CIVIC
DIGITAL
FELLOWSHIP

HIV/AIDS Review-Verify-Extract Source Tool (HARVEST)

Internal Tool & ETL Pipeline for HIV/AIDS Dashboard Visualizing Country Reports, PEPFAR, and UNAIDS Data

Ari Israel

Supervised by Derek Azar & Timothy Fowler

Population Division - Health Studies Branch

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

Goals

- → Visualize HIV/AIDS pandemic in a dashboard similar to JHU COVID-19 map
- → Emphasize confirmed data made public by country governments
- → Collate public data from UNAIDS and PEPFAR in addition to country reports
- → Identify relevant sources to extract and verify statistics from annual reports
- → Monitor sources for updated reports or tables to include in database

Challenges

→ Limited universe of confirmed HIV/AIDS pandemic data
  - → Most countries don’t publish annual surveillance reports
  - → Most data consistently available (i.e. UNAIDS) is modeled
  - → Lots of broken MOH sites/dead links for developing countries

→ Available country-reported data in unstructured format i.e. PDFs

→ Identifying relevant sources and extracting statistics not easily automated

→ Extracted statistics need to be manually verified by second analyst

Solutions

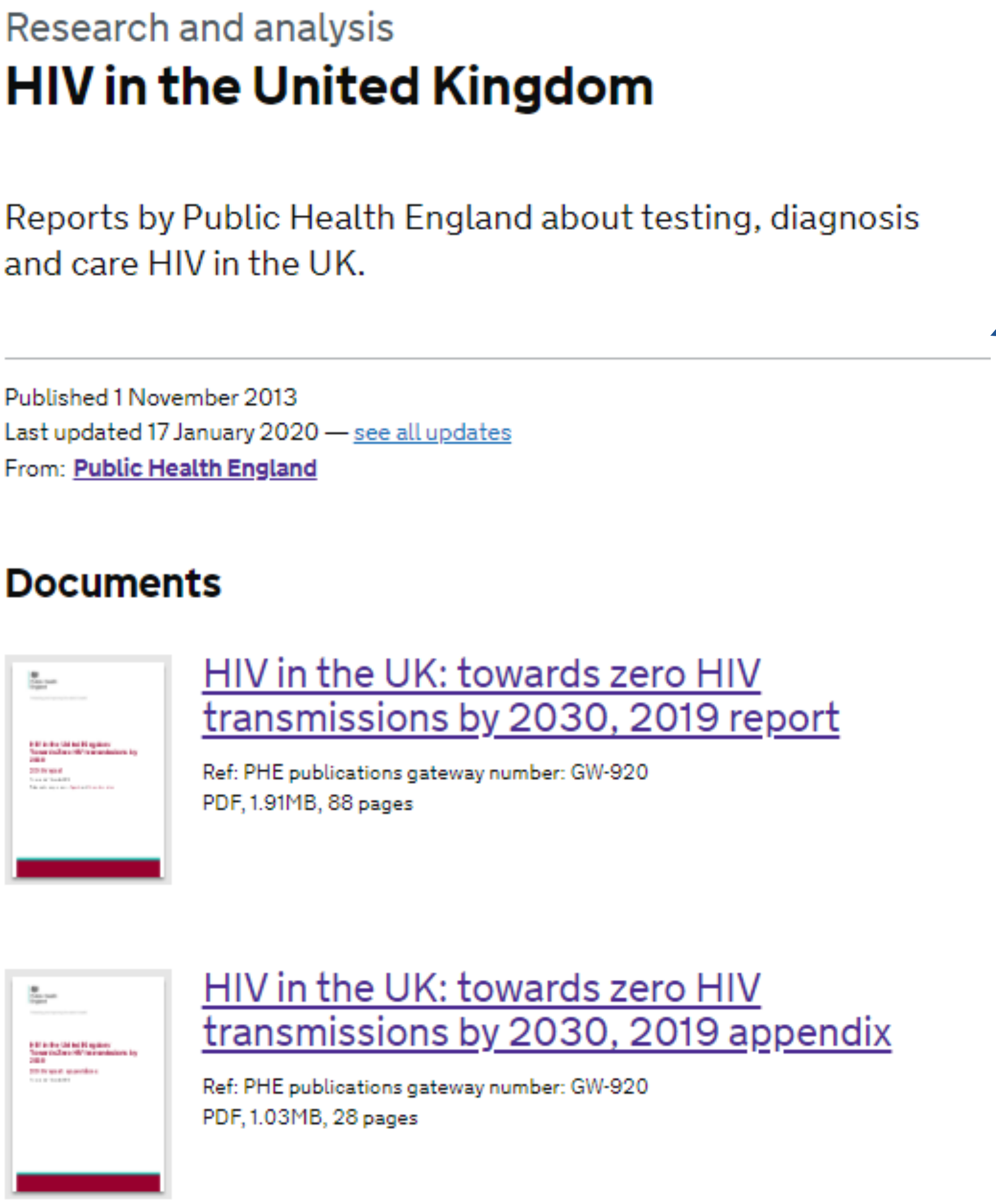
→ Automated crawl/scrape of manually-identified sites with annual reports

