



Transforming Everyday Life into Extraordinary Ideas




Application of Text Data Mining Approaches to the Canadian Longitudinal Study on Aging (CLSA) Medication Dataset: Development and Validation

Benoit Cossette, BPharm, PhD

Lauren Griffith BSc, MSc, PhD

Lisa Dolovich BScPhm, PharmD, MSc

DSECT/DSEN Seminar Series
Nov 22, 2018

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

CLSA Medication Coding Group

Benoit Cossette

Lauren Griffith

Lisa Dolovich

Patrick Emond

Dee Mangin

Jinhui Ma

Medication Coders: Helga Weigelin and Bevonie Brown

With acknowledgment of Kathryn Nicholson, Joanne Ho and Kasia Makara

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Overview

- Overview CLSA
- Challenges with medication collection
- Medication cleaning protocol development, results to date, validation, and lessons learned

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

What is the Canadian Longitudinal Study on Aging (CLSA)?

A research platform – infrastructure to enable state-of-the-art, interdisciplinary population-based *research* and *evidenced-based* decision-making that will lead to better health and quality of life for Canadians.

Canadian Longitudinal Study on Aging (CLSA)

- Strategic initiative of CIHR; on Canadian research agenda since 2001
- More than 160 researchers and collaborators – 26 institutions
- Multidisciplinary – biology, genetics, medicine, psychology, sociology, demography, economics, epidemiology, nutrition, health services
- Largest research platform of its kind in Canada for breadth and depth
- Following 50,000+ Canadians aged 45-85 at baseline for 20 years

CLSA Leads

Lead Principal Investigator
Parminder Raina (McMaster)

Co-principal Investigator
Christina Wolfson (McGill)

Co-principal Investigator
Susan Kirkland (Dalhousie)

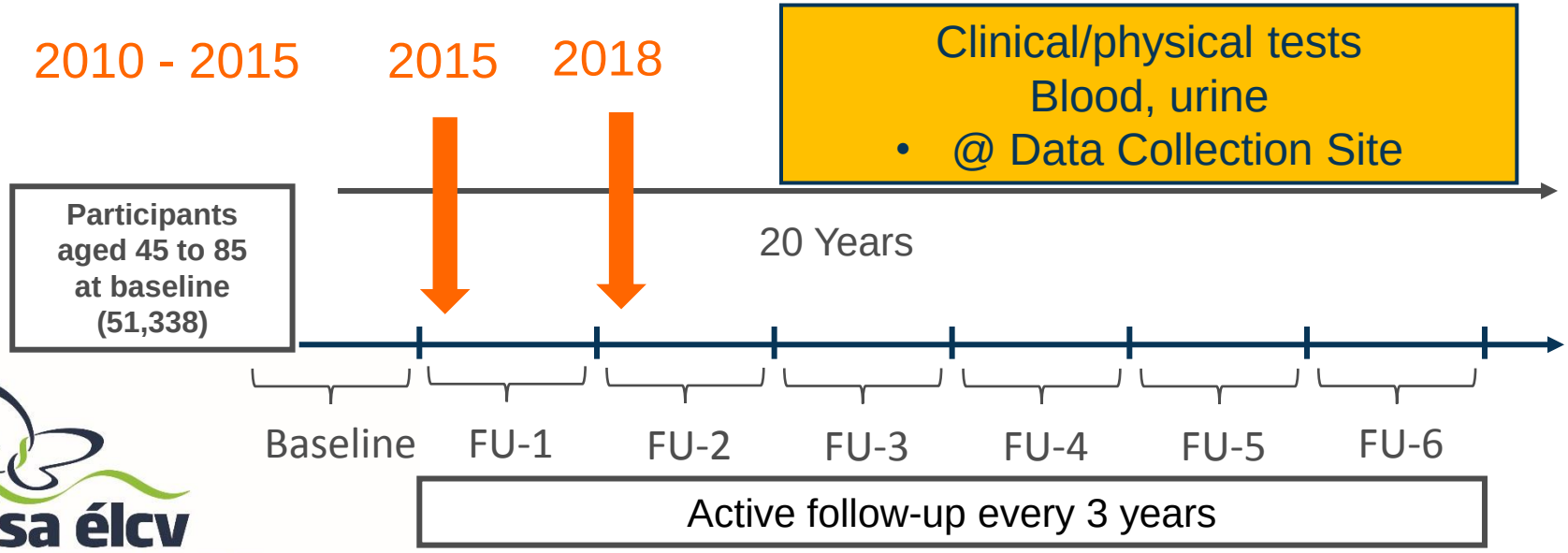
clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

