



Making a Complex Data Repository Work: Nuts and Bolts

Leslie L. Roos

Research Centres-Innovations

- 1) Building on registry resources
- 2) Record linkage frequent
- 3) Cooperation across agencies
- 4) File complexity
- 5) Long-term perspective
- 6) Knowledge management

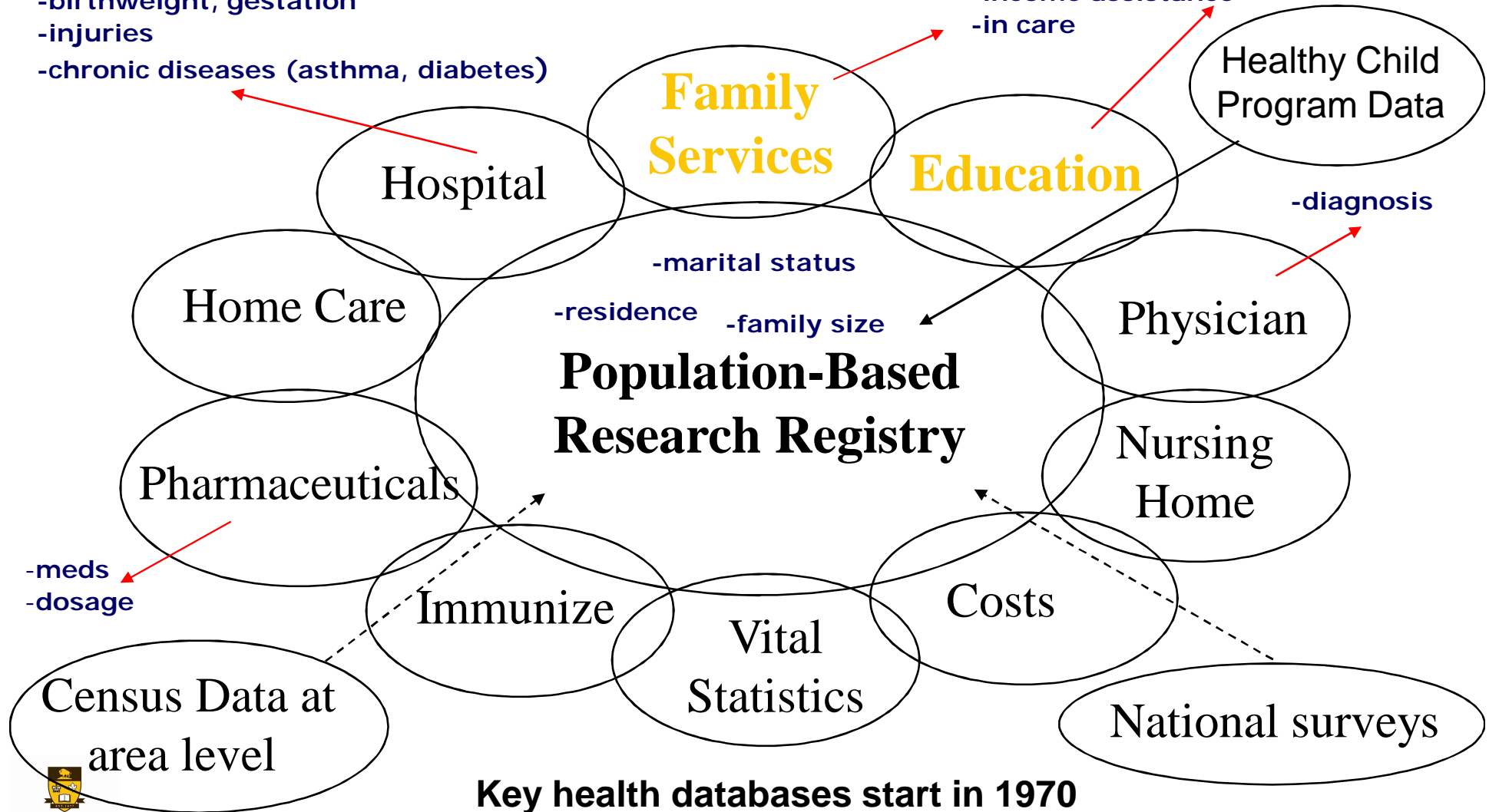


Manitoba Population Health Research Data Repository

- standards tests
- high school marks
- graduation
- retention

- birthweight, gestation
- injuries
- chronic diseases (asthma, diabetes)

- income assistance
- in care



Key health databases start in 1970

Informational Capital

- Great efforts to create ‘informational capital’
- Organize these efforts
- Preserve this capital: passage of time (memory loss), turnover of staff
- The more complex the data, the more documentation required

Different Perspectives

- Documentation via a data dictionary
- Working knowledge
- Open software
- Human/computer interface
- Variables and codes for each data set
- What is helpful for local users
- Let everyone read
- How do people get wanted information

Complexity

- Complexity increases with each file and with passage of time
- Each supplier has own standards
- Changes in government policy (ex. educational testing) and coding
- Data may be missing/different for some fields and some years
- Go beyond 'standard' data dictionary

Knowledge Management

- Knowledge management activities create readily retrievable information
- Informational capital can transform academic and policy-oriented research
- Local support, freely available to the international community (see MCHP website)

Three Main Sections

- Research Protocol: Critical steps in using Manitoba Repository
- Glossary: Terms commonly used
- Concept Dictionary: Detailed operational definitions of variables-choices outlined

Research Pillar

- Present rules for access, given privacy and confidentiality concerns
- Describe data sets
- Aid in funding applications
- Suggestions re. methods
- Effort to document by project's end

Repository Pillar

Describe data sets

Detailed organizational memory to find material after several years' hiatus.