→ Internal tool for analysts to review sources, extract and verify statistics

→ Save data collected via tool to database with ability to export master table

Technologies

- → **Anaconda** for environment/package management
- → **GitLab** for source control and wiki docs
- → **SQLite** to store prototype data
- → **Scrapy** to crawl and scrape sites
- → **Requests** to fetch data via APIs
- → **Python** for backend and Django
- → **Django** for web app framework and ORM
- → **JavaScript** for review interface frontend
- → **HTML/CSS** for Django templates and styling
- → **PDF.js** for interactive PDFs in review interface



Research and analysis

HIV in the United Kingdom

Reports by Public Health England about testing, diagnosis and care HIV in the UK.

Published 1 November 2013
Last updated 17 January 2020 — see all updates
From: **Public Health England**

Documents

HIV in the UK: towards zero HIV transmissions by 2030, 2019 report

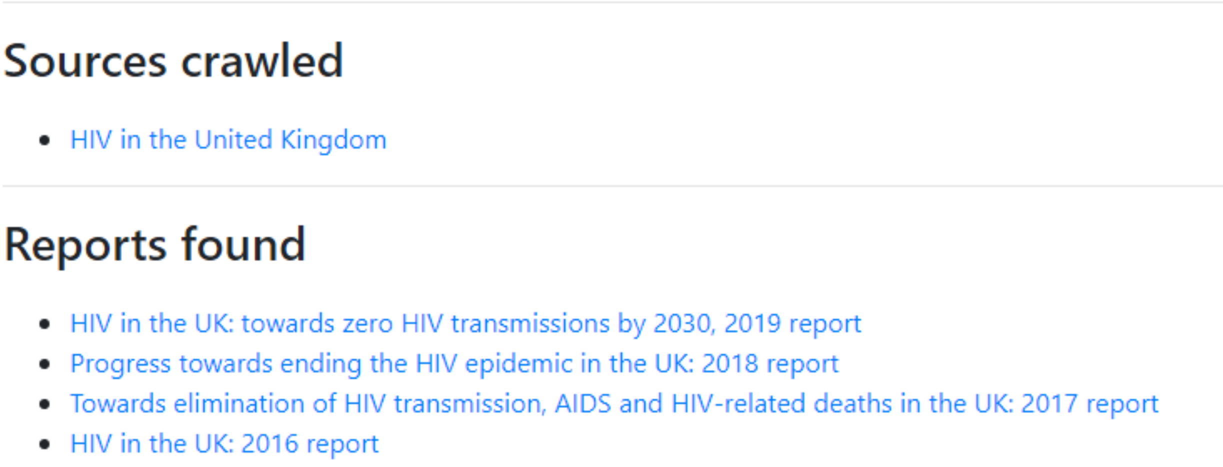
Ref: PHE publications gateway number: GW-920
PDF, 1.91MB, 88 pages