CLSA Research Platform

50,000 women and men aged 45 - 85 at baseline

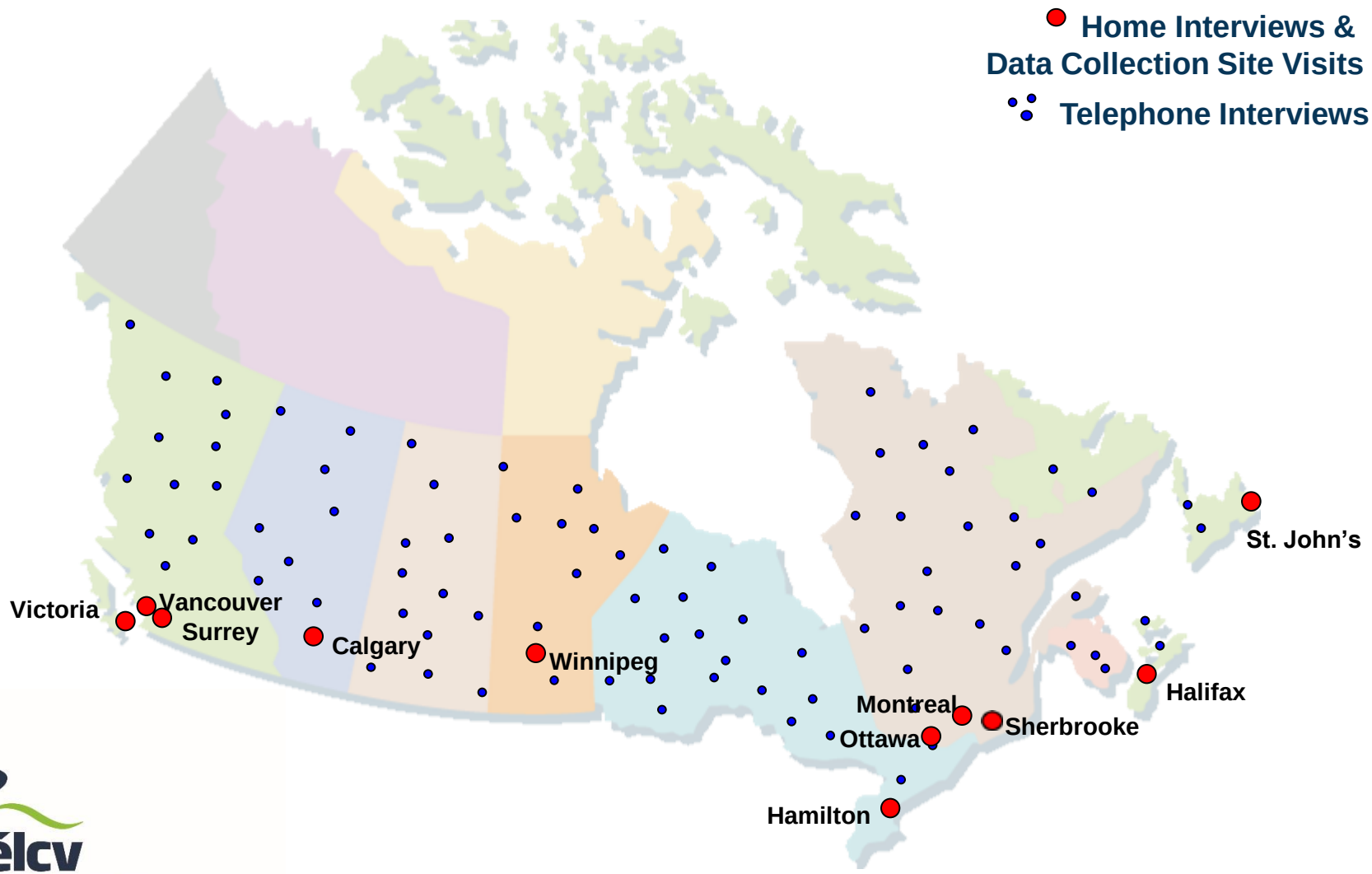
Target: 20,000 Actual: 21,241 Randomly selected within provinces	Target: 30,000 Actual: 30,097 Randomly selected within 25-50 km of 11 sites

Questionnaire • By telephone (CATI)	Questionnaire • In person, in home (CAPI)



clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

National Scope



clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Depth and Breadth of CLSA

PHYSICAL & COGNITIVE MEASUREMENTS

- Height & weight
- Waist and hip measurements
- Blood pressure
- Grip strength, timed up-and-go, chair raise, 4-m walk
- Standing balance
- Vision (retinal imaging, tonometer & visual acuity)
- Hearing (audiometer)
- Spirometry
- Body composition (DEXA)
- Bone density (DEXA)
- Aortic calcification (DEXA)
- ECG
- Carotid intima-media thickness (ultrasound)
- Cognitive assessment (30-minute battery)
- Biospecimen collection (blood and urine)

HEALTH INFORMATION

- Chronic disease symptoms (11 chronic conditions)
- Medication and supplement intake & compliance
- Women's health
- Self-reported health-care utilization
- Oral health
- Administrative data linkage health services, drugs and other administrative databases

PSYCHOSOCIAL

- Social participation
- Social networks and support
- Caregiving and care receiving
- Mood, psychological distress
- PTSD
- Injuries and consumer products
- Work-to-retirement transitions
- Personality traits
- Retirement planning
- Social inequalities
- Mobility-lifespace
- Built environments and contextual factors
- Income, wealth and assets

LIFESTYLE & SOCIODEMOGRAPHIC

- Smoking
- Alcohol consumption
- Physical activity (PASE)
- Nutrition (nutrition risk and food frequency)
- Ethnicity/race/gender
- Birth location
- Marital status
- Education

Terminology

- Tracking Cohort
  - Target - 20,000 participants from all 10 provinces, followed through Computer-Assisted Telephone Interviews (60 minutes at baseline)
  - **21,241 recruited***
- Comprehensive Cohort
  - Target - 30,000 participants living within 25 km (or 50 km) of a CLSA Data Collection Site (DCS)
  - Followed through in-home interviews (60 minutes) and physical assessments (2-3 hours) at a DCS
  - **30,097 recruited***

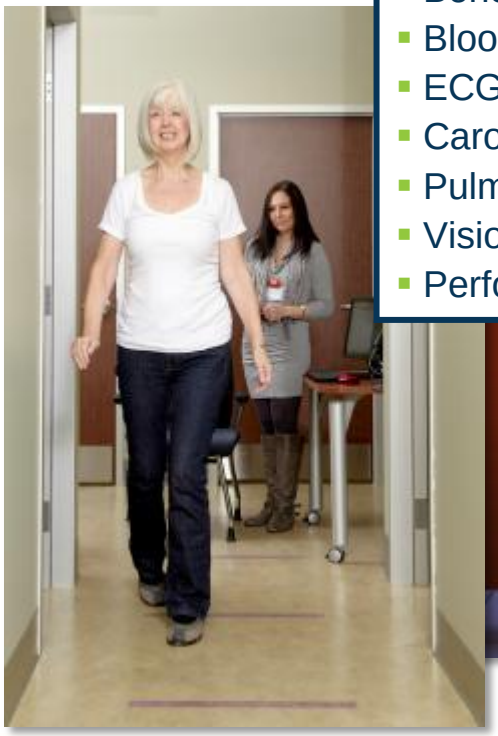
clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

CLSA Data Collection

Data Collection Site

Physical Assessments:

- Height, Weight, BMI
- Bone Density, Body Composition, Aortic Calcification
- Blood Pressure
- ECG
- Carotid Intimal-Medial Thickness
- Pulmonary Function
- Vision & Hearing
- Performance testing



Biospecimen Collection:

- Blood
- Urine

Cognitive Assessments:

