Over a Century of Infectious Disease Surveillance in Canada: Introducing the Canadian Disease Incidence Dataset (CANDID)

David J. D. Earn[1,3] Gabrielle MacKinnon[4] Samara Manzin[5]
Michael Roswell[2,6] Steve Cygu[7] Chyunfung Shi[2]
Benjamin M. Bolker[1,2] Jonathan Dushoff[2,3] Steven C. Walker[1]

[1]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada, L8S 4L8;

[2]Department of Biology, McMaster University, Hamilton, Ontario, Canada, L8S 4L8;

[3]M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada, L8S 4L8;

[4]Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada;

[5]Department of Biology, McGill University, Montreal, Quebec, Canada;

[6]Department of Biology, University of Maryland, College Park, Maryland, USA;

[7]African Population Health and Research Center;

Friday 20th December, 2024

Abstract

Background: Canadian notifiable disease surveillance provides data on the incidence of communicable diseases, dating back to the late 19th century. The Public Health Agency of Canada offers summaries of these data (from 1924–2022) through an online portal. These summaries provide historical context for Canadian health researchers, but lack information on intra-annual and inter-provincial patterns. Sub-annual and sub-national data can be found in published documents, or requested from archives of government agencies, but are only available in typewritten or handwritten hard copies. We digitized and collated these data sources to create a resource for epidemiology and public health.

Methods: We manually entered data from scans of hard copies into spreadsheets resembling the originals, facilitating accurate transcription through easier cross-checking. We developed open-source pipelines to harmonize these spreadsheets into CSV files that blend data across sources.

Results: We assembled and processed 1,631,380 incidence values from 1903–2021. Focusing on sub-annual and sub-national data and removing redundancy yielded 934,009 weekly, monthly, or quarterly incidence values broken down by province/territory, containing 139 diseases. We give two examples of sub-annual and sub-national patterns: strong annual cycles of poliomyelitis that peaked simultaneously across provinces, and spatially heterogeneous resurgence of whooping cough in the 1990s.

Interpretation: Canada's history of infectious disease surveillance has produced a detailed record of sub-annual and sub-national disease incidence patterns that remains largely unexplored. This important record is now available as the Canadian Disease Incidence Dataset (CANDID), hosted on a publicly accessible website along with the pipelines used to create it and scans of the original sources.

1 Introduction

Learning from data on past communicable disease outbreaks and epidemics is an important component of public health planning [1]. 2024 marks the 100th anniversary since the Canadian federal government began collecting such data through notifiable disease surveillance programmes [2]. Several provinces conducted surveillance before 1924, including Ontario, going back at least as far as 1903. The Public Health Agency of Canada (PHAC) provides summaries of these data as annual and national totals from 1924–2022 through an online portal (`https://diseases.canada.ca/notifiable`). These annual and national data are useful for understanding broad trends, but provide no information on seasonal patterns of incidence within years or on spatial patterns across provinces. For example, annual data cannot be used to estimate the timing and shape of an outbreak curve, and national data cannot be used to assess whether provinces displayed different patterns of spread in the vaccine era.

The national notification system has at times published more informative weekly incidence data, broken down by disease and province/territory, but these data are not available through the portal. It has generally been prohibitively challenging for researchers to find and access these sub-annual and sub-national data. Still, a great deal of this historical surveillance information can be found in government publications (either as hard copy reference material or on-line), as well as by directly contacting government agencies. Here, we introduce CANDID (Canadian Disease Incidence Dataset), a curated dataset that integrates and cleans these sources to create a comprehensive and accessible digital record of Canadian notifiable disease incidence.

These CANDID data have the potential to drive many studies by the broader research community. Our goal here is to announce their existence on a publicly available web site (Appendix A.2) and illustrate their value. Two examples illustrate the advantages of the sub-annual and sub-national data provided by CANDID. First, we describe how poliomyelitis incidence was strongly and consistently seasonal from 1933 to 1963, and that the yearly peaks were synchronous across provinces. Second, the well-studied apparent resurgence of whooping cough in the 1990s showed significant regional variation. While the territories, prairie provinces, and Québec experienced clear increases in incidence, this pattern was not evident in British Columbia, Ontario, and the Atlantic region. We use a simple graphical approach in these examples, illustrating how these data can inform fundamental questions in epidemiology. Formal statistical analyses that dig deeper into specific questions will follow in subsequent publications.

2 Methods

2.1 Data sources

We began searching for published and unpublished Canadian historical infectious disease notification data in 2000. We focused on sub-annual and sub-national data collected through

the surveillance programs described in Appendix A.1, particularly incidence over entire provinces and territories, since finer spatial resolution data were rare. We compiled provincial and territorial population data [3–5] to compute comparable incidence rates. We acquired data in any of three formats: (1) paper hard copies, (2) digitally produced PDF files, or (3) spreadsheets (including CSV files). We scanned all of the hard copies. We found that Optical Character Recognition (OCR) was unable to convert scans into digital spreadsheets with sufficient accuracy, and so we entered the data manually (§2.2). We used PDF extraction tools [6] to avoid manual entry of digitally-produced PDF pages.

2.2 Data entry

We manually entered the information in scans into replica Excel spreadsheets (i.e., digital spreadsheets in which the layout of each spreadsheet matches the original), to facilitate comparing reproductions with their sources (Figure 1). The Ontario Ministry of Health data were only partially entered as replicas, as they predate our systematic effort and are the sole exception. Reading scans and interpreting handwritten sources (e.g., Figure 2) was often a slow, and occasionally error-prone, process. We took an iterative approach to resolving unclear numbers by guessing and then refining our guesses if necessary when quality checks (§2.5) revealed discrepancies. Scripted data preparation pipelines (§2.3) facilitated updates during this process. Where available in these sources, we also entered reported annual and national data to facilitate quality control (§2.5).

2.3 Data preparation

We developed open-source [7] pipelines that convert replica spreadsheets (Figure 1) into CSV files, using common variables to combine data from different sources (details in Appendix A.3). The focal variable is the number of new cases of a specific disease reported in a specific location over a specific time period. CANDID consists of three CSV files, each corresponding to a stage (Figure A6) in the data processing pipeline, ranging from a minimally processed file offering maximum flexibility in data preparation to a heavily processed file prioritizing convenience.

The ***unharmonized*** file preserves the raw, digitized data, giving researchers the freedom to apply their own data preparation methods. Descriptors in the unharmonized file use original names (e.g., "infantile paralysis" for poliomyelitis before 1924) to minimize historical information loss [8]. The ***harmonized*** file removes low quality data, aggregates some municipal data to provincial levels, and adds harmonized location and disease descriptors that simplify the combination of data from different sources (e.g., poliomyelitis whenever infantile paralysis is reported). Harmonized CSV files are convenient for querying diseases, provinces, and time periods, but they cannot be used directly for analysis due to overlapping incidence values. In data science terms, the harmonized data are not ***normalized*** [9, 10]. For example, the harmonized data include total polio incidence alongside separate values for polio with and without paralysis (see Appendix A.3.3 for more examples). Such overlapping data are useful for quality control (§2.5), but data analysis requires removing overlaps to prevent

No.	Disease (For rare diseases, see page 9)		CANADA					
			Report Week – Semaine du rapport	Previous Week – Semaine précédente	Median 1960-1964 Médiane	Cumulative Total – Total cumulatif		
						1965	1964	Median – Médiane 1960-1964
1	Brucellosis (Undulant fever)	(044)	—	1	1	—	2	1
2	Diarrhoea of the newborn, epidemic	(764)	2	2	1	2	1	1
3	Diphtheria	(055)	—	—	1	—	—	1
4	Dysentery:		16	17r	39	16	55	39
5	(a) Amoebic	(046)	—	—	—	—	—	—
6	(b) Bacillary	(045)	10	17r	21	10	26	21
7	(c) Other and unspecified	(048)	6	—	18	6	29	18
8	Encephalitis, infectious	(082.0)	—	—	—	—	—	—
9	Food poisoning:		14	11	11	14	20	11
10	(a) Staphylococcus intoxication	(049.0)	—	7	—	—	—	—
11	(b) Salmonella with food as vehicle of infection.	(042.1)	2	4	11	2	20	11
12	(c) Unspecified	(049.2)	12	—	—	12	—	—
13	Hepatitis, infectious (including serum hepatitis).	(092,N998.5)	118	132	140	118	111	140

	A	B	C	D	E	F	G	H
3			CANADA					
4						Cumulative Total		
5	No.	Disease (For rare diseases, see page 9)	Report Week - Semaine du rapport	Previous Week - Semaine précédente	Median 1960-1964 Médiane	1965	1964	Median - Médiane 1960-1964
6	1	Brucellosis (Undulant fever) (044)	-	1	1	-	2	1
7	2	Diarrhoea of the newborn, epidemic (764)	2	2	1	2	1	1
8	3	Diphtheria (055)	-	-	1	-	-	1
9	4	Dysentery	16	17r	39	16	55	39
10	5	(a) Amoebic (046)	-	-	-	-	-	-
11	6	(b) Bacillary (045)	10	17r	21	10	26	21
12	7	(c) Other and unspecified (048)	6	-	18	6	29	18
13	8	Encephalitis, infectious [1] (082.0)	-	-	-	-	-	-
14	9	Food poisoning:	14	11	11	14	20	11
15	10	(a) Staphylococcus intoxication (049.0)	-	7	-	-	-	-
16	11	(b) Salmonella with food as vehicle of infection. (042.1)	2	4	11	2	20	11
17	12	(c) Unspecified (049.2)	12	-	-	12	-	-
18	13	Hepatitis, infectious (including serum hepatitis) (092,N988.5)	118	132	140	118	111	140

Figure 1: Example of a typewritten source document prepared using a typewriter. The top panel shows a scan of part of this source document, and the bottom panel shows our replica in Microsoft Excel of this same part.

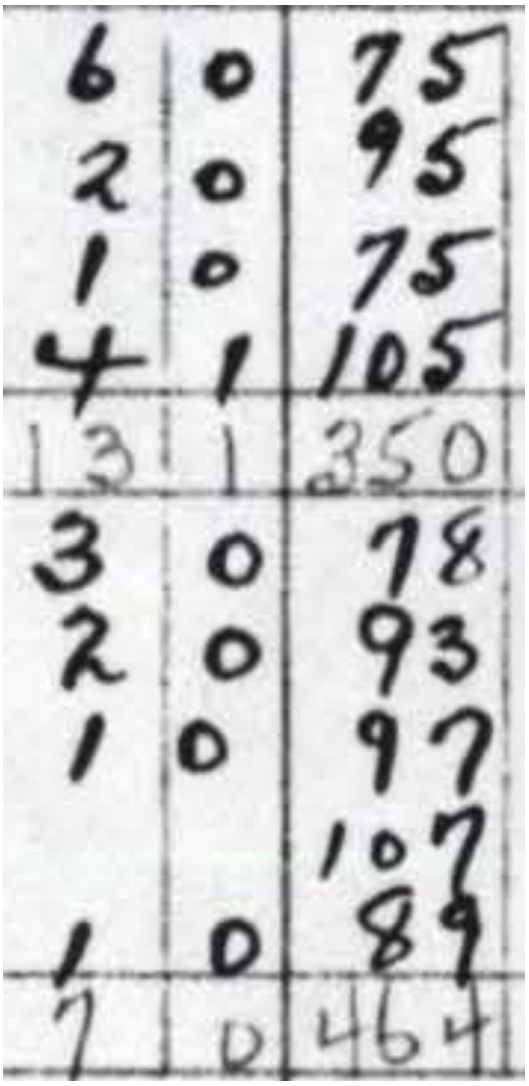

Figure 2a: Easy to read.