- Reflections on what has and has not worked.
- Synthesis of multiple sources of information over time.

Knowledge Translation

- Help build “community of practice” despite other differences
- For teaching and sharing information with other centres and investigators
- New collaborators can use
- SAS videos
- Epidemiology Supercourse lectures

Knowledge Communities

- 1) Great diversity of disciplines and collaborators
- 2) Different types of data
- 3) Different modeling techniques
- 4) Knowledge sharing is essential



Manitoba Centre for Health Policy

- 1) Magnet for new data sets
- 2) Collaboration takes advantage of information-rich environment
- 3) Collaboration raises many issues (see Nature article)
- 4) Collaboration increases need for knowledge management

Example: Complexity in Collaboration re. Child Development (6 Disciplines)

- Local Investigators
 - Noralou Roos, Les Roos, Randy Fransoo
- Local (but resides in Ontario)
 - Marni Brownell
- U of T / OISE
 - Ben Levin
- U of T / School of Public Policy
 - Marc Stabile
- Columbia University
 - Janet Currie
- University of California, Berkeley and SF
 - Doug Jutte

Complexity/Change Imposed on Organization

1. Annual updates
2. Coding changes and new variables
3. Education—policy changes re. testing
4. Education—major curriculum changes
5. Intellectual advances (statistical techniques, research designs)
6. New deliverables—using data differently

Concepts

- 1) Provide the 'basics'
- 2) Date written
- 3) Clarify choices
- 4) What can and can not be done
- 5) Approach supports material used infrequently—the 'Long Tail'
- 6) References and cross-references

Recent Activities

- Research Protocol (to help new users)
- New Statistical Methods
- Educational Data
- Characterizing Families
- Change in Coding (ICD-9-CM to ICD-10-CA)

Unfinished Business

- Organization and Search
- Research protocol is structured but not currently searchable
- Users may ‘wander in’ to the Concept Dictionary when they want the protocol

Unfinished Business: Organize Concepts Better

- 1) Time/Longitudinal Studies: years, coverage, age, changes by year, follow-up
- 2) Place: geography, areas, neighborhoods, maps
- 3) Measurement: loss to follow up, building indices, small area measures
- 4) Family: siblings, parents, family structure and life events

Unfinished Business...

- 5) Topic: health, education, income assistance, well being
- 6) Variables: dependent and independent
- 7) Problems: an overview for each topic

Structure beyond key words? How to aid teaching?

Ongoing Issues

- Longitudinal studies are great but tricky; often several shifts in coding/reporting.
- Change in perspective for survey researchers used to 'cleaned up', well-maintained data sets.

Knowns and Unknowns

- Known problems
 - Approach by sensitivity testing
 - Example: Biases with inhospital collection of ER data and physician billing re. these data

Knowns and Unknowns

- Communication of Known Problems
 - Changing Formats over Time
 - Example: Variables miscalculated in complex cohort study using hospital and physician claims
- <Some of us knew but not others>

Knowns and Unknowns

- Unknown Problems
 - Blending information from different ministries
 - Example: Fiscal and Calendar year when health data underwent coding change at start of fiscal year

Possibility: Moving to Wikis

- “Wiki” Provides Framework and Software Allowing Groups to Communally Contribute
- Inherently Collaborative
- Facilitates Linking of Internal Web Pages and External References



Moving to Wikis

- Pages owned by Community of Volunteers
- Vested Interest in Accuracy
- Change can be Recorded and Vetted

Moving to Wikis

- Expertise brought together in a dynamic model:
- Easily cross-referenced
- Can be expanded as wanted

New Model from ‘WikiGenes’

- Hybrid of Traditional Scientific and Collaborative Dynamic Publishing
- Traditional—unambiguous authorship with credit acknowledged
- Multiple authorship—allows assembly of extensive knowledge and revisions

New Model

- Information can be easily reviewed and criticized
- Multiple perspectives can be easily integrated
- Authorship preserves community and individual interests



Possibility

- Seeking Funding for this New Distribution System
- Will Serve to Aggregate, Review, and Disseminate Knowledge with Involvement of a Larger Community
- Model of Collaborative Innovation

Ongoing Issues

- Costs -- How much is enough documentation? A long-term versus short-term perspective
- Cost-sharing of a public good: Project funding creates problem of 'free-riders', no contribution to documentation, complaints about costs

Ongoing Issues

- ‘Business model’ must find ongoing balance among:
- Bringing in new types of data/research
- Updates of data sets already held
- Documentation of all types
- Funding and new projects