- Neuropsychological Battery
  - Memory
  - Executive function
  - Reaction time

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

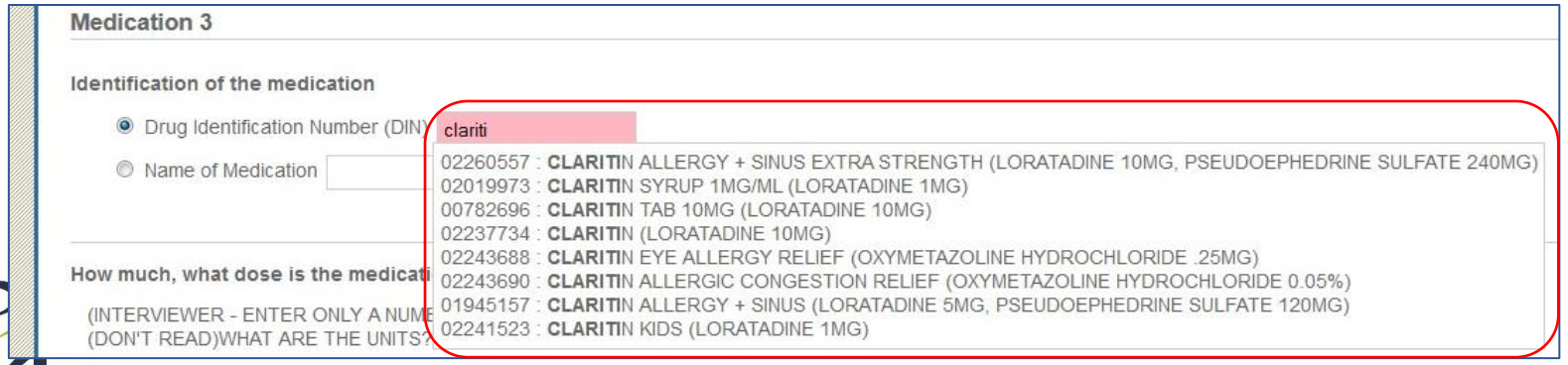
Medication Data Collection

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Entering DIN Numbers & Drug Names

Always use the DIN number field when entering a drug.

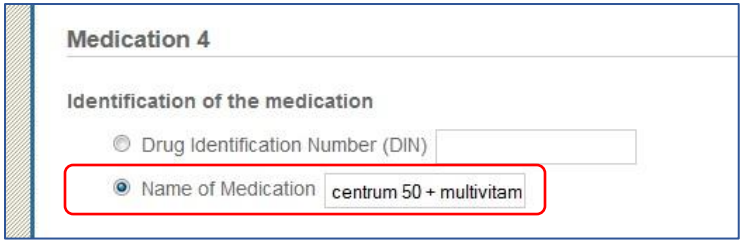
If the DIN number isn't available type the name of the medication in the DIN field, as you are typing, a list of drugs will appear, locate the correct one and click on it to select it.



Entering DIN Numbers & Drug Names

Multivitamins and herbal supplements will not have DIN's.

Enter the name of the supplement in the *"Name of Medication"* field as shown below



clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

New Question Added (Dec. 2014)

Was this medication prescribed by a doctor or is it a non-prescription medication?

- Yes
- No
- Don’t know/No answer
- Refused

NB :This questions was added to try a determine why we have so many “medication” without DINs

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Additional Information Collected

- Dosage (quantity, units)
- Frequency (e.g., once a day, once a week)
- Duration (< 6 mo, 6 mo-1 yr, >1 yr)
- Start date
- Reason(s) for use

Health Canada
Drug Product Database

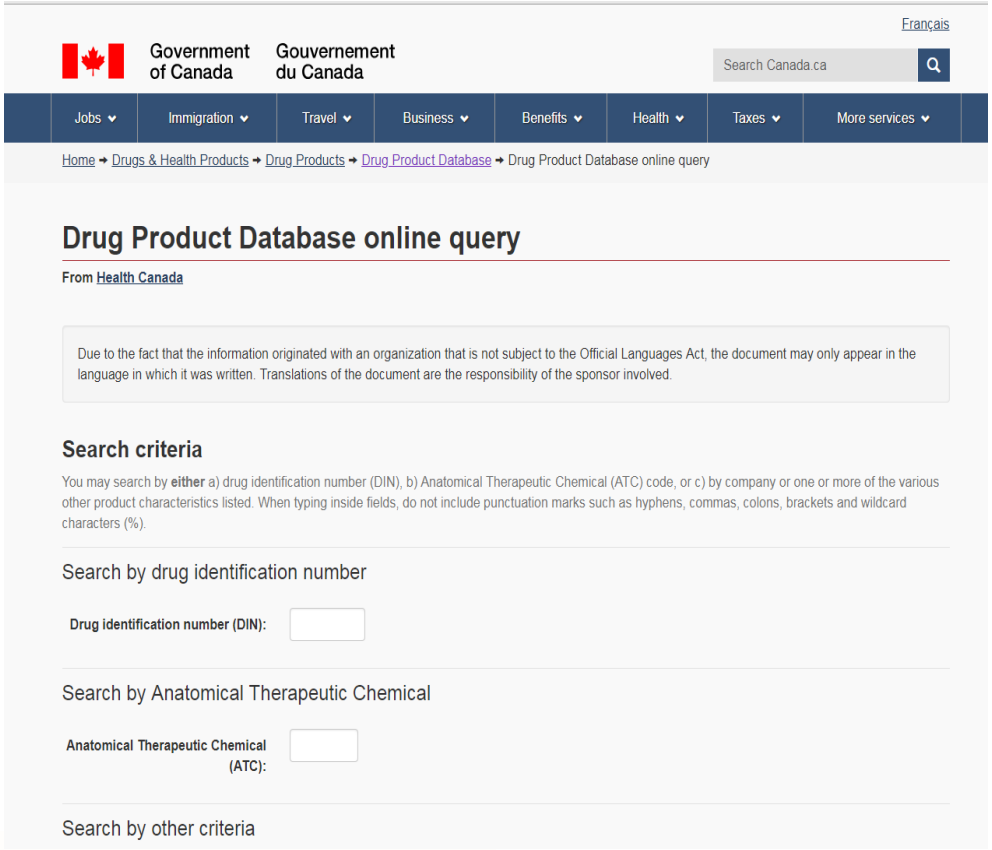
The Drug Product Database (DPD) is maintained by Health Canada.