HIV in the UK: towards zero HIV transmissions by 2030, 2019 appendix

Ref: PHE publications gateway number: GW-920
PDF, 1.03MB, 28 pages

Scrapy used to crawl and scrape UK government site with list of reports

Results shown in HARVEST **Django** app



Sources crawled

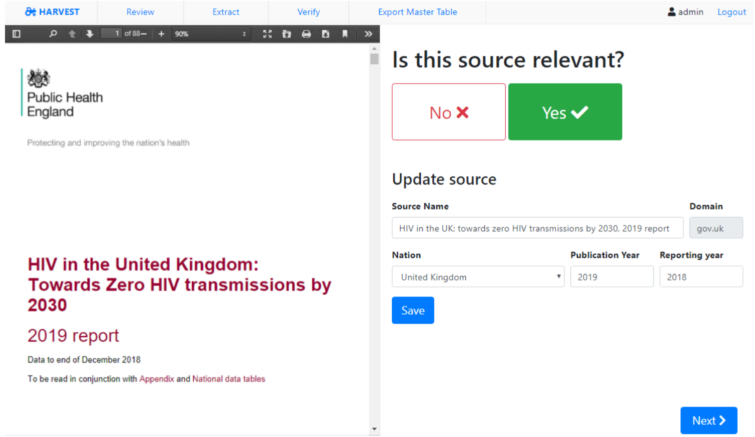
- HIV in the United Kingdom

Reports found

- HIV in the UK: towards zero HIV transmissions by 2030, 2019 report
- Progress towards ending the HIV epidemic in the UK: 2018 report
- Towards elimination of HIV transmission, AIDS and HIV-related deaths in the UK: 2017 report
- HIV in the UK: 2016 report

Workflow

1. **Review** source for relevance → 2. **Extract** statistics from source → 3. **Verify** extracted statistics



Public Health England

Protecting and improving the nation's health

HIV in the United Kingdom: Towards Zero HIV transmissions by 2030

2019 report

Data to end of December 2018

To be read in conjunction with Appendix and National data tables

Is this source relevant?

Update source

Source Name: HIV in the UK: towards zero HIV transmissions by 2030, 2019 report

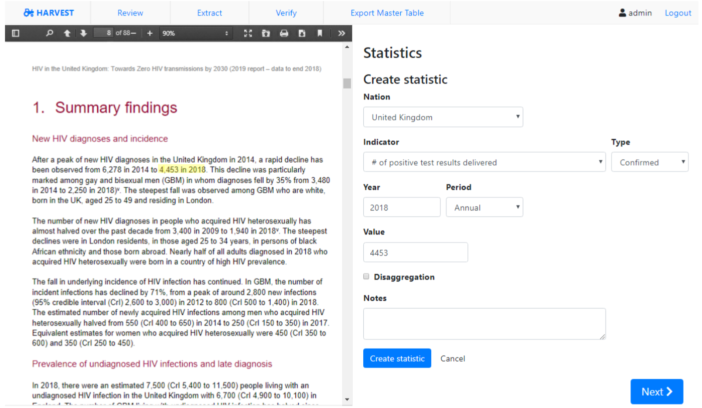
Domain: gov.uk

Nation: United Kingdom

Publication Year: 2019

Reporting year: 2018

Analyst 1



HIV in the United Kingdom: Towards Zero HIV transmissions by 2030 (2019 report – data to end 2018)

1. Summary findings

New HIV diagnoses and incidence

After a peak of new HIV diagnoses in the United Kingdom in 2014, a rapid decline has been observed from 6,278 in 2014 to 4,453 in 2018. This decline was particularly marked among gay and bisexual men (GBM) in whom diagnoses fell by 35% from 3,480 in 2014 to 2,250 in 2018)[v]. The steepest fall was observed among GBM who are white, born in the UK, aged 25 to 49 and residing in London.

The number of new HIV diagnoses in people who acquired HIV heterosexually has almost halved over the past decade from 3,400 in 2009 to 1,940 in 2018[v]. The steepest declines were in London residents, in those aged 25 to 34 years, in persons of black African ethnicity and those born abroad. Nearly half of all adults diagnosed in 2018 who acquired HIV heterosexually were born in a country of high HIV prevalence.

The fall in underlying incidence of HIV infection has continued. In GBM, the number of incident infections has declined by 71%, from a peak of around 2,800 new infections (95% credible interval (CrI) 2,600 to 3,000) in 2012 to 800 (CrI 500 to 1,400) in 2018. The estimated number of newly acquired HIV infections among men who acquired HIV heterosexually halved from 550 (CrI 400 to 650) in 2014 to 250 (CrI 150 to 350) in 2017. Equivalent estimates for women who acquired HIV heterosexually were 450 (CrI 350 to 600) and 350 (CrI 250 to 450).

