Appendix A: Organisations and initiatives that tested the min specs

Data trust, Data repository, or Data collaboration	Description
Actionable Intelligence for Social Policy (AISP)	Actionable Intelligence for Social Policy (AISP) helps state and local governments collaborate and responsibly use data to improve lives. AISP does not hold data. AISP is a network and peer learning community of 36 distinct state and local data collaborations, each of which holds cross-sector data on health, education, and social service provision. Collectively, the AISP network covers over 50% of the US population. Website: https://www.aisp.upenn.edu/
The Centre for Addiction and Mental Health (CAMH)-wide BrainHealth Databank	The Centre for Addiction and Mental Health (CAMH)-wide BrainHealth Databank initiative accelerates research and improves care by collecting and studying the full spectrum of data that individuals choose to share to advance mental health. The BrainHealth Databank holds data from diverse multi-modal research projects and integrated clinical programs in Ontario, Canada, and includes records for over 25,000 participants. Website: https://www.camh.ca/en/science-and-research/discovery-fund/brainhealth-databank
Canadian Partnership for Tomorrow's Health (CanPath)	The Canadian Partnership for Tomorrow's Health (CanPath) is a large-scale, prospective cohort platform used by the research community to study the biology, behaviours, and environments of Canadians to advance knowledge of the causes and control of chronic diseases and cancer. CanPath holds and regularly updates data provided with consent by more than 300,000 Canadians who are actively involved and followed in every province. CanPath data include surveys, health measures, environmental exposures, and linkages to health system records. Many participants also provided physical measures and biological samples from which genomic data can be derived. Website: https://canpath.ca/
Canadian Institute for Health Information (CIHI)	CIHI is an independent, not-for-profit organisation funded by the Canadian federal government and Canadian provinces and territories to deliver comparable and actionable information to accelerate improvements in health care, health system performance and population health across the continuum of care. CIHI is a secondary data collector of health information and holder of health data. CIHI data holdings include population-level hospital data for almost the entire population of Canada as well as data from other health care facilities, long-term care homes, regional health authorities, medical practitioners and governments. Website: https://www.cihi.ca/en
Canadian Research Data Centre Network (CRDCN)	The CRDCN is a partnership between Statistics Canada and a consortium of more than 30 Canadian universities. Through the CRDCN, approved researchers in university, government, and other sectors are able to access linked but anonymised social, economic and health datasets prepared by Statistics Canada that, in some cases, include entire populations of provinces or territories. Website: https://crdcn.ca/
Sherbrooke University Hospital Center (CHUS) Biobank	The CHUS Biobank is located in Quebec, Canada, and provides data for approved research projects from researchers. Specimens include blood, tissue, fluid, stool, microbiome and living derivatives. Data include demographic data, health status (weight, blood test, illnesses, drugs, medical images, etc.), lifestyle habits (diet, alcohol consumption, physical activity, etc.), medical and family history, clinical data as well as data from genome-wide sequencing. Website: https://biobanking.org/biobanks/view/123