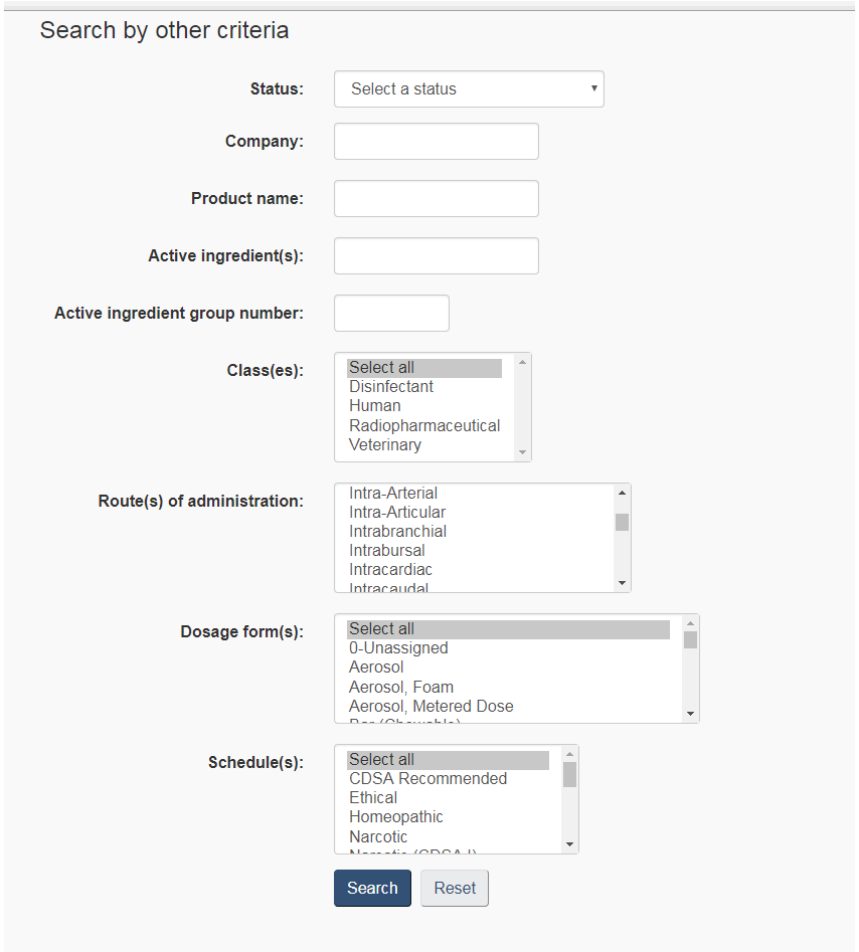
Included drugs:

- Drugs for humans and animals
- Disinfectants
- Radiopharmaceutical drugs
- Biological drugs for humans

Source: https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database/more-info.html

Health Canada DPD - online query





clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Health Canada DPD - online query

Search results summary

From Health Canada

New search

Search criteria

- Status: Marketed
- Product name: lipitor
- Class(es): Select all
- Route(s) of administration: Select all
- Dosage form(s): Select all
- Schedule(s): Select all

Search results

Filter items | Showing 1 to 4 of 4 entries | **Show** 10 **entries**

List of returned drug products

Status	DIN	Company	Product	Class	PM 1	Schedule	# 2	A.I. name 3	Strength
Marketed	02230711	PFIZER CANADA INC	LIPITOR	Human	Yes	Prescription	1	ATORVASTATIN (ATORVASTATIN CALCIUM)	10 MG
Marketed	02230713	PFIZER CANADA INC	LIPITOR	Human	Yes	Prescription	1	ATORVASTATIN (ATORVASTATIN CALCIUM)	20 MG
Marketed	02230714	PFIZER CANADA INC	LIPITOR	Human	Yes	Prescription	1	ATORVASTATIN (ATORVASTATIN CALCIUM)	40 MG
Marketed	02243097	PFIZER CANADA INC	LIPITOR	Human	Yes	Prescription	1	ATORVASTATIN (ATORVASTATIN CALCIUM)	80 MG

Health Canada DPD - extract

1	ingredient	Brand_name	TC_ATC_NUMBER	TC_AHFS_NUMBER	strength	strength_unit	Drug_identification_number
16925	CEFAZOLIN (CEFAZOLIN SODIUM)	CEFAZOLIN FOR INJECTION, USP	J01DB04	08:12.06.04	500	MG	2237137
16926	CEFAZOLIN (CEFAZOLIN SODIUM)	KEFZOL	J01DB04	08:12.06.04	1	G	322296
16927	CEFAZOLIN (CEFAZOLIN SODIUM)	KEFZOL	J01DB04	08:12.06.04	10	G	411450
16928	CEFAZOLIN (CEFAZOLIN SODIUM)	KEFZOL	J01DB04	08:12.06.04	500	MG	322288
16929	CEFAZOLIN (CEFAZOLIN SODIUM)	KEFZOL ADD-VANTAGE INJ 1GM/VIAL	J01DB04	08:12.06.04	1	G	411442
16930	CEFAZOLIN (CEFAZOLIN SODIUM)	KEFZOL ADD-VANTAGE INJ 500MG/VIAL	J01DB04	08:12.06.04	500	MG	411434
16931	CEFAZOLIN SODIUM	CEFAZOLIN FOR INJECTION USP	J01DB04	08:12.06.04	20	G	2318830
16932	CEFEPIME (CEFEPIME HYDROCHLORIDE)	CEFEPIME FOR INJECTION	J01DE01	08:12.06.16	2	G	2319039
16933	CEFEPIME (CEFEPIME HYDROCHLORIDE)	MAXIPIME FOR INJECTION -PWS IV IM 1G	J01DE01	08:12.06.16	1	G	2163632
16934	CEFEPIME (CEFEPIME HYDROCHLORIDE)	MAXIPIME FOR INJECTION-PWS IV 2GM	J01DE01	08:12.06.16	2	G	2163640
16935	CEFIXIME	AURO-CEFIXIME	J01DD08	08:12.06.12	400	MG	2432773
16936	CEFIXIME	SUPRAX	J01DD08	08:12.06.12	100	MG	868965
16937	CEFIXIME	SUPRAX	J01DD08	08:12.06.12	100	MG	2195992
16938	CEFIXIME	SUPRAX	J01DD08	08:12.06.12	200	MG	868973
16939	CEFIXIME	SUPRAX	J01DD08	08:12.06.12	400	MG	868981
16940	CEFIXIME	SUPRAX	J01DD08	08:12.06.12	400	MG	2195984
16941	CEFIXIME	SUPRAX - TAB 200MG	J01DD08	08:12.06.12	200	MG	2195976
16942	CEFOTAXIME (CEFOTAXIME SODIUM)	CEFOTAXIME SODIUM FOR INJECTION, BP	J01DD01	08:12.06.12	1.0	G	2261510
16943	CEFOTAXIME (CEFOTAXIME SODIUM)	CEFOTAXIME SODIUM FOR INJECTION, BP	J01DD01	08:12.06.12	2.0	G	2261529
16944	CEFOTAXIME (CEFOTAXIME SODIUM)	CEFOTAXIME SODIUM FOR INJECTION, BP	J01DD01	08:12.06.12	500	MG	2261502
16945	CEFOTAXIME (CEFOTAXIME SODIUM)	CLAFORAN 0.5GM/VIAL	J01DD01	08:12.06.12	500	MG	546208
16946	CEFOTAXIME (CEFOTAXIME SODIUM)	CLAFORAN 1GM/VIAL	J01DD01	08:12.06.12	1	G	546216
16947	CEFOTAXIME (CEFOTAXIME SODIUM)	CLAFORAN 2GM/VIAL	J01DD01	08:12.06.12	2	G	546224
16948	CEFOTAXIME (CEFOTAXIME SODIUM)	CLAFORAN ADD-VANTAGE VIALS 1GM/VIAL	J01DD01	08:12.06.12	1	G	1989758