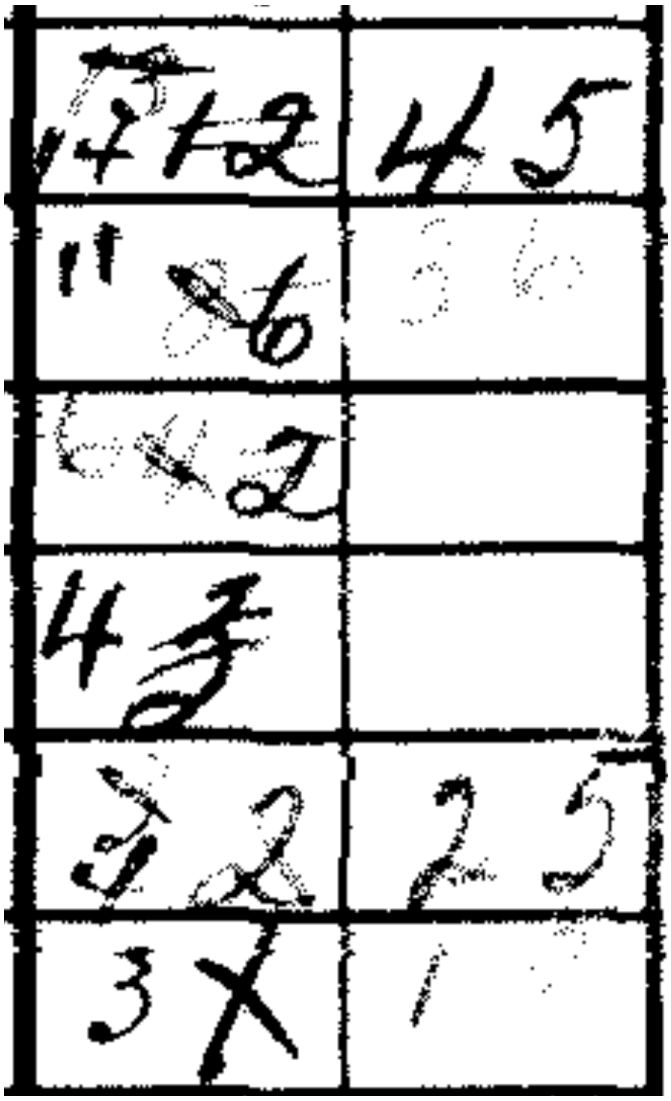

Figure 2b: Difficult to read.

Figure 2: Examples of handwritten data. Most handwritten hard copies, such as the 1939 erysipelas and gonococcal data from the Ontario Ministry of Health (left), are easy to read. Others, like the 1955 poliomyelitis data from Statistics Canada (right), are difficult to read, posing challenges for digitization.

double-counting cases. The normalized file does not contain overlapping incidence values, enabling aggregation without double-counting. For convenience when computing incidence rates, we joined provincial population sizes to the normalized data. All figures in this paper are based on the normalized file. Researchers who wish to make different harmonization or normalization choices can use the upstream files (see Appendix A.3.3 for details).

2.4 Data provenance

Each record can be traced back to the relevant original scan, replica spreadsheet, and/or script used to produce it, using information in the CSV files (Appendix A.3.4). We follow research data management practices by distributing DataCite [11] (version 4.3) metadata with each dataset in the archive. These metadata will make it easier to deposit future versions of CANDID into a research data repository, which we plan to do (§4.3).

2.5 Quality control

We compared sums of incidence values with marginal totals reported both in CANDID data sources and in the PHAC portal [2]. Discrepancies suggest possible data-entry or scripting errors. We investigated discrepancies and fixed those that appeared to be due to digitization error. These investigations were simplified using our open data provenance tools (§2.4) and digitized data source replicas (Figure 1). More detail is in Appendix A.4.

3 Results

As of December 2024 CANDID is based on 13 sources (Table 1). Appendix A.2 provides information on how to access these 1,631,380 unharmonized, 1,186,827 harmonized, and 934,009 normalized incidence values.

Figure 3 lists the 139 diseases that appear in the normalized dataset, and highlights the time periods in which weekly, monthly, or quarterly incidence data were found in each location. The case numbers for many of these diseases are aggregated from 315 "sub-diseases" (Appendix A.3.2), harmonizing 929 unique historical name variants. The stratification of each disease into sub-diseases varied across sources (details in Appendix A.3.3).

3.1 Examples

3.1.1 Poliomyelitis cases peaked at the same time each year across all provinces

Poliomyelitis incidence displayed a strongly seasonal pattern (upper panel Figure 4), with the peak occurring consistently between week 31 and 40 of each year between 1933 and 1963, after which cycles disappeared. These annual cycles were synchronous across provinces, as each province experienced peak incidence around the same time (provincial panels in Figure 4). Detecting this pattern of spatial synchrony requires incidence data at both sub-annual and sub-national scales.

Years	Provinces	Frequency	Organization	Received As
1903-1939	Ontario	monthly	The Sanitary Journal	Hard copy
1910, 1921-1927	Saskatchewan	monthly	Saskatchewan Bureau of Public Health	Hard copy
1915-1925	Quebec	monthly	Quebec Ministry of Health and Social Services	Hard copy
1924-1955	All	weekly, monthly	Statistics Canada	Hard copy (with handwriting)
1930-1947	Ontario	monthly	Ontario Department of Health	Hard copy
1939-1989	Ontario	weekly	Ontario Ministry of Health	Hard copy (with handwriting)
1956-1978	All	weekly	Statistics Canada	Hard copy
1979-1989	All	4-weekly	Statistics Canada	Hard copy
1990-2001	All	monthly, quarterly	Health Canada	Hard copy
1990-2021	Ontario	weekly	Public Health Ontario	Spreadsheet
2001-2006	All	quarterly	Canada Communicable Disease Report	PDF
2004-2017	Manitoba	monthly	Public Health Manitoba	PDF
2004-2019	Alberta	weekly	Alberta Health	Spreadsheet

Table 1: Data sources. The Frequency column gives the shortest period over which incidence counts were reported for all diseases and locations in the source (if not all disease-location combinations have the shortest period, multiple frequencies are given). Sources that include handwritten data are indicated in the Received As column.

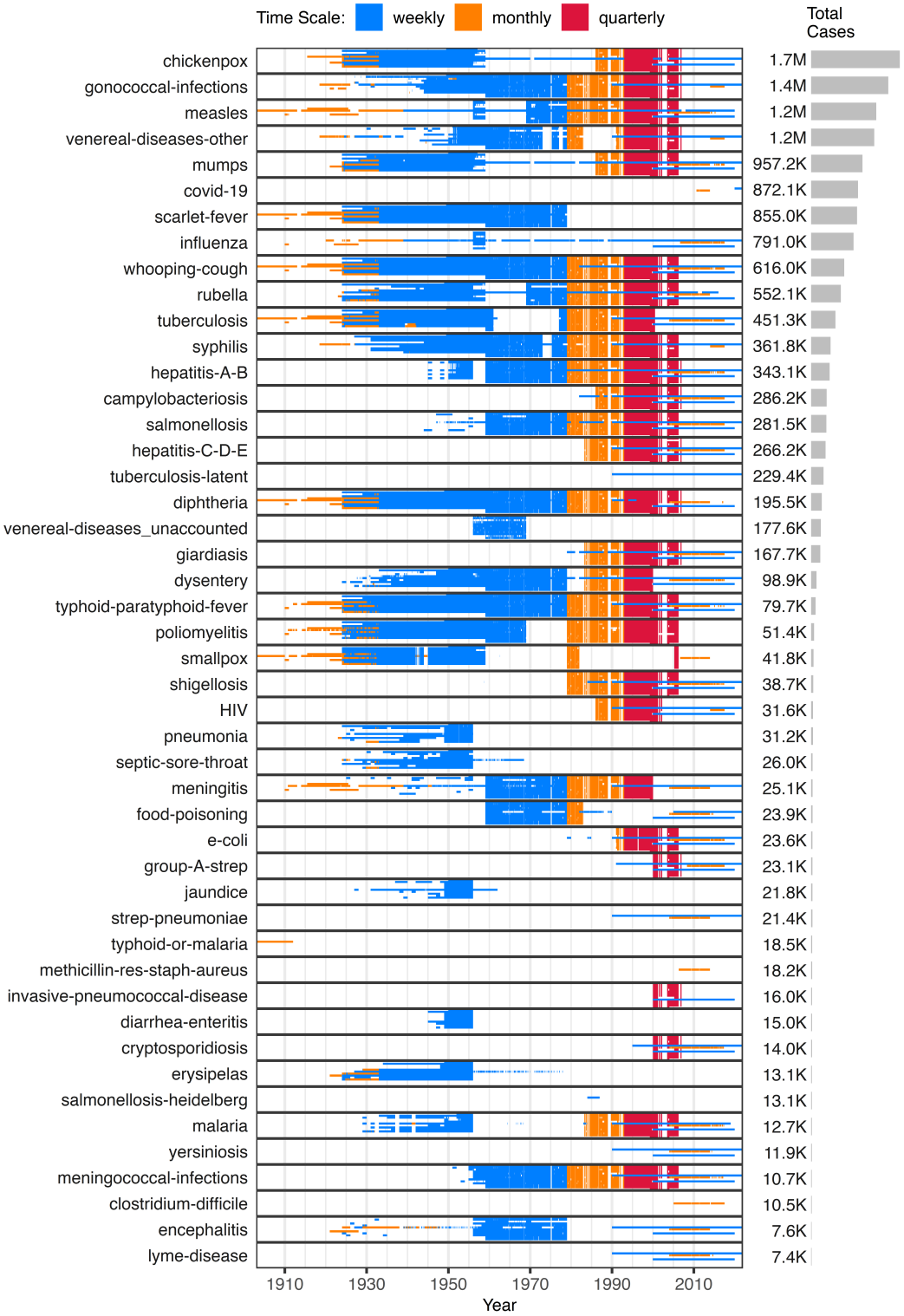

Figure 3a: Data availability for highly-reported diseases (top 47 by total reported cases). The remaining diseases and the full figure caption are below.

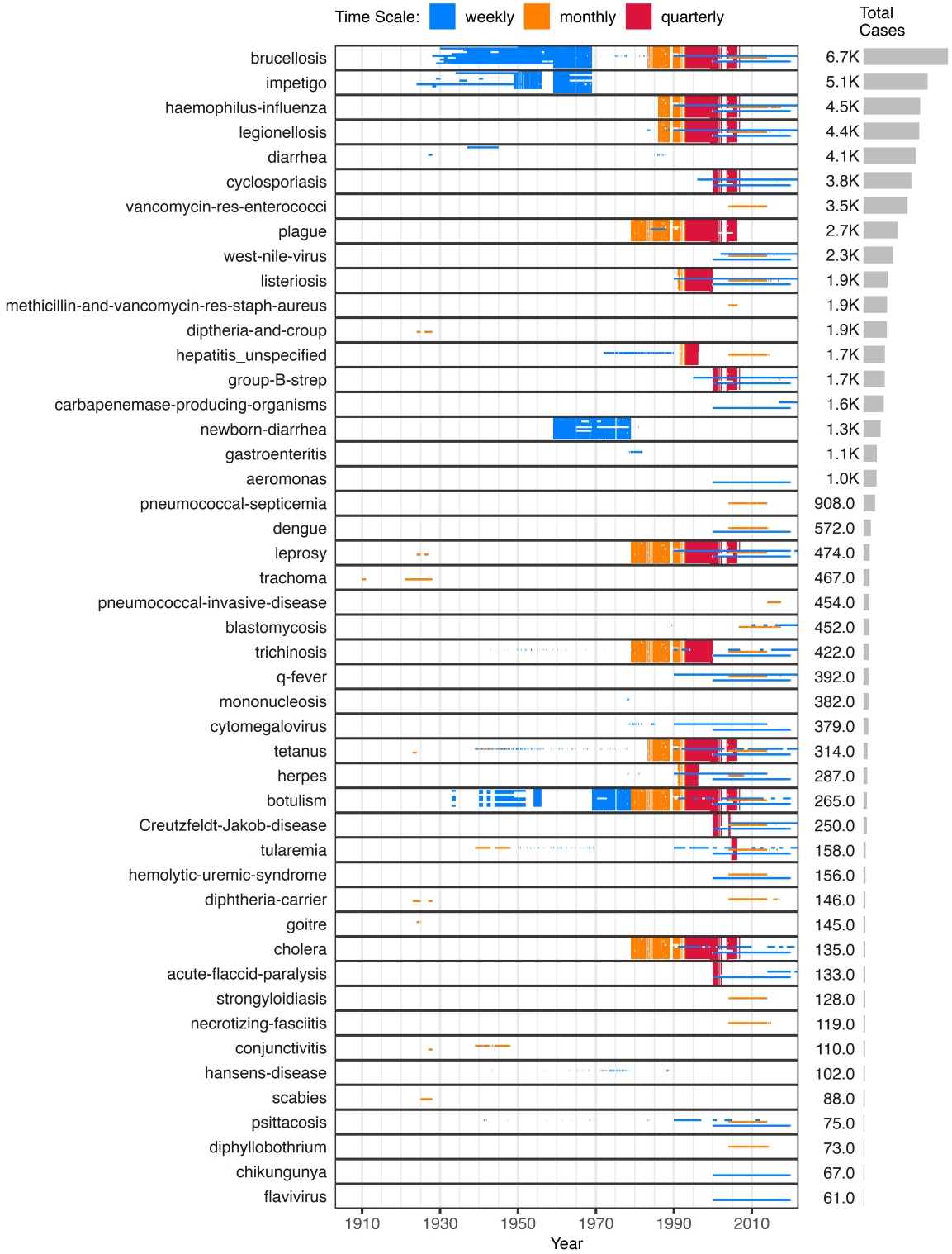

Figure 3b: Data availability for moderately-reported diseases (middle 47 by total reported cases). The remaining rarely reported diseases and the full figure caption are below.

