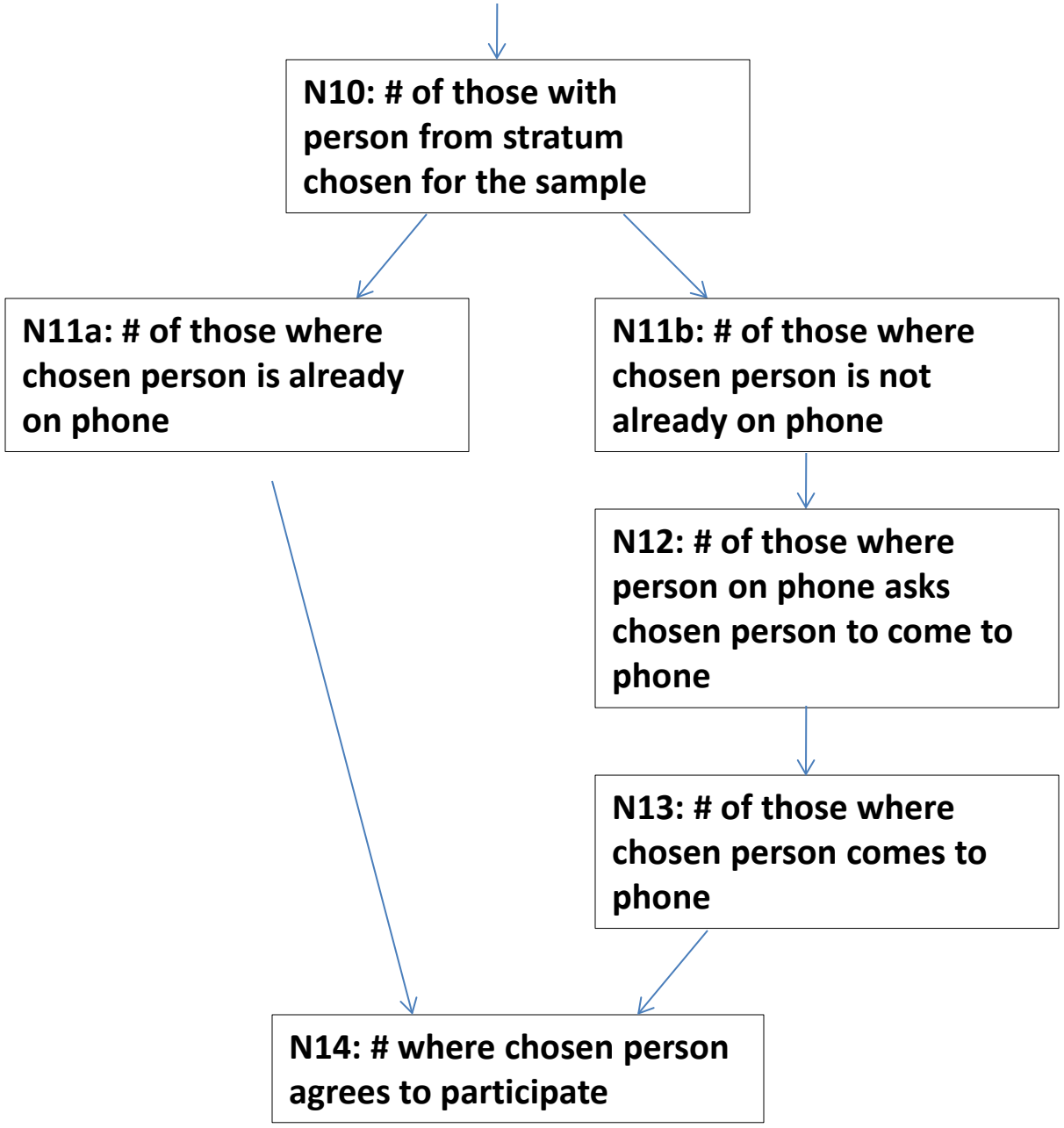
# Probabilities for the health registries

- 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 HR

# 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