Licensed Natural Health Products Database (LNHPD)

Managed by Health Canada and includes information on **licensed** natural health products, including:

- vitamin and mineral supplements
- herb and plant-based remedies
- traditional medicines like Traditional Chinese Medicines or Ayurvedic (Indian) Medicines
- omega 3 and essential fatty acids
- probiotics
- homeopathic medicines
- many everyday consumer products, like certain toothpastes, antiperspirants, shampoos, facial products and mouthwashes

Source: https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/applications-submissions/product-licensing/licensed-natural-health-products-database.html

Findings from the CLSA baseline evaluation

- 42% of medications were coded by the interviewers using a medication name / DIN proposed by the CLSA software linked to an extract from the Health Canada Drug Product Database (DPD).
- The remaining entries were entered in the '*name of medication*' field in free text.
- Based on a sample of 100 participants, 73% of the free text entries were for natural products.

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Start date / duration

Decision to drop the start date

- Start date:
  - Selected from calendar
  - 28% in January
  - 5% to 8% in February – December

- Decision to drop the start date Duration:
  - 6 months to one year
  - More than one year
  - …

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Protocol: Computer algorithm development Objectives

- To describe the process used to develop an original computer generated approach to clean medication name data that have been recorded in the CLSA database.
  - Agreement for same medication name is not the same as agreement for same DIN (DIN match).
- To conduct manual re-coding on the remaining uncleaned medication data to inform the refinement of the computer generated approach.
- To identify common medication data entry errors that can be used to support more accurate recording of medications data in the CLSA data in future stages of data collection.

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Protocol: Computer algorithm development Methods

- Computer generated approach was applied to all CLSA input entries that required cleaning. The intent of this application was to match or clean the prescribed medication or natural product number entered during a participant's interview, (an input) to one or more product(s) in the DPD or NHP.
- The approach uses multiple distinct algorithms, known hereafter as a test, for matching inputs to drugs or naturals.
- All tests are ordered sequentially such that once an input is matched it is no longer considered in the remaining tests.

Protocol: Computer algorithm development Methods

- Programming conducted using SQL (a database scripting language) and PHP (a general programming language).
- The Health Canada and the CLSA data repository (Opal) data were loaded into a MySQL database using SQL.
- Some pre-processing was conducted on these databases before using PHP to enhance performance and increased speed of matching.
  - For instance, the "simple" test compares the free text to drug names by ignoring non-alpha-numeric characters.

Results so far: Applying computer based algorithm tests

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Algorithms

- **Direct**: matches the brand or product name (not case sensitive)
- **Code**: matches 6 or more consecutive digits which matches a DIN or NPN (with or without leading 0's)
- **Word**: matches the brand or product name as a sub-string of the input
- **Simple**: matches ignoring non alpha-numeric characters
- **No parentheses**: matches ignoring any words found in (parentheses)
- **No-units**: matches ignoring any numerical units

Algorithm - direct

Direct: matches the brand or product name (not case sensitive)

Examples:

Input	Match
acetaminophen 500mg	ACETAMINOPHEN 500MG
acetaminophene	ACETAMINOPHENE
Acid Reducer	ACID REDUCER
actonel	ACTONEL
adalat	ADALAT

Algorithm - code

Code: matches 6 or more consecutive digits which matches a DIN or NPN (with or without leading 0's)

Examples:

Input	Match
00026158 Senokot Sennosides 8.6 mg	Senokot Tablets
02237556 Eur-Fer 300mg	Euro-Fer
02367335 sandoz latanoprost 1 drop	SANDOZ LATANOPROST
02371987 Mar-Atenolol 50mg	MAR-ATENOLOL
02374420 Apo-Anastrozole 1mg	APO-ANASTROZOLE

Algorithm - word

Word: matches the brand or product name as a sub-string of the input

Examples:

Input	Match
adalat XL	ADALAT
acetaminophen extra fort	ACETAMINOPHEN
Alendronate sodium (Fosamax)	ALENDRONATE
Altace HCL	ALTACE
apo metoprolol	METOPROLOL

Algorithms - simple

Simple: matches ignoring non alpha-numeric characters (including spaces)

Examples:

Input	Match
Act-Atorvastatin	ACT ATORVASTATIN
advil liqui gels	ADVIL LIQUI-GELS
apo hydro	APO-HYDRO
Apo- Hydro	APO-HYDRO
apometoprolol sr	APO-METOPROLOL SR