Figure 3c: Data availability for rarely reported diseases (bottom 45 by total reported cases). The full figure caption is below.

Figure 3: Data availability of all diseases in the normalized dataset. The diseases are ranked by the total number of cases (right panels), summed over all provinces for which data were obtained. The diseases are ordered with the largest number of cases at the top. Each incidence value, including zeros, is shown as a tiny coloured rectangle. The y-axis labels identify the disease, while the rectangle's length along the x-axis represents the temporal extent. Colours indicate time scale, with shorter periods in blue (short wavelength) and longer periods in longer wavelengths. White spaces represent missing data (see Appendix A.3.1 for details on the varied reasons for missing data). The vertical position within each disease denotes the province/territory, arranged roughly clockwise: NL-NS-PE-NB-QC-ON-MB-SK-AB-BC-YT-NT-NU. Data from provincial sources (QC, ON, MB, SK, AB) during periods where national sources (Table 1) are unavailable results in visually prominent horizontal lines.

3.1.2 Regional differences in whooping cough incidence

Aggregating provincial whooping cough data to the national level (Figure 5) reveals a pattern consistent with previous analyses that lacked provincial data [12]. One feature of this pattern is an apparent resurgence of whooping cough in the 1990s (highlighted in Figure 5). We find that this much discussed resurgence (e.g., [12]) was not uniformly expressed across the country (Figure 5 bottom six panels), and is clearly apparent only in the territories, the prairies, and Québec. Throughout the 1990s, yearly incidence peaked at 21 cases per $100,000$ in Ontario, but in the territories the peak was 293. The original study [12] did not have the sub-national data required to explore these regional differences.

4 Interpretation

4.1 Brief summary

We digitized and collated sub-annual (weekly, monthly, quarterly) and sub-national (provincial, territorial) Canadian communicable disease incidence counts from 13 sources (1903–2021), covering 139 diseases and 315 sub-diseases. Two examples illustrate patterns detectable only with such granular data: (1) annual poliomyelitis incidence cycles peaked consistently across provinces and territories each year (Figure 4); (2) the resurgence of whooping cough in the 1990s showed regional differences (Figure 5). The cleaned data, the processing pipeline that produced them, scans of the original source documents, and Excel spreadsheet replicas of those scans are all publicly available online, ensuring full transparency and accessibility (Appendix A.2).

4.2 Explanation

CANDID complements existing Canadian notification data. The Public Health Agency of Canada (PHAC) provides an online portal [2] (`https://diseases.canada.ca/notifiable`) with annual, national incidence counts often used in retrospective analyses (e.g., [12, 14–

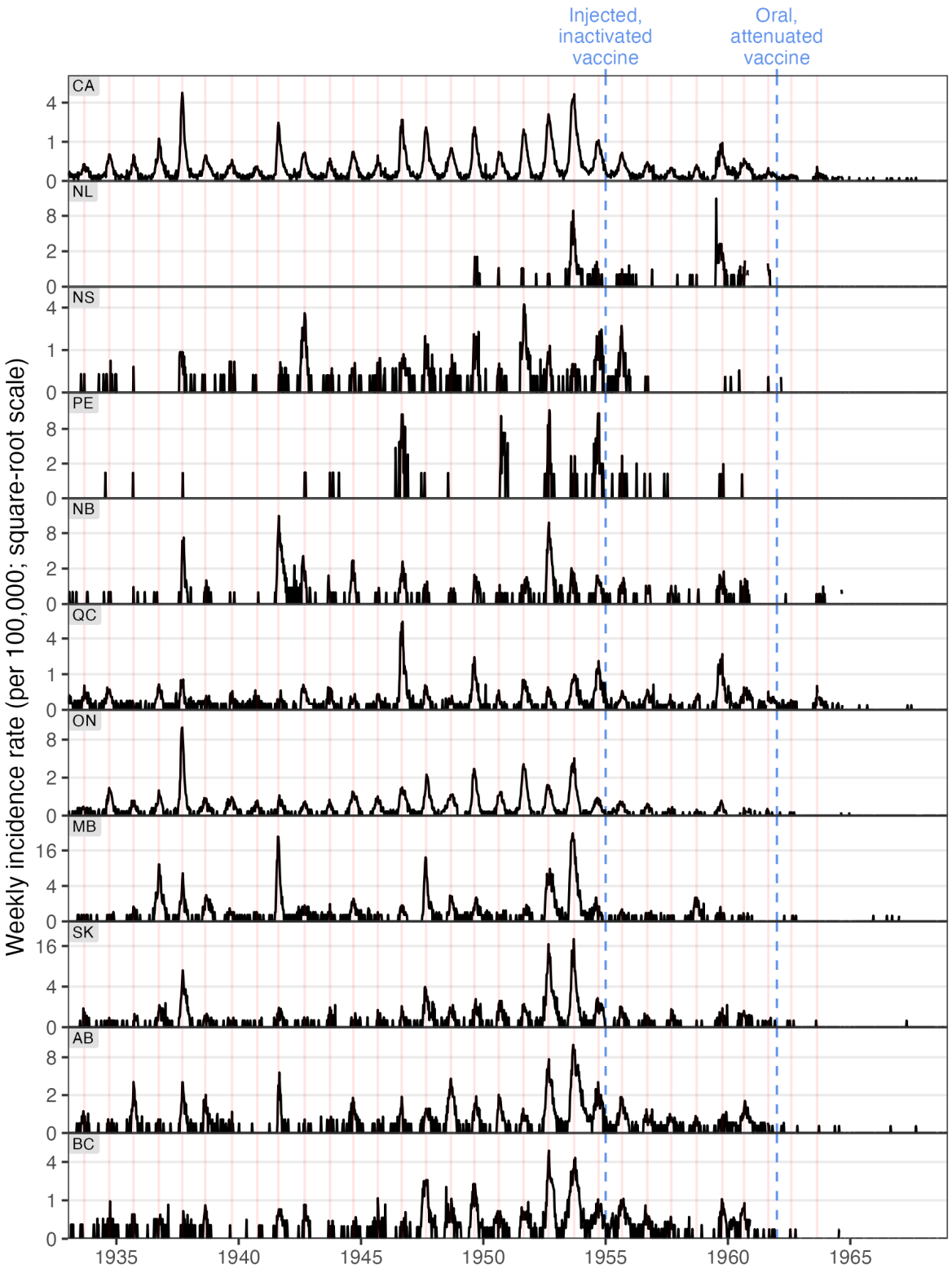

Figure 4: Caption on next page.

Figure 4: Weekly poliomyelitis incidence from 1933 to 1968 (square root scale). The vertical lines do not indicate the start of each year but mark the week of peak national incidence (top panel) in years with more than 20 cases. Provincial peaks closely align with these national peaks, indicating strong spatial synchrony. This pattern could not have been detected with annual and national data. The introduction of two important vaccination programmes are shown as blue vertical dashed lines. Methods for producing this plot are described in Appendix A.5.

18]). However, these data lack the detail needed to study outbreak patterns, seasonality, or geographic variation. Our sub-annual, sub-national data enable research on intra-annual and inter-provincial patterns in Canada and facilitate comparisons with U.S. data [19]. While Public Health Ontario provides online monthly data since 2012 [20], our archive includes weekly Ontario data (1990–2021) and extends back before 1924, including Ontario (1903), Saskatchewan (1910), and Québec (1915). By consolidating federal and provincial sources into a standardized format, we simplify integration of new data, enabling researchers to focus on analysis rather than curation.

Our project parallels Project Tycho [21], which curated weekly U.S. incidence data and whose impact is summarized by [19]. Unlike Tycho, we open-sourced our data preparation pipelines to enable community-driven quality improvements. These pipelines include scans of original documents, spreadsheet replicas, and scripts to convert them into tidy CSV files. To our knowledge, no other studies publish such replicas (Figure 1), which help identify and correct data-entry errors. Our open-science approach enables researchers to trace incidence counts back to original sources (Appendix A.3.4) and improve data quality over time.

4.3 Future directions

In this initial release of CANDID, we focused on sub-annual and sub-national disease incidence, with plans to expand further. First, we will include data stratified by age, sex, and municipality for available time ranges, provinces, and diseases. Second, we will extend the dataset's time range, disease coverage, and geographic detail as finer-scale or corrected data become available. Third, we will curate population-level information useful for epidemiological analyses, such as birth rates, mortality, vaccination, and school-term dates. Beyond expansion, we will address open questions in Canadian infectious disease history. For example, we will refine the regional stratification of whooping cough (Figure 5) to test existing hypotheses [12] on its 1990s resurgence. Finally, we will explore AI-powered optical character recognition, using our archive as a training dataset to enhance its efficiency and accuracy.

4.4 Limitations

Analyzing incidence data involves challenges in standardization and comparability. Comparing regional incidence patterns requires adjusting for under-reporting differences among provinces. Complex, evolving sub-disease hierarchies further complicate interpretation (e.g.,

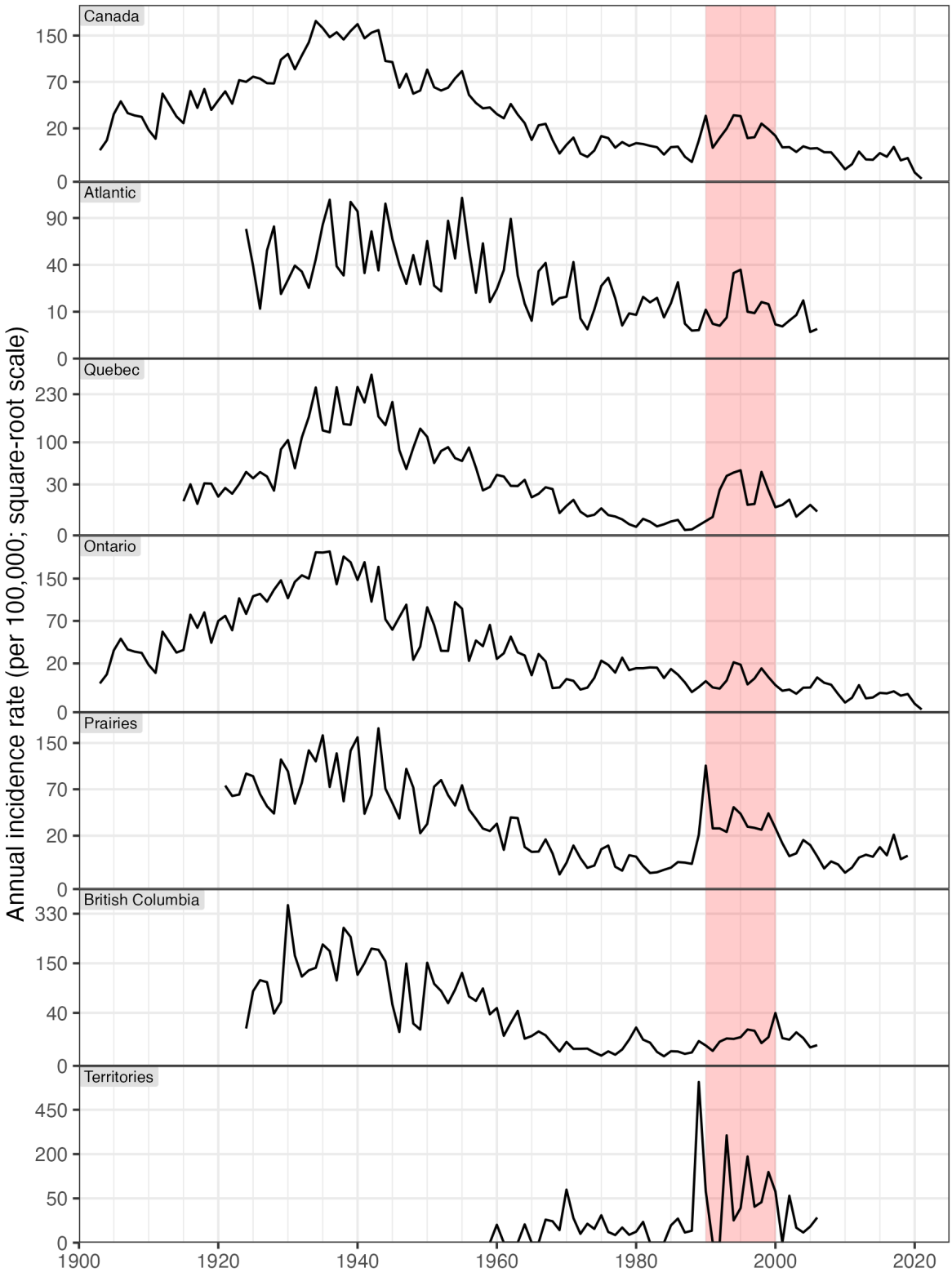

Figure 5: Caption on next page.