Continued

Appendix A: Continued

Data trust, Data repository, or Data collaboration	Description
Digital Cardiac Health Platform (DCHP)	The Digital Cardiac Health Platform (DCHP) is a secure and high-performance data integration platform for secondary use of clinical data located in Ontario, Canada. Website: https://www.uhn.ca/PMCC/Documents/Strategic-Expansion-of-Digital-Cardiac-Health-Platform-into-UHN-Digital-Health-Platform.pdf
GEMINI	The GEMINI Study collects routinely-generated administrative and clinical data for research and quality improvement. GEMINI holds data from over 30 participating hospitals in Ontario, Canada, and includes records for over 2 million hospitalisations. Website: https://www.geminimedicine.ca/
Hartford Data Collaborative (HDC)	The HDC is part of the Connecticut Data Collaborative. It includes a network of non-profit organisations, government agencies, and philanthropic partners that facilitates data sharing and data integration among its partners in the Hartford, Connecticut area. HDC does not hold data, it links data for others. HDC works with data on demographics, health, education and social service provision. Website: https://www.ctdata.org/about-hdc
Health Data Research Network (HDRN) Canada	HDRN Canada brings together provincial, territorial and federal organisations which hold and manage data. HDRN Canada does not hold data. HDRN Canada is a network of provincial, territorial and federal data centres which hold data covering the entire Canadian population. Data covers topics including health, education, demographics, and immigration. Website: https://www.hdrn.ca/
Health Intervention and Technology Assessment Program - National Health Security Office (HITAP-NHSO), Thailand Ministry of Public Health	The Health Intervention and Technology Assessment Program (HITAP) and National Health Security Office (NHSO) worked together on a research data initiative in Thailand. This initiative was created to explore the impact of COVID-19 on the healthcare system such as health service utilisation. Data includes demographics, inpatient and outpatient information. Over 1 billion records are held in the NHSO database which represent health service utilisation record from a universal coverage scheme which provides health care to approximately 80% of Thai citizens. Website: https://www.hitap.net/research/179967
ICES (formerly the Institute for Clinical Evaluative Sciences, now using the name ICES in public communications)	The ICES Data Repository includes over 100 health and population-related data assets in which direct personal identifiers (such as name and health card number, if applicable) have been removed and replaced with a confidential code. Most holdings are administrative data at population-level, including decades of longitudinal health records for 21 million people who are, or have previously been, eligible for publicly funded health care services in Ontario, Canada. The ICES Data Repository also includes some data for publicly funded services outside of health, population survey, investigator-led research project data sets (e.g., clinical trial data sets), and registries that have been linked to ICES' core administrative data holdings. Website: https://www.ices.on.ca/

Continued

Appendix A: Continued

Data trust, Data repository, or Data collaboration	Description
Manitoba Centre for Health Policy (MCHP)	The Manitoba Centre for Health Policy (MCHP) is a population health data centre located at the University of Manitoba in Winnipeg, Canada. MCHP conducts research on health and the social determinants of health. MCHP holds population-level data for approximately 1.37 million people, as of 2022, derived from Manitoba Government departments, provincial laboratories, clinical programs, community and social outreach organisations, and Indigenous governance bodies. Website: https://umanitoba.ca/manitoba-centre-for-health-policy/
National Diabetes Repository	The National Diabetes Repository is a scalable secure analytics platform and multi-jurisdictional data repository. Through enhanced informatics, the repository applies privacy-aware methods to safely contain data and make it available to approved investigators for analytics. The data are de-identified and include primary care electronic medical records (e.g., Encounters, Prescriptions, Vital Stats) from the Canadian provinces of Ontario, Alberta, Manitoba, Newfoundland and Labrador, and Quebec. The data represent over 150,000 individuals from over 1,500 physicians. Website: https://repository.diabetesaction.ca/
New Brunswick Institute for Research Data and Training (NB-IRDT)	The NB-IRDT at the University of New Brunswick is a provincial data custodian that hosts and provides researcher access to linkable and prepared person-level data on health, education, labour market training, social assistance, immigration, and other related topics. Most of the datasets at NB-IRDT are population-level, providing information on almost all of New Brunswick's approximately 800,000 residents. Website: https://www.unb.ca/nbirdt/
Newfoundland and Labrador Centre for Health Information (NLCHI)	NLCHI is a crown corporation responsible for managing provincial health data and information assets and providing data access and health analytics services, whose staff also have expertise in data security and privacy. NLCHI holds population-level data for the province's roughly 520,000 residents. Website: https://www.nlchi.nl.ca/
Ontario Brain Institute – Brain CODE	The Ontario Brain Institute's Brain-CODE is a large-scale informatics platform that manages the acquisition and storage of multidimensional data collected from participants with a variety of brain disorders. Brain-CODE currently hosts data from over 20,000 participants with a wide array of data types, including clinical, neuroimaging and molecular. Website: https://www.braincode.ca/content/about-brain-code
Ontario Health Data Platform (OHDP)	The Ontario Health Data Platform (OHDP) is a highly secure high performance computing environment that was established by the government of Ontario support scientific investigation of COVID-19 by enabling private and secure record level linkage to personal health information. The OHDP is comprised of datasets containing personal health information collected from ICES and/or Ontario Health. The OHDP holds population-level data for the province of Ontario. Website: https://ohdp.ca
PHEMI Health DataLab	The PHEMI Health DataLab is a cloud-based system for privacy, security & governance. PHEMI has one central system to manage data stored in many locations (cloud or hybrid cloud), process data in many locations (cloud or hybrid cloud) while controlling access and governing data all in one place. Website: https://www.phemi.com/data-privacy-management/