Algorithm – no parentheses

No parentheses: matches ignoring any words found in (parentheses)

Examples:

Input	Match
EVISTA (60MG)	evista
MYLAN-METOPROLOL (TYPE L)	mylan-metoprolol
Vitalux(areds)	Vitalux
Whey Powder (Vanilla flavour/Chocolate Flavour)	Whey powder

Algorithm – no unit

No-units: matches ignoring any numerical units

Examples:

Input	Match
APO-SALVENT 5MG/ML	apo-salvent
ASPIRIN 81MG	aspirin
calcium/magnésium 500 mg	Calcium/Magnesium
synthroid.05 mg	SYNTHROID
synthroid.75mg	SYNTHROID

		drug	%	natural	%	drug+nat	%	total	%
direct	**single**	3079	32.8%	1907	20.3%	4986	53.1%	5076	54.1%
	multiple	90	1.0%	0	0.0%	90	1.0%		
code	**single**	31	0.3%	18	0.2%	49	0.5%	51	0.5%
	multiple	2	0.0%	0	0.0%	2	0.0%		
word	**single**	1066	11.4%	441	4.7%	1507	16.1%	2106	22.4%
	multiple	282	3.0%	317	3.4%	599	6.4%		
simple	**single**	97	1.0%	84	0.9%	181	1.9%	224	2.4%
	multiple	22	0.2%	21	0.2%	43	0.5%		
no-parens	**single**	4	0.0%	4	0.0%	8	0.1%	12	0.1%
	multiple	2	0.0%	2	0.0%	4	0.0%		
no-units	**single**	114	1.2%	22	0.2%	136	1.4%	323	3.4%
	multiple	186	2.0%	1	0.0%	187	2.0%		
total matched		4669	49.8%	2788	29.7%	7457	79.5%	7457	79.5%
						total unmatched		1926	20.5%
						total inputs		9383	

direct: Matches the brand or product name (not case sensitive)

code: Matches 6 or more consecutive digits which matches a DIN or NPN (with or without leading 0's)

word: Matches the brand or product name as a sub-string of the input

simple: Matches ignoring non alpha-numeric characters

no-parens: Matches ignoring any words found in (parentheses)

no-units: Matches ignoring any numerical units

single: The input matches a single DIN or NPN

multiple: The input matches more than one DIN or NPN

Spelling errors

- Acetaminophen
  - Acetamin**a**phen
  - Acetom**e**n**o**phen
- Actonel
  - Act**a**nel
  - Act**a**n**o**l
  - Acton**a**l
- Timulol (timolol)
- Exium (Nexium)
- Tamsulosin hydrochlorathiazide (hydrochloride)
- Apothyazide
  - Apo-triazide (triamterene / hydrochlorothiazide)

  OR
  - Apo-hydro (Hydrochlorothiazide)

Examples of manual coding decision rules to code for DRUG NAME– based on iterative coding exercises (not all shown)

Coding Issue	Manual Code
EXACT MATCH (ie has DIN there, exact product name, if dosage available this might help)	code exact DIN from DPD/NHPD
BRAND NAME but no dosage	code DIN from DPD that is the lowest number DIN
Generic NAME but multiple manufacturers	code DIN from DPD that is the lowest number DIN
Incorrect Spelling	Put in correct spelling and then choose lowest number DIN
Drug class do not code further	Drug class do not code further
Drug class – code further	Drug class – code further
Not a drug	Not a drug
Not enough information	Not enough information

Study Two: Validation (in progress)

- To determine the extent of agreement between the original and refined computer generated approach (test) and the manual re-coding (gold standard) for cleaning the medication data in the CLSA database.
- To assess the extent of agreement when comparing the original computer generated approach (test) with the refined computer generated approach (gold standard), as well as when the order of the algorithms are applied in different sequences.
- To assess whether the extent of agreement changes when re-coding prescribed medication vs. natural products or when re-coding among participants with specific chronic diseases.

Sample Size

- Based on 1) the presence of free text medication data; and 2) the presence of at least one prescribed medication or natural product within the alpha-numeric data.

	Agreement	Precision	Confidence Interval	Sample Size
Best Scenario	0.95	0.02	0.95	457
Worst Scenario	0.85	0.03	0.95	545

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Lessons learned (so far)

- Manual entry by interviews derives many varied data entry issues (despite training)
- Many entries do not match to DPD or NHPD and so manual coding will always be needed
  - Many international products
  - Complex combination products, esp NHPD, vitamins
  - Formatting such as parentheses, brackets cannot always be dealt with by computer
  - Nondrug products
- If more than 1 manual coder they all need to use the exact same drug/NHP database
- Will always be behind with new products unless DPD/NHPD can be accessed in real time (versus updated in the back end every 3-6 months)

Conclusions

- The process of developing and validating the algorithm is challenging
- A well working computer algorithm will be a great asset to CLSA for retrospective and prospective medication data coding
- A clean CLSA medications database will be a great asset to Canadian researchers

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Extra Slides re Data Access

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Data Access

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

CLSA Data Access

- Designed as a research study, funded as a **research platform**
- Data **available** to researchers and trainees based in academic settings and research institutes in Canada and elsewhere
- **2019 application deadlines:**
  - February 25
  - June 3
  - September 23

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

CLSA Data Access Guiding Principles

- **The rights, privacy and consent of participants** must be protected
- **Confidentiality and security of data and biospecimens** must be safeguarded
- CLSA data and biospecimens are resources to be used optimally to **support research to benefit all Canadians**