Figure 5: Regional [13] differences in average annual whooping cough incidence in Canada over twelve decades (square root scale). The national data (top panel) are very similar to the first figure in a review based on different data sources [12]. The red region (1990–1999) highlights the first resurgence in national whooping cough incidence since widespread vaccination began in 1943. This plot shows that not all regions peaked in the 1990s, a pattern that could not have been detected with the national data used in the original study. Methods for producing this plot are described in Appendix A.6.

the normalized dataset includes 37 meningitis sub-diseases, with 1–15 reported in any given year; Figure A8 in Appendix A.3.3). Varying time scales for incidence measurement (e.g., weekly, monthly, quarterly; Figure 3) hinder the creation of evenly spaced time series.

Ideally, CANDID would provide weekly case counts for every notifiable [22] disease in all provinces and territories. Historical gaps persist due to surveillance program changes (e.g., chickenpox was not notifiable from 1959–1985) and incomplete source coverage, particularly outside Québec, Ontario, and Saskatchewan before 1924. Gaps range from missing weeks (e.g., lost book pages) to data-entry errors, often from unclear handwritten records (§2.5). We have addressed known discrepancies for our primary example diseases (whooping cough, poliomyelitis). By releasing open data pipelines, we aim to establish a foundation for collective efforts to improve the comprehensiveness and quality of CANDID over time (details in Appendix A.4).

4.5 Conclusion

More than a century of infectious disease surveillance in Canada has produced a valuable record of epidemic patterns that has been largely unexploited, but can now be easily accessed. Comprehensive sub-annual and sub-national Canadian infectious disease incidence data have previously been unavailable. Similar data from other countries have been critical to establishing the foundations of epidemiological modelling and continue to push the field forward (e.g., [23–27]).

CANDID will facilitate progress on public health research questions that require information on intra-annual and regional variation in incidence. In principle, it should be straightforward to keep the archive up to date, but doing so will require the cooperation of provincial and territorial public health agencies, which have released *less* data publicly since strictly digital data collection began in the 1990s. We contacted all these agencies but were able to obtain recent weekly data from only two provinces. We urge all Canadian governments to require their public health agencies to release weekly, aggregated counts of infectious disease notifications publicly as a matter of course as soon as possible.

Acknowledgements

We are deeply thankful to Alberta Health and Public Health Ontario for providing us with recent weekly incidence data. We are grateful for support from the Natural Sciences and

Engineering Research Council of Canada (NSERC) via an Emerging Infectious Disease Modelling (EIDM) grant to the Canadian Network for Modelling Infectious Diseases (CANMOD). DJDE, BMB and JD were supported by NSERC Discovery grants. GM was supported by an NSERC Undergraduate Student Research Award (USRA) held at the McMaster University Department of Mathematics and Statistics. Research assistants Jen Freeman, Frank Jin, Ronald Jin, and Steven Lee wrote code that we used during this project. Research assistants Jeanne Lin, Saul Widrich, Qinxian Zhu, Claire Lees, and Julia Maja entered some of the data. Research assistants Maya Earn, Arielle Earn, and Elizabeth O'Meara found, organized, and scanned source documents. We appreciate the enthusiastic encouragement we received from Caroline Colijn and many other CANMOD colleagues.

References

[1] Ogden NH, Acheson ES, Brown K, Champredon D, Colijn C, Diener A, et al. Mathematical modelling for pandemic preparedness in Canada: Learning from COVID-19. Canada Communicable Disease Report. 2024;50(10):345.

[2] Totten S, Medaglia A, McDermott S. Updates to Canadian Notifiable Disease Surveillance System. Canada Communicable Disease Report. 2019;45(10):257-61. Available from: `https://doi.org/10.14745/ccdr.v45i10a02`.

[3] Dominion Bureau of Statistics, Canada. Sixth Census of Canada, 1921. vol. 2. Dominion Bureau of Statistics; 1925. Available from: `https://publications.gc.ca/pub?id=9.830550&sl=0`.

[4] Statistics Canada. Population-1921-1971-Revised Annual Estimates of Population, by Sex and Age Group, Canada and the Provinces. Statistics Canada; 1973. Available from: `https://publications.gc.ca/pub?id=9.817507&sl=0`.

[5] Statistics Canada. Table 17-10-0005-01 Population estimates on July 1, by age and gender; 2021. Accessed: 2022-02-16. Available from: `https://doi.org/10.25318/1710000501-eng`.

[6] PDFTables. PDFTables; 2024. Accessed: 2024-09-10. `https://pdftables.com`.

[7] Thibault RT, Amaral OB, Argolo F, Bandrowski AE, Davidson AR, Drude NI. Open Science 2.0: Towards a truly collaborative research ecosystem. PLoS Biology. 2023;21(10):e3002362.

[8] Torres-Espín A, Ferguson AR. Harmonization-information trade-offs for sharing individual participant data in biomedicine. Harvard data science review. 2022;4(3). Available from: `https://doi.org/10.1162/99608f92.a9717b34`.

[9] Wickham H. Tidy data. The Journal of Statistical Software. 2014;59. Available from: `http://www.jstatsoft.org/v59/i10/`.

[10] Cheng C, Messerschmidt L, Bravo I, Waldbauer M, Bhavikatti R, Schenk C, et al. A general primer for data harmonization. Scientific data. 2024;11(1):152. Available from: https://doi.org/10.1038/s41597-024-02956-3.

[11] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data; 2019. Available from: https://doi.org/10.14454/7xq3-zf69.

[12] Thommes E, Wu J, Xiao Y, Tomovici A, Lee J, Chit A. Revisiting the epidemiology of pertussis in Canada, 1924–2015: a literature review, evidence synthesis, and modeling study. BMC Public Health. 2020;20:1-9. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09854-4.

[13] Statistics Canada. Standard Geographical Classification (SGC) 2021 - Volume I, The Classification. Statistics Canada; 2021. Available from: https://www150.statcan.gc.ca/n1/en/pub/12-571-x/12-571-x2021001-eng.pdf?st=ihui0H05.

[14] Payne E, Totten S, Laroche J, Archibald C. Hepatitis B: Hepatitis B surveillance in Canada: 2005-2012. Canada Communicable Disease Report. 2014;40(13):266.

[15] Gasmi S, Ogden N, Lindsay L, Burns S, Fleming S, Badcock J, et al. Emerging infections: surveillance for Lyme disease in Canada: 2009–2015. Canada Communicable Disease Report. 2017;43(10):194.

[16] Lin D, Fane BHM, Squires SG, Dickson C. Relaying the burden of diphtheria in Canada with hospital data. CCDR. 2021;47(10).

[17] Lin D, Fane BHM, Squires SG, Dickson C. Influenza Vaccine: Describing the burden of diphtheria in Canada from 2006 to 2017, using hospital administrative data and reportable disease data. Canada Communicable Disease Report. 2021;47(10):414.

[18] Golden AR, Griffith A, Tyrrell GJ, Kus JV, McGeer A, Domingo MC, et al. CCDR 50th Anniversary: Invasive group A streptococcal disease surveillance in Canada, 2021–2022. Canada Communicable Disease Report. 2024;50(5):135.

[19] van Panhuis WG, Cross A, Burke DS. Project Tycho 2.0: a repository to improve the integration and reuse of data for global population health. Journal of the American Medical Informatics Association. 2018;25(12):1608-17.

[20] for Health Protection OA, Ontario) PPH. Infectious diseases query: reference guide; 2023. Available from: https://www.publichealthontario.ca/-/media/Documents/Q/2018/query-id-reference.pdf?rev=16dc019860924876976548bcbe004913&sc_lang=en.

[21] Van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. The New England journal of medicine. 2013;369(22):2152.

[22] Doherty JA, et al. Establishing priorities for national communicable disease surveillance. Canadian Journal of Infectious Diseases and Medical Microbiology. 2000;11:21-4. Available from: https://www.hindawi.com/journals/cjidmm/2000/134624/.

[23] Bartlett MS. The Critical Community Size for Measles in the United States. Journal of the Royal Statistical Society Series A (General). 1960;123(1):37-44.

[24] London WP, Yorke JA. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. American Journal of Epidemiology. 1973;98(6):453-68.

[25] Anderson RM, May RM. Directly transmitted infections diseases: control by vaccination. Science. 1982;215(4536):1053-60.

[26] Fine PEM, Clarkson JA. Measles in England and Wales—I: An analysis of factors underlying seasonal patterns. International Journal of Epidemiology. 1982;11(1):5-14.

[27] Grenfell BT, Anderson RM. Pertussis in England and Wales: an investigation of transmission dynamics and control by mass vaccination. Proceedings of the Royal Society of London Series B Biological Sciences. 1985;226(1234):475-92.

[28] Sockett PN, Garnett MJ, Scott C, et al. Communicable disease surveillance: Notification of infectious diseases in Canada. Canadian Journal of Infectious Diseases and Medical Microbiology. 1996;7:293-5.

[29] Bauch CT, Earn DJD. Transients and attractors in epidemics. Proceedings of the Royal Society of London, Series B. 2003;270(1524):1573-8.

[30] Hooker G, Ellner SP, De Vargas Roditi L, Earn DJD. Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in Ontario. Journal of the Royal Society of London, Interface. 2011;8(60):961-74.

[31] Garmonsway D. unpivotr: Unpivot Complex and Irregular Data Layouts; 2023. R package version 0.6.3. Available from: https://CRAN.R-project.org/package=unpivotr.

APPENDIX

A Methodological Details

A.1 Data Sources

Since 1924, the Dominion Bureau of Statistics, now Statistics Canada, has collected communicable disease incidence data.

> The main objectives of this system are to provide a mechanism for monitoring the health of the population by identifying and responding to changes in reporting trends of specific diseases and to provide information that can contribute to the development of health policy and the planning of care, prevention, and control programs. [28]

This initiative persists today under the administration of the Public Health Agency of Canada (PHAC), as the Canadian Notifiable Disease Surveillance System (CNDSS). Data for the CNDSS are provided by provincial and territorial governments to monitor diseases of public health concern [2]. Prior to the onset of this federal initiative in 1924, several provinces were already collecting such data without reporting it to the federal government.

Our group's search for historical Canadian infectious disease notification data began in 2000. Initially, we acquired handwritten weekly counts for Ontario, covering five decades (1939–1989), from the Ontario Ministry of Health (Table 1). A significant breakthrough occurred during a visit to the chief medical officer of health in Manitoba, where he discovered a single page of notifications submitted to the Dominion Bureau of Statistics. We photocopied this document and sent it to Statistics Canada, as it was the first evidence that the data tables we were seeking existed.

In 2002 and later, we engaged in extensive communications with Statistics Canada via telephone and email. As a result, they provided photocopies of handwritten weekly and monthly notification spreadsheets from 1924–1955 that they had located in their archives. At the time, our resources were insufficient to digitize and clean all of these data. However, we did present analyses of a few disease time series in publications [29, 30]. Additionally, we located more published CNDSS data from 1956–2000 in the library collections of McGill University, McMaster University, and the University of Alberta. During this historical period, the publication of data transitioned from weekly to monthly and eventually to quarterly. Quarterly data up to 2007 are available online in the Canada Communicable Disease Report (CCDR). Unfortunately, we were unable to find any sub-annual and sub-national data source covering all provinces after 2007.

In 2021, we began reaching out to provincial and territorial public health agencies for more recent data. Although we corresponded via email with all such agencies, only Public Health Ontario and Alberta Health provided us with recent data. We were able to find some other provincial data in publications and online.