Prevalence of undiagnosed HIV infections and late diagnosis

In 2018, there were an estimated 7,500 (CrI 5,400 to 11,500) people living with an undiagnosed HIV infection in the United Kingdom with 6,700 (CrI 4,900 to 10,100) in

Statistics

Create statistic

Nation: United Kingdom

Indicator: # of positive test results delivered

Type: Confirmed

Year: 2018

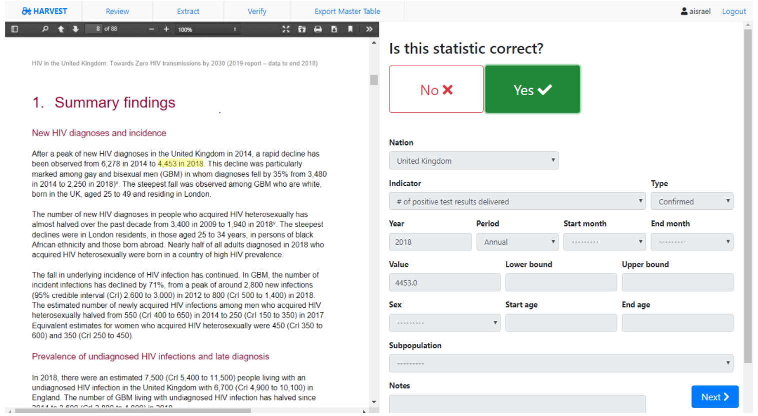
Period: Annual

Value: 4453

□ Disaggregation

Notes:

Analyst 1



HIV in the United Kingdom: Towards Zero HIV transmissions by 2030 (2019 report – data to end 2018)

1. Summary findings

New HIV diagnoses and incidence

After a peak of new HIV diagnoses in the United Kingdom in 2014, a rapid decline has been observed from 6,278 in 2014 to 4,453 in 2018. This decline was particularly marked among gay and bisexual men (GBM) in whom diagnoses fell by 35% from 3,480 in 2014 to 2,250 in 2018)[v]. The steepest fall was observed among GBM who are white, born in the UK, aged 25 to 49 and residing in London.

The number of new HIV diagnoses in people who acquired HIV heterosexually has almost halved over the past decade from 3,400 in 2009 to 1,940 in 2018[v]. The steepest declines were in London residents, in those aged 25 to 34 years, in persons of black African ethnicity and those born abroad. Nearly half of all adults diagnosed in 2018 who acquired HIV heterosexually were born in a country of high HIV prevalence.

The fall in underlying incidence of HIV infection has continued. In GBM, the number of incident infections has declined by 71%, from a peak of around 2,800 new infections (95% credible interval (CrI) 2,600 to 3,000) in 2012 to 800 (CrI 500 to 1,400) in 2018. The estimated number of newly acquired HIV infections among men who acquired HIV heterosexually halved from 550 (CrI 400 to 650) in 2014 to 250 (CrI 150 to 350) in 2017. Equivalent estimates for women who acquired HIV heterosexually were 450 (CrI 350 to 600) and 350 (CrI 250 to 450).

Prevalence of undiagnosed HIV infections and late diagnosis

In 2018, there were an estimated 7,500 (CrI 5,400 to 11,500) people living with an undiagnosed HIV infection in the United Kingdom with 6,700 (CrI 4,900 to 10,100) in England. The number of GBM living with undiagnosed HIV infection has halved since

Is this statistic correct?

Nation: United Kingdom

Indicator: # of positive test results delivered

Type: Confirmed

Year: 2018

Period: Annual

Start month: ---------

End month: ---------

Value: 4453.0

Lower bound:

Upper bound:

Sex: ---------

Start age:

End age:

Subpopulation: ---------

Notes:

Analyst 2

Shape your future
START HERE >

United States® Census 2020

Deliverables

- ✓ **System design and app documentation**
- ✓ **HARVEST internal tool prototype**

Next step: productionize software
Document business logic and language-agnostic architecture for potential DSD rewrite in .NET