Continued

Appendix A: Continued

Data trust, Data repository, or Data collaboration	Description
Population Data BC (PopData)	PopData is a data and education resource facilitating interdisciplinary research on the determinants of human health, well-being and development in British Columbia (BC), Canada. PopData supports access to population-level data for BC's approximately 5 million residents and from over 30 data sets from both federal and provincial sources. Website: https://www.popdata.bc.ca/
Primary care Ontario Practice-based Learning and Research Network (POPLAR)	POPLAR is an initiative of Ontario's six University Departments/Sections of Family Medicine and the Alliance for Healthier Communities. POPLAR securely collects and de-identifies electronic medical record (EMR) data to support practices in delivering optimal care across Ontario, and strengthen practice-based clinical research and quality improvement processes. Currently, over 1,000 family physicians are contributing EMR data for over 1.5 million patients to the POPLAR database. Website: https://www.poplarnetwork.ca/?
Rhode Island Ecosystem	The Rhode Island Ecosystem is an integrated data system that links data, at the person and family level, across various state agency and non-profit datasets to drive holistic improvements in the wellbeing of Rhode Island residents. The Rhode Island Ecosystem holds population-level data for the state's roughly 1 million residents. Website: https://eohhs.ri.gov/initiatives/data-ecosystem
ThinkData Works - External Data Catalog	ThinkData Works' Data Catalog is a scalable, cloud-agnostic end-to-end data management platform that helps organisations securely find, govern, and deliver data. With configurable data ingestion, active metadata management, role-based distribution and streaming integration to data science toolkits, the ThinkData Catalog ensures data preparedness for governance reporting and advanced analytics. With multi-cloud warehouse support and data virtualisation capabilities, the ThinkData Catalog lets clients access and collaborate on shared data while adhering to data residency requirements. The Catalog supports regional deployment on isolated infrastructure, letting users define where and what data is hosted on the platform. Website: https://www.thinkdataworks.com/platform/data-catalog