A.2 Data access

Download links to the datasets described in this paper, as well as others from our broader Canadian historical epidemiological data digitization project, can be found at the following web page:

`https://github.com/canmod/iidda/blob/main/README.md`

We have also written a small package for reading the data directly into R. An introduction to this package can be found here:

`https://canmod.github.io/iidda-tools/iidda.api/articles/Quickstart`

A.3 Data preparation pipelines

Each data preparation pipeline followed one of the paths outlined in Figure A6. Data sources were provided as hard copies, digitally produced PDF files, or digital spreadsheets. Digital spreadsheets were particularly advantageous because they enabled us to directly script the production of CSV files. When data were not in spreadsheet format, conversion was necessary before scripting could begin. For digitally produced PDF files, we tried to use automated tools like PDFTables (`https://pdftables.com/`) to convert them into spreadsheets. Hard copies, however, always required scanning followed by manual data-entry into spreadsheets. We were unable to find a viable optical character recognition (OCR) approach to avoid manual data-entry. However, in the years since we began this systematic effort to digitize Canadian incidence data there have been tremendous advances in artificial intelligence (AI), and so this situation may have changed. In future work on digitization we plan to replicate parts of this work using OCR, to test it as an efficiency tool in this area.

All CANDID preparation pipelines are available at:

`https://github.com/canmod/iidda/tree/main/pipelines`

The main scripts for producing each of the three datasets (unharmonized, harmonized, and normalized) discussed in this article are available at:

`https://github.com/canmod/iidda/tree/main/pipelines/canmod-compilations/prep-scripts`

These scripts depend on outputs produced by other pipelines that generate data from specific sources. Instructions on how to reproduce all of these outputs are provided in the `README.md` file on this GitHub repository:

`https://github.com/canmod/iidda`

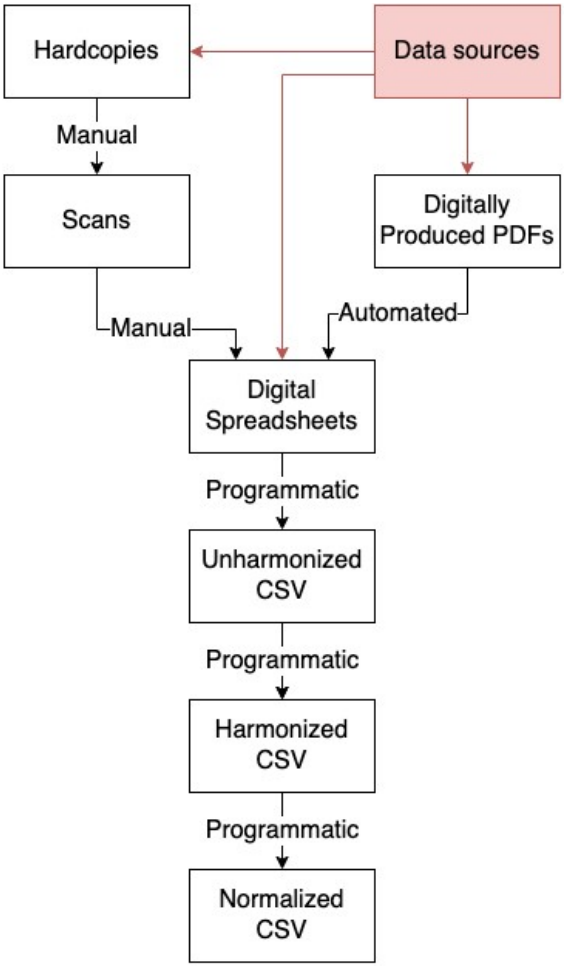

Figure A6: Data preparation pipeline overview. All products begin with data sources (red), which provide hard copies, digital spreadsheets, or digitally produced PDFs. Processing steps are classified as manual (e.g., data-entry), automated (e.g., PDF table extraction tools), or programmatic (e.g., customized R script).

A.3.1 Preparing Unharmonized CSV Files

We developed one open-source R script for each spreadsheet that converts it into a tidy CSV file [9]. These R scripts used the `unpivotr` package [31] to convert the wide-format data used in historical documents to long-format data that make it easier to manipulate using standard tools [9]. Long-format data also make it easier to combine data from different sources into a single data set, by ensuring they each consist of fields from the same standardized data dictionary. All of the resulting CSV files have been collected into a single file that we label *unharmonized* because it contains unharmonized historical disease and place names.

These unharmonized data also include information on why certain incidence values are not available, when we could determine the reason from the sources. In the column containing numbers of cases, `cases_this_period`, we allowed the following types of values:

- Non-missing numeric case numbers (non-negative integers).
- One of the following strings explaining why the case numbers are missing[1]:
  - The phrase '`Not available`', for unknown reasons.
  - The phrase '`Not reported`', for unknown reasons.
  - The phrase '`Not reportable`', presumably indicating that the jurisdiction was not required to report these numbers.
  - The word '`Missing`', typically indicating missing pages in the middle of a multi-page table.
  - The word '`Unclear`', meaning that the value is missing from CANDID because it is not legible.
  - The word '`Unclear`', with a special string format,

    ```
    {guess_1} - {guess_2} - ... - {guess_n} (unclear)
    ```

    (e.g., '`36-23-59 (unclear)`'), meaning that the value is missing because the number is difficult to read but we have one or more guesses. In the harmonized datasets (Appendix A.3.2) we use the first (i.e., best) guess.

A.3.2 Preparing Harmonized CSV Files

We developed one open-source R script for each data source that joins the unharmonized data with harmonized disease and place names. This was achieved by creating lookup tables containing all historical names, and then adding columns for harmonized names. Links to these lookup tables are available at:

`https://github.com/canmod/iidda/blob/main/README.md#canmod-digitization-project`

[1] Typically these strings were taken as is from the source, but in many instances before 1924 we had to make an educated guess about the reason why particular records were missing.

Because sources sometimes reported diseases hierarchically, our harmonized disease names were provided in two columns: `disease`, giving the name of the disease being reported, and `nesting_disease`, optionally giving the name of another disease within which the `disease` is nested. We refer to the values in these `disease` and `nesting_disease` columns collectively as ***disease names***. These disease names are organized into hierarchies, such that most disease names are nested within another disease name (e.g., hepatitis-A is nested within hepatitis-A-B). We refer to disease names that are not nested within any other disease name as ***basal diseases***. These basal diseases are plotted in Figure 3 of the main text. For a given combination of location and time period, all disease names at or below a specific basal disease in the hierarchy are referred to as its ***sub-diseases***.

In addition to joining harmonized names, the harmonization scripts also apply the following changes:

- Remove illegible data.
- Apply fixes to dates and locations that were obviously entered incorrectly in the original source documents (e.g., one whooping cough record from 1943 was for a week ending on a Sunday, while all other data from the same source were for weeks ending on a Saturday).
- Aggregate data that were stratified by age or city.
- Replace alternative characters for reporting zero cases with a literal `0` (e.g., often a dash was used).
- Remove records containing missing values or text strings indicating the type of missing value (e.g., '`Unclear`', '`Missing`').

These fixes add convenience at the expense of removing information contained in the original source, but this information remains accessible in the unharmonized data Appendix A.3.1.

We collected all of the resulting harmonized CSV files into a single file that we label *harmonized*.

A.3.3 Preparing the Normalized CSV File

We developed an open-source R script to remove redundant data and to add data that are implied but not explicitly provided by the original source documents. Our goal was to ensure that each apparently reported case is represented by a single incidence value in the resulting CSV file. We refer to the resulting file as *normalized* aligning with the concept that normalized databases represent each 'fact' only once [9].

There are five sources of overlap that could cause cases to be counted more than once:

- Locations (e.g., provincial data being reported along with national data)
- Data sources (e.g., weekly data for Ontario between 1939–1978 being reported by both Statistics Canada and the Ontario Ministry of Health)

- Time periods (e.g., weekly data being reported along with monthly data)
- Disease hierarchies (e.g., polio with and without paralysis being reported along with total polio)
- Mixtures of the previous sources of overlap (e.g., some provinces have only monthly data, while others have only weekly data.)

In addition to removing overlapping historical records, we also add records that are implied by the information in the sources. There are two types of implied information in the harmonized data that we have made explicit in the normalized data:

- Unaccounted cases that are detected when the reported number of cases for a disease is greater than the total over its sub-diseases.
- Missing data at finer time-scales (e.g., weekly, monthly) can be assumed to be zero if zeros are reported at a coarser time-scale (e.g., yearly data) that temporally bound the finer scale.

The CANDID archive curates data from a wide variety of sources (Table 1) and diseases, each with different biases and quality issues. It is therefore impossible to produce a perfectly normalized dataset, nor is it our goal to do so here. Instead we aim (1) to make reasonable choices so that the analyses in this paper respect basic principles of consistency (e.g., each case is counted at most once), (2) to set up a data processing pipeline that allows sustained work to improve the quality of the normalization process, and (3) to make the complete and un-normalized data easy to access so that others with expertise in a particular area can make improved normalization choices.

In the remainder of this section, we describe the normalization steps that we take to produce our posted normalized files. Users are free to modify these steps by modifying the scripts and other files these scripts depend on (see Appendix A.3.4 for information on how to find these scripts and files).

Add unaccounted cases: Sometimes the total for a `nesting_disease` is reported along with some, but not all, of its sub-diseases. In these instances, after having ruled out other known data quality issues, we produced records with incidence counts given by the associated reported total minus the sum of the reported sub-diseases. These incidence values can be identified in the normalized dataset by a value in the `disease` column of the form `nesting-disease_unaccounted`, and with `derived-unaccounted` in the `record_origin` column. These sums were computed by grouping by time-period, province, data source ID, and `nesting_disease`.

Join population data: The estimated provincial population for each incidence value was joined to the normalized dataset. As a result, population numbers are repeated in the dataset because incidence values for different diseases are linked to the same population size within

a specific period and province. While this repetition technically violates our normalization principle, the added convenience justifies this step. The `population` column provides linearly interpolated estimates of the intercensal populations for each province at the mid-point of each period, using census-derived data from [3–5].

Resolve overlapping locations: Filtering out all data for the entire country easily solves this source of overlap.

Compute implied zeros: In some instances, there are zeros reported at a coarse time-scale (i.e., for a year), but the data at a finer timescale (weekly/monthly) for the same disease and location is empty or not available. We replaced weekly data that were missing and/or not available in national data sources with zeros when they were implied by a zero at a coarser timescale for the same disease and location. These incidence values can be identified in the normalized dataset by a value of `derived-implied-zeros` in the `record_origin` column. These implied zeros were given lower priority than other weekly data when resolving overlap, as we describe next.

Resolve overlapping data sources and time-scales: We generally prioritize national data sources that report for all provinces (e.g., Statistics Canada) over provincial data sources that report for a single province (e.g., Saskatchewan Bureau of Public Health). We always prioritize finer time-scales (e.g., weekly) over coarser ones (e.g., quarterly). For example, if monthly data from a national source overlaps with weekly data from a provincial source, we will choose the weekly provincial data.

We handle data source and time-scale overlap sequentially, starting with an empty dataset and adding records from a dataset produced by applying the previous normalization processes to the harmonized data. At each step in this sequence, we consider new candidate records and only add those that do not overlap temporally with the existing ones. Being added first therefore indicates a higher priority:

1. All weekly data from national sources.
2. Non-overlapping weekly data from provincial sources.
3. Non-overlapping two-weekly data from national sources.
4. Non-overlapping two-weekly data from provincial sources.
5. Non-overlapping weekly implied zeros from national sources.
6. Non-overlapping monthly data from national sources.
7. Non-overlapping monthly data from provincial sources.
8. Non-overlapping quarterly data from national sources.

For any time period and province, we have at most two data sources: one from a federal organization (e.g., Statistics Canada) and one from a provincial organization (e.g., Saskatchewan Bureau of Public Health). To resolve such overlap when it occurs, we prefer national sources to those from provincial sources. This choice has the advantage of being easy to apply and also has a better chance of producing provincial data streams that are comparable because we can inherit the choices that the federal organization made when publishing data from different provinces.

If the two sources produced identical results then this choice would be irrelevant. Although there are periods for which national and provincial sources reported identical counts, this is not typically the case. Figure A7 gives an example comparing 37 years of weekly whooping cough data in Ontario as reported by Statistics Canada and the Ontario Ministry of Health. This figure shows that until 1970 the two agencies were reporting virtually identical numbers, with the occasional deviation. In contrast, there are deviations consistently from 1970 to 1977, although the qualitative pattern is still similar.