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Data Access Timeline

Data and Sample Access Committee & Scientific Management Team Review

Dataset Preparation & Release

≤ 12 weeks

Submission

Final Report

6-9 weeks

5-7 days

Administrative & Statistical Review

Access Agreement Signed; Ethics Approval

clsa elcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

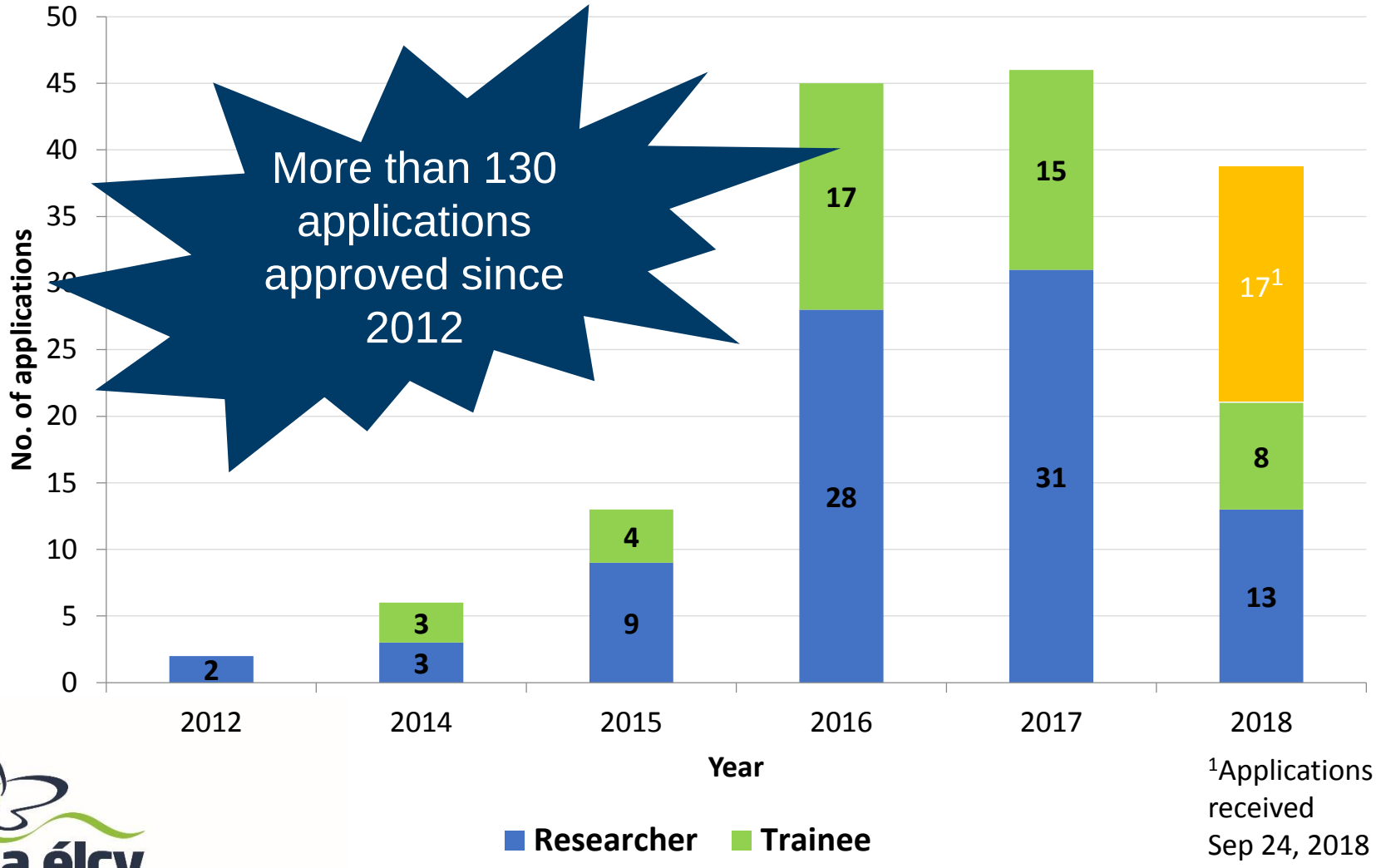
Applicants advised to plan on receiving data 6 months after submission deadline

Data Access Fees

- **Partial Cost Recovery Model**
  - **Alphanumeric data**
    - **$3,000** for a straightforward alphanumeric dataset
    - **Graduate students** using data solely for thesis research & **Postdoctoral fellows** using data solely for the postdoctoral project are eligible for a **fee-waiver**. Trainees must be enrolled at a Canadian institution or be supported by Canadian funds if working outside Canada.
  - **Images & raw data**
    - Additional fees of $1,000 per application are associated with the request for images & raw data.

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Approved Applications



[1]Applications received Sep 24, 2018

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Approved Projects Keywords

Resources: www.clsa-elcv.ca

Font Size + - | Language en fr | Search

Update your Contact Details | Contact Us

clsa élcv

Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

About Us | Participants | Partners | Researchers | Data Access | Stay Informed

Participants

- Update your contact details
- Protecting your privacy
- Find out how the CLSA platform is being used