Appendix B: Main changes to min specs and reasons for those changes

Refined 2023 min spec (changes relative to the 2020 min spec are identified with italic font)	**Min spec published in 2020**	**Main reason(s) for change(s)**
Legal 1) The data trust, *data repository, or data collaboration* must fulfill all legal requirements including, *as required,* authority*(ies)* to collect, *retain, use, disclose, and/or destroy* data	The data trust must fulfill all legal requirements, including the authority to collect, share and hold data	"Authority" changed to "authorities" because multiple authorities may apply. Expanded the list of activities that there must be authority(ies) for. Replaced the term "share" with "disclose" to be inclusive of activities to make data available/accessible and/or release it to another organisation or person.
Governance 2a) The data trust, *data repository, or data collaboration* must have a stated purpose *that specifically addresses why its activities are necessary or beneficial*	The data trust must have a stated purpose	Original min spec was vague and could be fulfilled by any organisation/initiative regardless of its purpose. Revised min spec requires that a data-related purpose is either necessary (e.g., required by law) or producing some public benefit.
Governance 2b) The data trust, *data repository, or data collaboration* must have an accountable governance body *that is answerable for its decisions*	The data trust must have an accountable governing body	Added "answerable" to ensure that the governance body(ies) have mechanisms in place to respond to questions and concerns.
Governance 2c) The data trust, *data repository, or data collaboration* must be transparent *about its purpose, governance body membership, data holdings, policies regarding who has access to what data for what purposes, and other information that is requested*	The data trust must be transparent in its activities	Original min spec was too vague; responses often did not provide the information that data partners, stakeholders and members of the public would typically seek.
Governance 2d) The data trust, data repository, or data collaboration must acknowledge and respect Indigenous Data Sovereignty	Not applicable	New requirement added consistent with the right to self-determination under the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP).
Governance 2e) Governance must be adaptive *and responsive to risks, opportunities, and the concerns of stakeholders*	Governance must be adaptive	Added "adaptive" to ensure that, in addition to establishing proactive measures, the governance body(ies) has/have mechanisms in place to change. Added detail on what organisations/initiatives might need to adapt and respond to, i.e., response to risks, opportunities, questions, and concerns of stakeholders.
Management 3a) There must be well-defined policies, processes, and *procedures covering the entire data_lifecycle*	There must be well-defined policies and processes for the collection, storage, use and disclosure of data	Added "procedures" and "covering the data lifecycle" to make the min spec more technically complete.
Management 3b) There must be policies, processes, and/or *procedures* for *cybersecurity and* data protection safeguards which are reviewed and updated regularly	Policies and processes must include data protection safeguards which are reviewed and updated regularly	Added "procedures" to make the min spec more technically complete. Given public concerns about the topic, specifically included the term "cybersecurity" (though it could be considered part of data protection broadly).
Management 3c) There must be *policies*, processes *and*/or *procedures* to identify, assess, and manage risks *on an ongoing basis*	There must be an ongoing process to identify, assess and manage risks	Added policies and procedures to make the min spec more technically complete and to make it clear that risk management is an ongoing (vs. one-time) requirement.

Continued

Appendix C: Continued

Refined 2023 min spec (changes relative to the 2020 min spec are identified with italic font)	**Min spec published in 2020**	**Main reason(s) for change(s)**
Management 3d) There must be policies, processes and/or procedures to create and maintain metadata and data documentation which provides sufficient information for potential users to find, understand, use, and reuse data holdings	Not applicable	New requirement based on the practical observation that you cannot have a functional data trust, data repository, or data collaboration without metadata and other documentation about the data that it collects, retains, uses, discloses, and/or destroys.
Data Users 4a) Data users must complete *privacy and security* training before they access data	All data users must complete training before they access data	Specified that privacy and security training is mandatory. For consistency with other min specs language, removed the word "all".
Data Users 4b) Data users must *acknowledge* that there *may be* consequences for non-compliance	All data users must agree to a data user agreement that acknowledges that data use will be monitored and includes consequences for non-compliance	Changed the min spec to reflect the practice that many organisations have data users acknowledge terms vs. sign agreements. For consistency with other min specs language, removed the word "all". Changed text to read "may be" consequences vs. implying that there will always be consequences because depending on the severity and intentionality of the non-compliance, there may not be consequences. Removed text about the data user agreeing to monitoring/auditing policies/practices because it is understood that non-compliance would be identified by monitoring/ auditing and, in practice, monitoring/auditing policies and procedures are often described in other documents vs. data user agreements/terms.
Stakeholder & Public Engagement *5a) There must be ongoing engagement with stakeholders.*	There must be early and ongoing engagement with stakeholders including members of the public	Separated stakeholder and public engagement requirements to clarify that both are essential and because they are, by their nature, different.
Stakeholder & Public Engagement *5b) Stakeholder engagement must include ongoing engagement with members of the public*	There must be early and ongoing engagement with stakeholders including members of the public	Separated stakeholder and public engagement requirements to clarify that both are essential and because they are, by their nature, different.
Stakeholder & Public Engagement 5c) Where there is a reasonable expectation that specific subpopulations or groups would have a particular interest in, *and*/or would be affected by, an activity *or decision*, there must be direct engagement tailored for that subpopulation/group	Where there is a reasonable expectation that specific subpopulations or groups would have a particular interest in, or would be affected by, an activity of the data trust, there must be direct engagement tailored for that subpopulation/group	Added the requirement to engage and involve affected groups at key decision points.

Note: the revised min specs use the language "data trust, data repository, or data collaboration" instead of relying on an uncommon interpretation of the term "data trust."