Resolve overlap caused by disease hierarchies: To address how this type of overlap is resolved, it is necessary to further define terminology related to disease hierarchies, building on the concepts introduced in Appendix A.3.2. The ***global hierarchy*** of a basal disease includes all sub-diseases that appear at least once in the harmonized dataset, while the ***local hierarchy*** is specific to a particular location and time period. Some of these global hierarchies are simple (e.g., whooping cough has no sub-diseases at all) whereas others are complex (e.g., meningitis has 37 sub-diseases in the global hierarchy) with local hierarchies changing over time.

We will dig into the meningitis hierarchy a little to give a sense of the complexity. The harmonized dataset contains 33 different local hierarchies of meningitis. Figure A8a to Figure A8d give the meningitis global hierarchy, with each figure highlighting the local-hierarchy associated with a specific set of locations and times. From 1921 to 1967 Statistics Canada reported a total meningitis count without any sub-diseases (Figure A8a). The sources give no indication of what kind of meningitis is being reported, possibly because it was not known. From 1969–1978 only viral meningitis was reported and this was stratified by coxsackie, echo, and virus-unspecified (Figure A8b). Even if totals for meningitis-viral or meningitis were given in these sources, they were excluded from the normalized data to avoid overlap. From 1979 to 1985 the collection of sub-diseases changed completely to report only meningitis associated with encephalitis (both viral and bacterial, Figure A8c). Collections of sub-diseases from provincial data sources could sometimes be quite complex (e.g., Figure A8d). Even in this complex case, there is no overlap in observed diseases in the normalized data.

The above examples illustrate that we resolved overlap caused by disease hierarchies by keeping only the most detailed sub-diseases of each basal disease, and setting the nesting diseases of these sub-diseases to be the basal disease. Intermediate nesting disease counts were then removed, keeping only the finest stratification of each basal disease reported in a given time period and location. For example, the stratification of meningitis illustrated

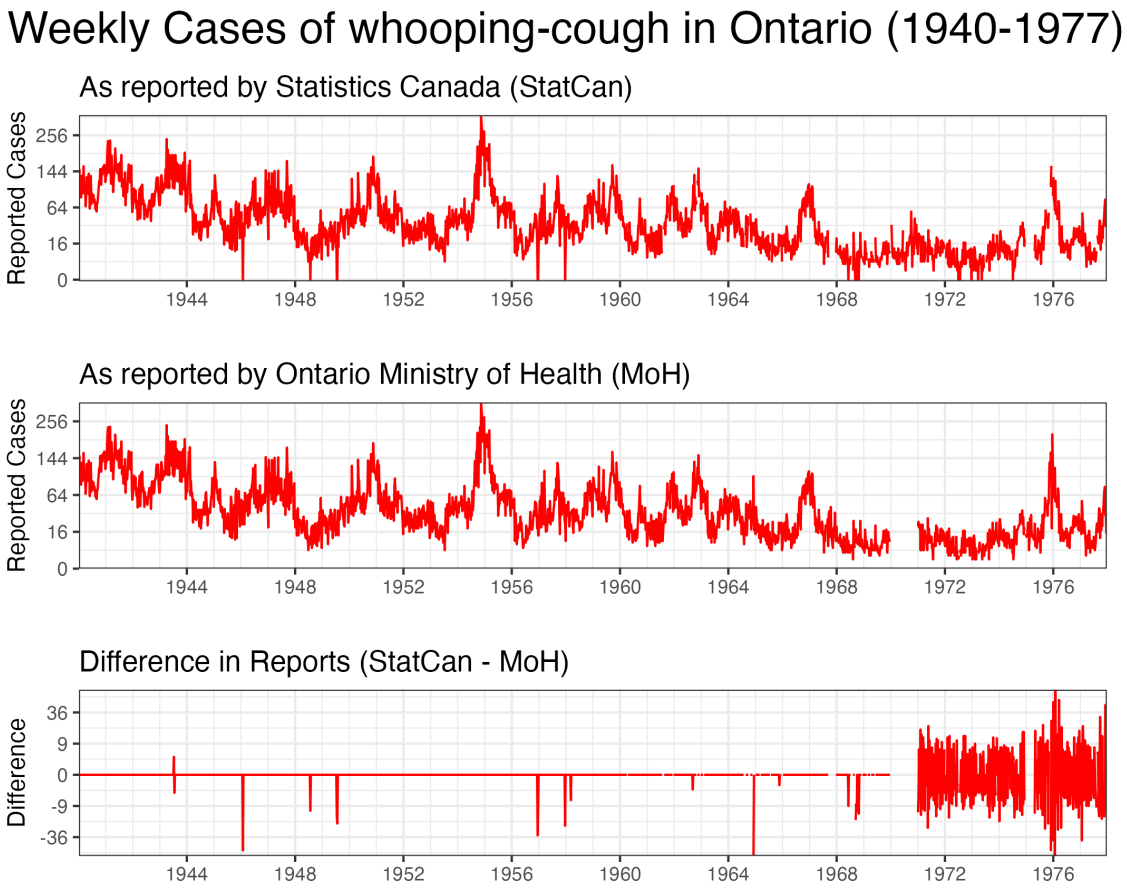

Figure A7: Comparing reported Ontario whooping cough incidence from Statistics Canada with the Ontario Ministry of Health. The difference between the two sources is given on the bottom panel.

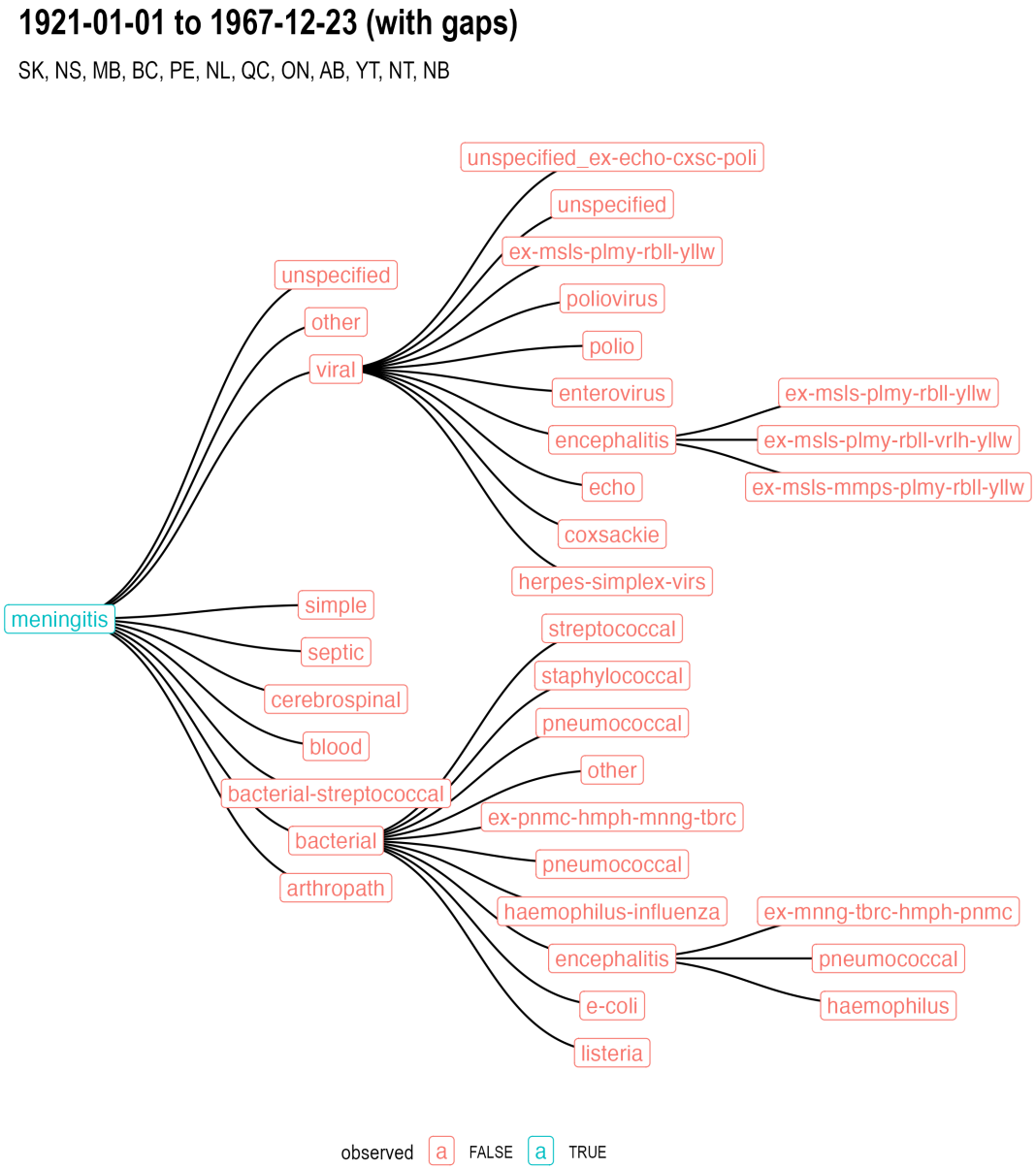

Figure A8a: Global meningitis disease hierarchy highlighting in blue a particular local hierarchy of reported sub-diseases between 1921 and 1967. In this local hierarchy, the only reported sub-disease is the basal disease itself. Full caption below.

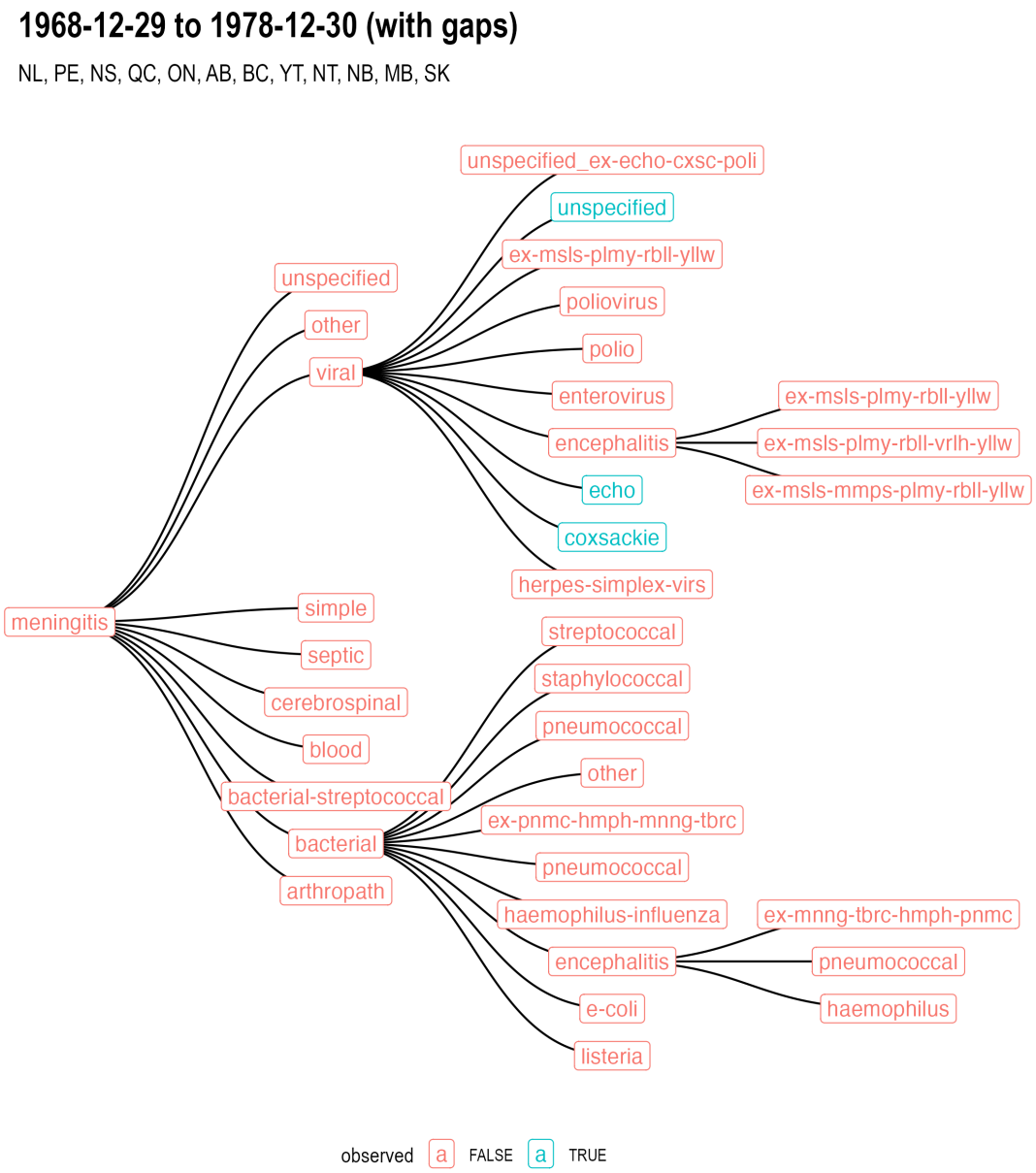

Figure A8b: Global meningitis disease hierarchy highlighting in blue a particular local hierarchy of reported sub-diseases between 1968 and 1978. Full caption below.

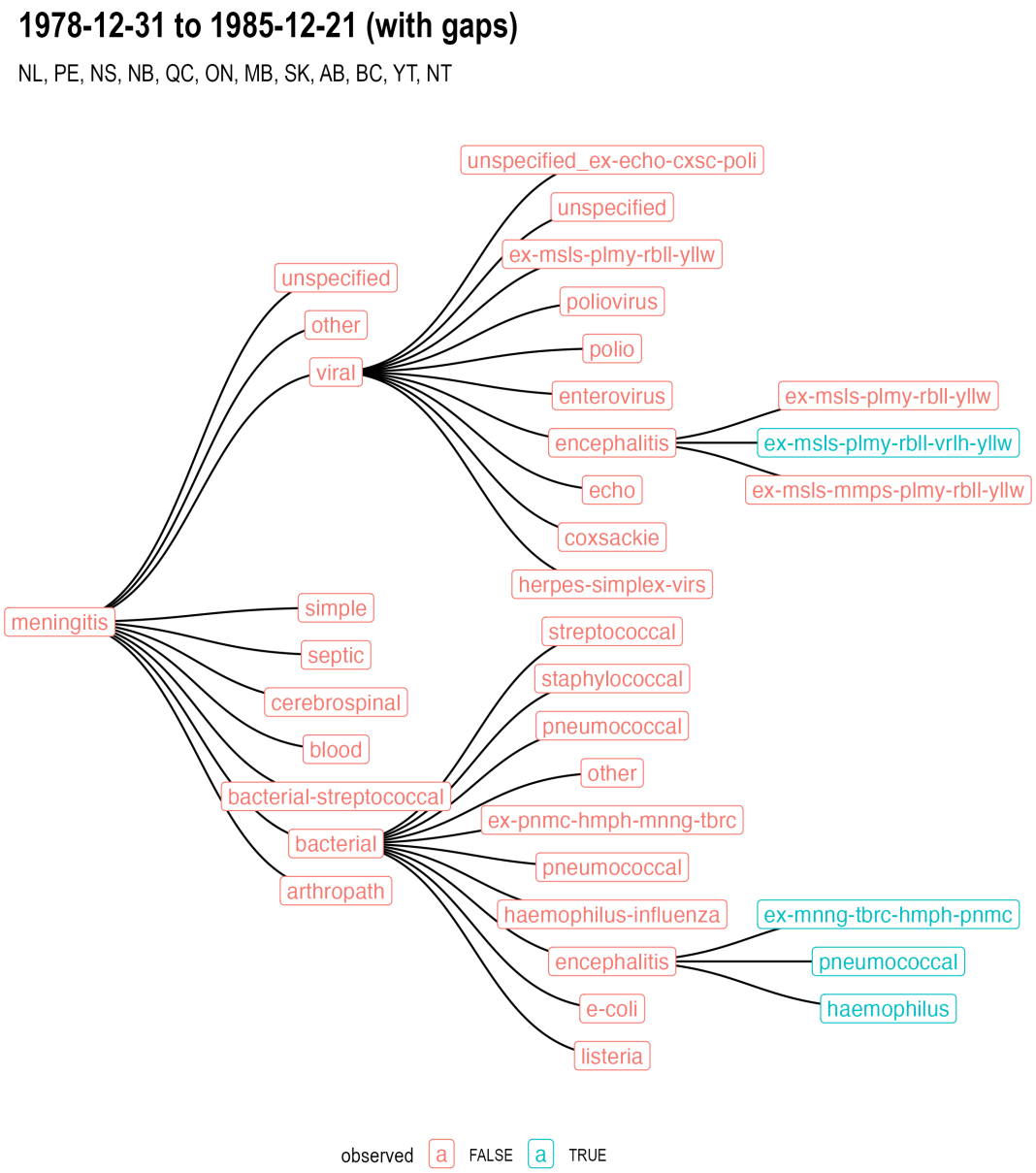

Figure A8c: Global meningitis disease hierarchy highlighting in blue a particular local hierarchy of reported sub-diseases between 1979 and 1985. Full caption below.

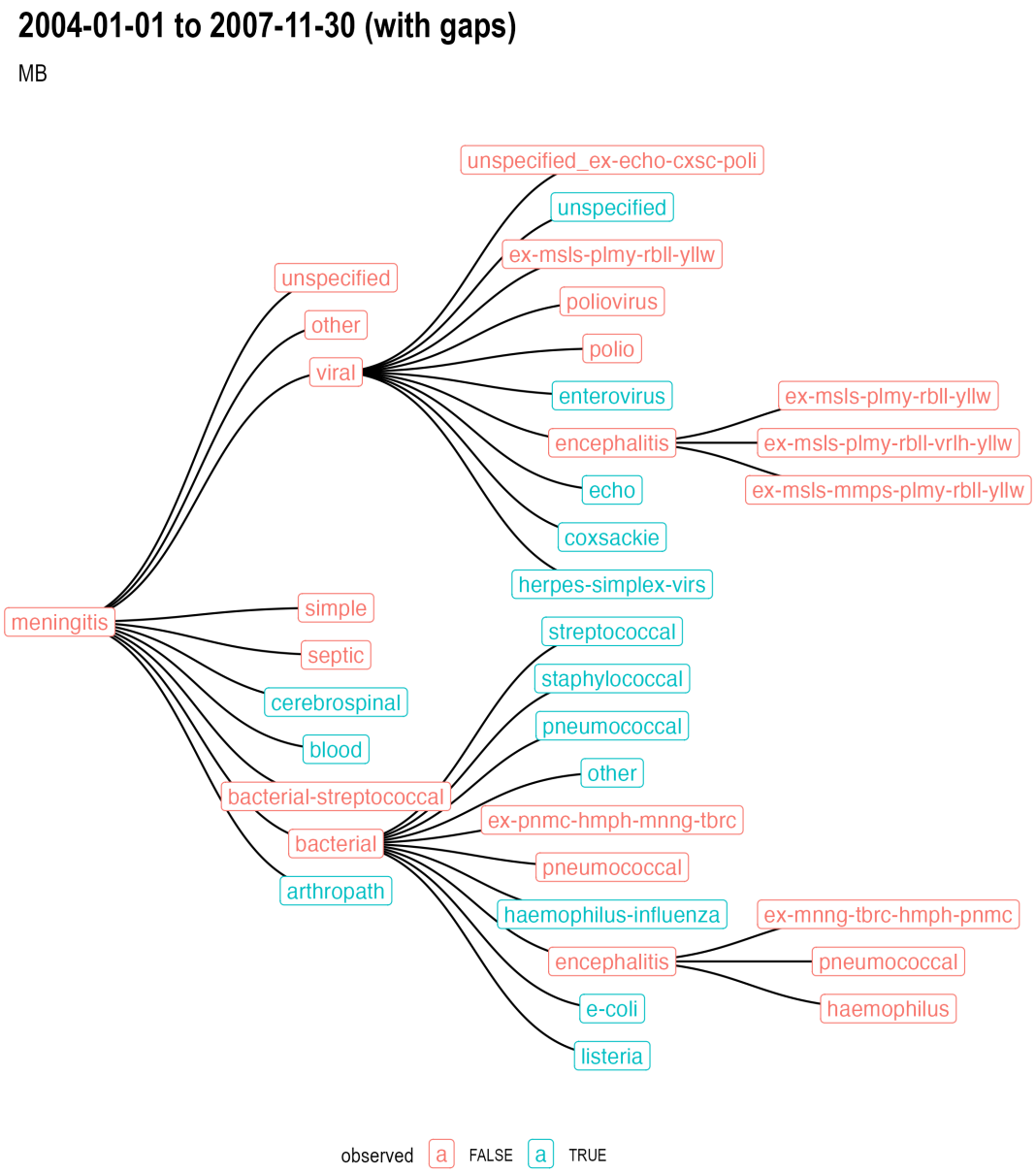

Figure A8d: Global meningitis disease hierarchy highlighting the sub-diseases reported by Public Health Manitoba in the date ranges given. Full caption below.

Figure A8: Global hierarchy of meningitis diseases. Each sub-figure corresponds to a specific date and location range, with a local hierarchy of sub-diseases for that range highlighted in blue and others in red. Disease names starting with "ex" indicate counts that exclude certain disease types, abbreviated in the figure. Full disease names are constructed by concatenating node names along the hierarchy with dashes (e.g., meningitis-bacterial-haemophilus-influenza).

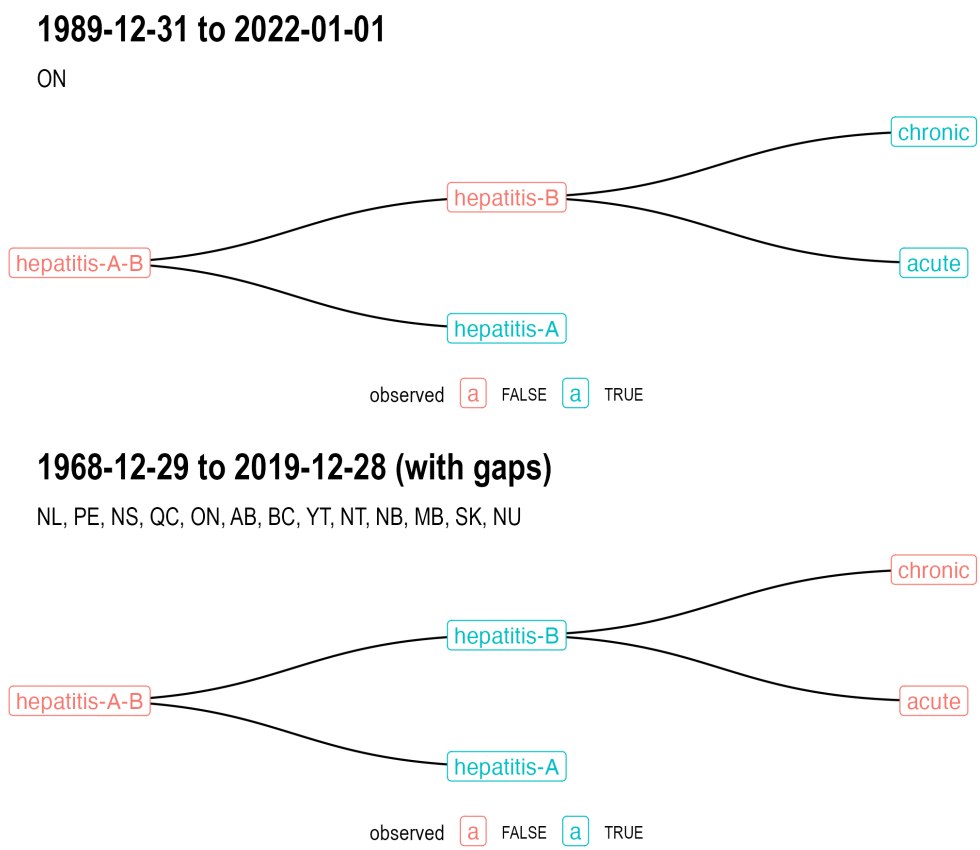

Figure A9: Global hepatitis A and B disease hierarchy, with two panels highlighting different local hierarchies in blue. In the top panel hepatitis B is stratified by acute and chronic sub-diseases, but in the bottom panel hepatitis B is not stratified. See Figure A8 for interpretation details.

in Figure A8b came from a data source (Statistics Canada) that also reported a total for viral meningitis (historically called aseptic meningitis), but this total was removed in the normalization process and so the viral node is coloured red to indicate that this sub-disease cannot be observed in the normalized data. For another example, Public Health Ontario stratified hepatitis B by acute and chronic sub-diseases (Figure A9, top) whereas Statistics Canada and other provincial agencies did not (Figure A9, bottom).

A.3.4 Data Provenance

The harmonized and normalized data both contain the following columns with unique identifiers to resources used to produce each record:

- `original_dataset_id`: Uniquely identifies the unharmonized dataset containing each record.

- `digitization_id`: Uniquely identifies the digitization (typically an Excel file) containing the data for each record.
- `scan_id`: Uniquely identifies the scan of the original document containing each record.

Records that were implied from associated data (see Appendix A.3.3), as opposed to explicitly reported, do not have entries in the first two of these columns. Data that we received in digital form typically do not have a scan associated with them, and so such records do not have a `scan_id`.

Information on the above identifiers can be found at:

`https://github.com/canmod/iidda/blob/main/README.md#identifiers`

The following vignette describes how to investigate the provenance of a record using these identifiers:

`https://canmod.github.io/iidda-tools/iidda.api/articles/Provenance`

Metadata for the unharmonized, harmonized, and normalized versions of our prepared data are available at:

`https://github.com/canmod/iidda/blob/main/README.md#canmod-digitization-project`

These metadata contain links to all of the resources (R scripts, Excel/CSV files, and PDFs) used to produce the datasets.

A.4 Quality control

Marginal total cross checks: Using marginal totals from the comprehensive dataset, we performed cross checks over the three available stratifications of data: time scales, locations, and diseases. The analysis involved several comparisons at different levels of data aggregation.

For time-scales, sub-annual records, such as weekly, monthly, or quarterly data, were aggregated to the annual scale and compared to the available annual records. When multiple sub-annual time scales were present, their annual sums were also compared to one another.

For locations, sub-national data were aggregated to the national level, and this sum was compared with the reported national total. Both the national and sub-national totals were aggregated to yearly scales for comparison with records in the PHAC portal.

For diseases, sub-class totals were aggregated and compared to the overall disease totals. In cases where the sum of sub-diseases was less than the reported disease total, the difference was labeled as `disease_unaccounted` and included in the normalized data as a derived entry with a `record-origin` of `derived-unaccounted-cases`.

For each cross check, we put all records with discrepancies into a CSV file with provenance information for finding the original scans and excel files to more easily fix potential errors. The scripts for producing these CSV files can be found at:

https://github.com/canmod/iidda/tree/main/pipelines/canmod-cross-checks/prep-scripts

A link to download the current state of these CSV files, along with the data themselves, is located at:

https://github.com/canmod/iidda/blob/main/README.md#canmod-digitization-project

We plan to continue addressing these potential issues, but if you use the archive, please check the discrepancy files to see if any data of interest is flagged. Given that our pipelines and tools are open, users are encouraged to fix issues and submit pull requests to our GitHub repository:

https://github.com/canmod/iidda

The following list summarizes the current status of each cross-check:

- Time-scale cross checks
  - Total number of year-location-disease combinations with discrepancies: 373
  - Percentage of these combinations without these discrepancies: >99%
  - Percentage of these discrepancies that are from handwritten data: 100%
- Location cross checks
  - Total number of period-disease pairs with discrepancies: 2,067
  - Percentage of these pairs without these discrepancies: >99%
  - Percentage of these discrepancies that are from handwritten data: 63%
- Disease cross checks
  - Total number of period-location pairs with discrepancies: 187
  - Percentage of these pairs without these discrepancies: >99%
  - Percentage of these discrepancies that are from handwritten data: 88%

It is difficult to combine this information into an estimate of the overall proportion of the harmonized dataset that is error free, because the above estimates are for different stratifications of the data. We can expect errors to yield discrepancies in more than one cross check. For example, a single data-entry error in a sub-disease, province, and week could trigger a discrepancy in all three cross checks if the disease, national, and annual data were reported as well. Not all of these discrepancies represent our data-entry errors because sometimes the original sources are inconsistent or unclear (especially in the handwritten data). Sources also varied in the quality of their marginal totals, and so we removed these totals from our cross checks. But overall, given that the percentages of the combinations of factors that are free of known discrepancies are all greater than 99%, we do not expect that many more data-entry errors remain.

Comparing with the PHAC portal: We also compared the national and yearly data on the PHAC portal (`https://diseases.canada.ca/notifiable`) with aggregated national and yearly totals in CANDID for what we believe are the same diseases. We made these comparisons visually using line plots. Figure A10 gives an example using chickenpox.

We do not expect the data sources to match exactly, for a variety of reasons. Before 1924 and after 2000 all of the data come from provincial data sources, and so our spatial coverage is limited at these times, whereas the PHAC portal typically reports using data from more provinces after 1924 (though not always, as is evident in the bottom panel of Figure A10). CANDID historical source documents presumably included the best numbers at the time. PHAC may have updated these numbers to account for data quality corrections or changes in criteria for determining whether a case qualifies as a specific disease. Discrepancies could also occur because the sub-diseases included in a particular disease change over time.

Fixing data: These data quality checks have allowed us to correct errors, and to continue to do so, including data-entry typos (e.g., if 500 cases should have been 50), resolving bugs in preparation scripts, and rethinking the interpretation of data source organization. For instance, the difference between zero incidence and missing incidence was not clear in some of the handwritten data (Figure 2). Additionally, we needed to change how we aggregated sub-diseases to be comparable with such aggregations on the PHAC portal. Reasons for irreducible discrepancies are discussed in the section on limitations (§4.4).

A.5 Polio methods

We analyzed weekly provincial polio data. For each province and week, we calculated the weekly incidence rate by multiplying the number of new cases by 100,000 and dividing by the interpolated population for that week. To estimate the national incidence rate, we summed the provincial cases and divided by the population of provinces with available data, then multiplied by 100,000. Vertical lines in the plot Figure 4 indicate the week with the highest national incidence in a given polio year, which we defined as the 52-week period centered around epidemiological week 34, the typical yearly peak for polio. We plotted peaks only for years with more than 20 total cases reported in the country.

A.6 Whooping cough methods

We estimated the annual incidence rate in each of the six geographic regions of Canada, and Canada as a whole, as follows. We first computed the average daily number of whooping cough cases for each province or territory by summing the reported cases over all available time periods within a year and dividing by the total number of observed days. We then estimated the total annual cases for each province by scaling this daily rate by the total number of days in the year, accounting for any missing data. Summing these estimates across the provinces within each geographical region gave us the total estimated cases per region. Finally, we calculated the incidence rate per 100,000 individuals by dividing the regional total estimated cases by the total regional population and scaling appropriately. These sums

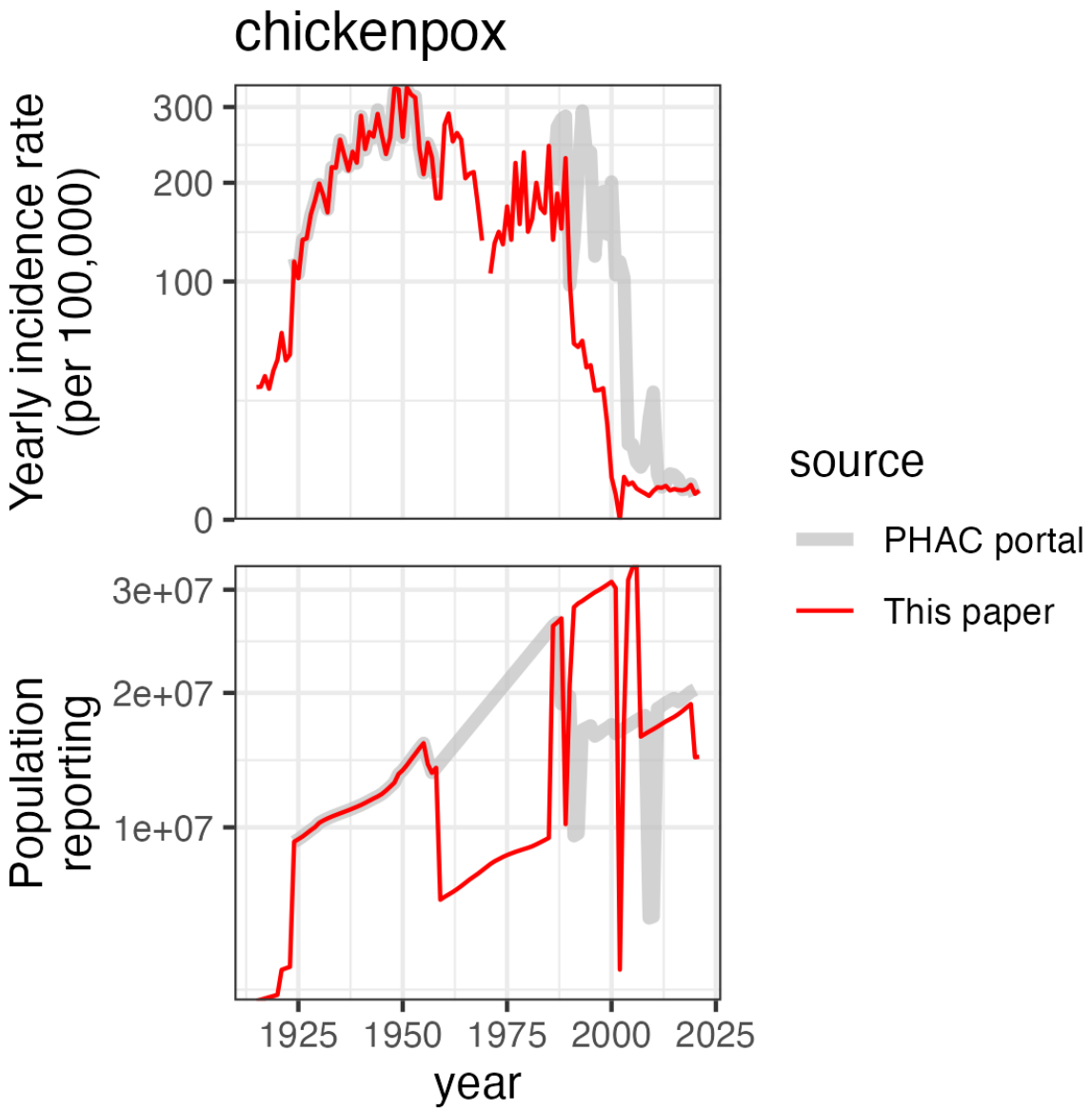

Figure A10: Chickenpox as an example of a comparison between our archive and the PHAC portal. The top panel gives yearly incidence from CANDID in red and the PHAC portal in grey. The bottom panel gives the total population of all reporting provinces for each source. Jumps in this bottom panel are caused by provinces being added or dropped from each source. For chickenpox the reported incidence is very similar in both sources during times when the reporting population is identical.

and averages were not contaminated by double counting, as each reported or implied case is represented only once in the normalized dataset (see Appendix A.3.3). However, our method could be affected by within-year variation in incidence rates that did not average out over the sample of available time periods.

Details for these calculations are as follows:

- Let x_{ij} be the number of new whooping cough cases reported during time period i (e.g., week, month, or quarter) within province or territory j of Canada.
- Let n_{ij} be the number of days in time period i within province or territory j.
- Let Ω_k be the set of all time periods i within year k. Note that Ω_k may not include all possible periods if data are missing for some weeks, months, or quarters.
- Let Ψ_l be the set of provinces and territories j that are contained within geographical region l of Canada (e.g., Atlantic, Quebec, Ontario, Prairies, British Columbia, Territories).
- Let N_k be the total number of days in year k (either 365 or 366).
- Let p_{kj} be the population of province or territory j during year k.

1. **Average Daily Cases in Year k and Province j:**

$$\text{Average Daily Cases}_{kj} = \frac{\sum_{i \in \Omega_k} x_{ij}}{\sum_{i \in \Omega_k} n_{ij}}$$

2. **Estimated Total Annual Cases in year k in Province j:**

$$\text{Estimated Annual Cases}_{kj} = N_k \times \left(\frac{\sum_{i \in \Omega_k} x_{ij}}{\sum_{i \in \Omega_k} n_{ij}} \right)$$

3. **Total Estimated Cases in Year k and Region l:**

   Sum the estimated annual cases over all provinces and territories j within region l:

$$\text{Total Estimated Cases}_{kl} = N_k \times \sum_{j \in \Psi_l} \left(\frac{\sum_{i \in \Omega_k} x_{ij}}{\sum_{i \in \Omega_k} n_{ij}} \right)$$

4. **Incidence Rate per 100,000 Individuals in Year k and Region l:**

   Calculate the incidence rate by dividing the total estimated cases by the total population of the region and scaling by 100,000:

$$\text{Incidence Rate}_{kl} = 100,000 \times N_k \times \frac{\sum_{j \in \Psi_l} \left(\frac{\sum_{i \in \Omega_k} x_{ij}}{\sum_{i \in \Omega_k} n_{ij}} \right)}{\sum_{j \in \Psi_l} p_{kj}}$$