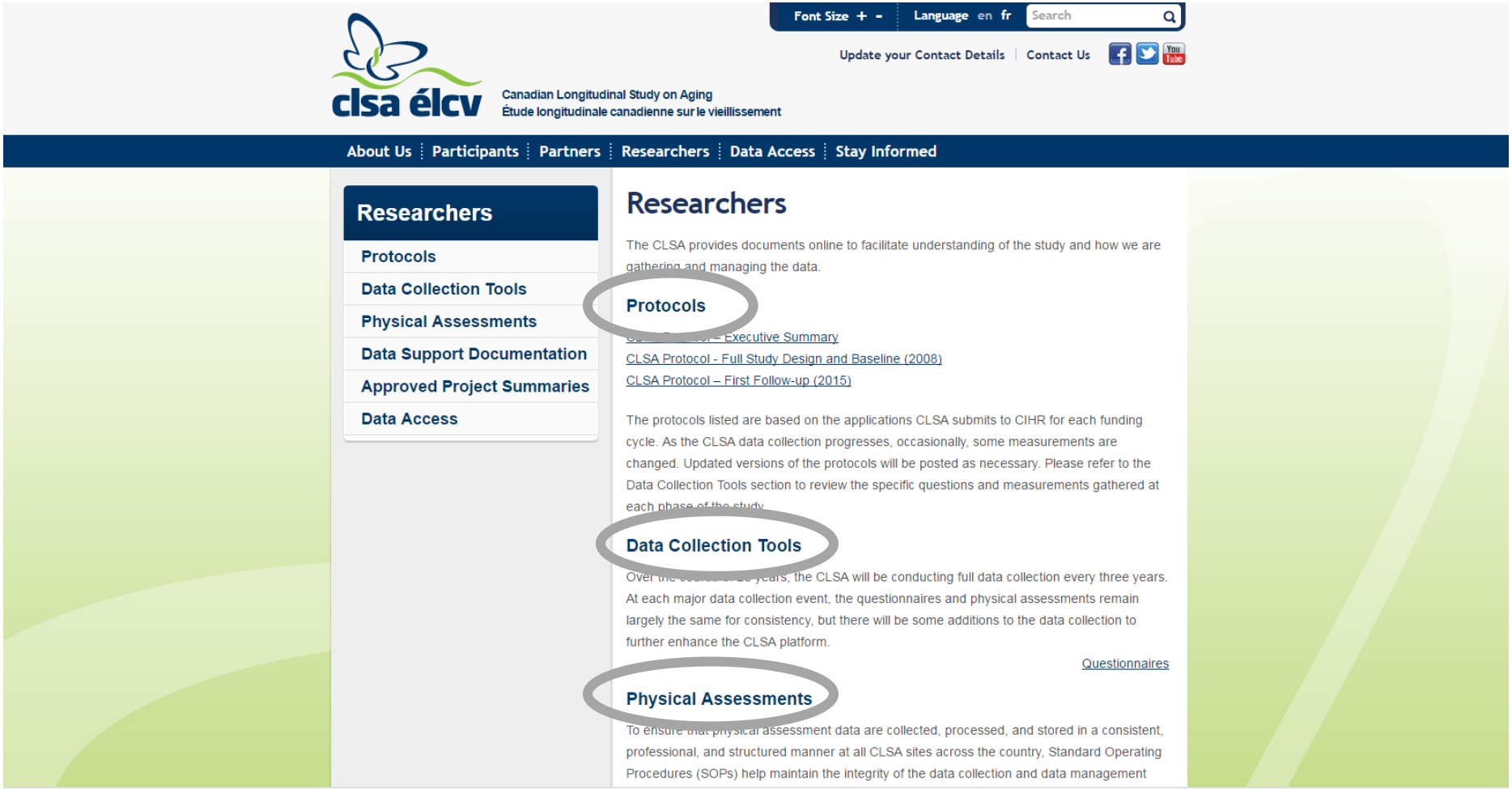
Researchers

- Spring 2016 data release
- DataPreview Portal
- Approved Projects

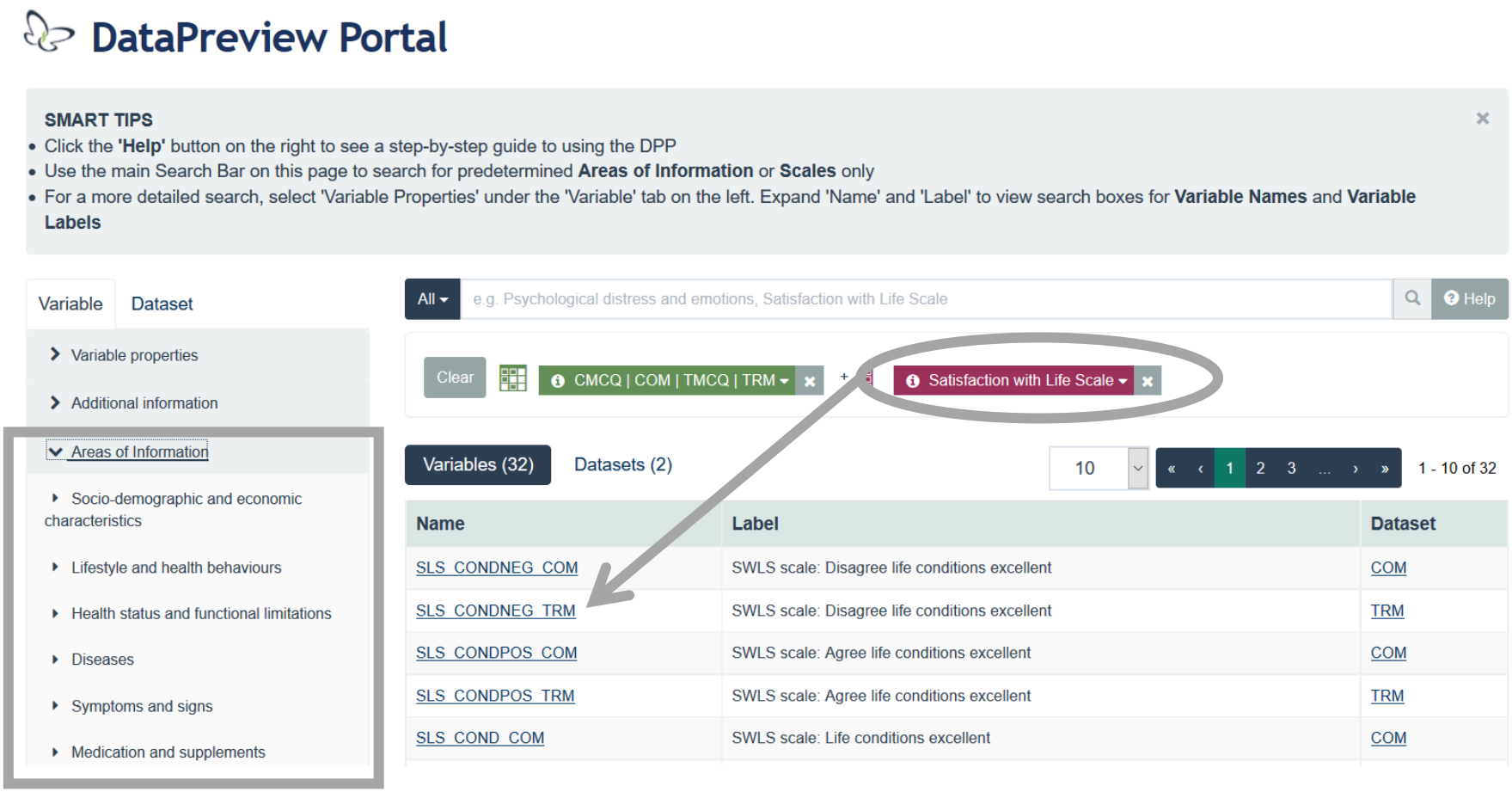
Partners

- Partners & Supporters
- Partnering with the CLSA
- Collaborate and Innovate

Information for Researchers & Trainees



Data Preview Portal



clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

FAQs

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement