Scandinavian Journal of Public Health, 2007; 35(Suppl 69): 21–25

informa
healthcare

ORIGINAL ARTICLE

An introduction to the General Temporal Data Model and the Structured Population Event History Register (SPEHR)

SAMUEL J. CLARK

Department of Sociology, University of Washington; Institute of Behavioral Science (IBS), University of Colorado at Boulder; MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, University of the Witwatersrand, South Africa

Abstract
There are some 37 demographic surveillance system sites active in sub-Saharan Africa, Asia and Central America. These sites, and other longitudinal population and health research projects, generate data over time in order to describe and explain the event histories of individuals and the populations they constitute. This note addresses key data management challenges presented by such complex temporal data-gathering efforts. Ideas supporting a standard definition for temporal population data, and a standard design for temporal databases to improve management of longitudinal population data, are presented and briefly discussed.

Key Words: *Longitudinal, data, GTDM, SPEHR, relational database, temporal data, data model*

The investigation of many public health questions in the developing world requires large population-based data sets collected over long periods of time. Intervention trials of many sorts – vaccine, behavioral modification, mosquito control, poverty alleviation, micronutrient supplementation, etc. – require the long-term study of well-defined populations so that events can be correctly sequenced and the effects of interventions can be properly identified and disentangled from possible confounding factors. Gathering, managing, and analyzing complex longitudinal data of this sort creates challenges on many levels, especially in the resource-poor settings where the data are needed most. Managing and analyzing the data to the best of their potential is one of the most significant challenges facing many long-term population-based studies in the developing world. Poor-quality data management results in corrupt data that are difficult to access and analyze, and this reduces the overall productivity of studies and makes it difficult to share the data or pool and compare it with data collected by other similar studies.

At the core of all longitudinal data-management systems is a *temporal database* that is able to store and manipulate the data collected by a longitudinal project – and it is often the case that this database is based on an idiosyncratic design that has evolved in an ad hoc fashion over many years. For individual projects this results in poorly functioning databases that allow complex inaccuracies and errors to accumulate in the data, and for the group of longitudinal projects as a whole the result is a collection of largely incompatible longitudinal databases whose data cannot be easily shared and analyzed together. This limits the usefulness of data collected by individual projects and largely denies the potential synergy available from being able to easily share, compare, and pool data from multiple longitudinal studies.

Solutions to both challenges are: (1) standard definitions for temporal data describing human populations and (2) a standard temporal database design for databases that store such information. Because there is such variety in the studies conducted both within and across individual longitudinal projects, both the standard data definition and the standard database design must be sufficiently general and flexible to define and manage a

Correspondence: Samuel J. Clark, Department of Sociology, University of Washington, USA. E-mail: samclark@u.washington.edu

ISSN 1403-4956 print/ISSN 1651-2553 online/07/010021-5 © 2007 Taylor & Francis
DOI: 10.1080/14034950701355452

wide variety of data and be able to easily accommodate changes to the overall set of data collected and managed by an individual project as time progresses. Standards meeting these criteria would enable individual projects to manage their data in a conceptually consistent, accurate, and well-documented way throughout their period of investigation, and this in turn would lead to greater accuracy and productivity and provide the potential to easily share data with and among other sites utilizing the same standards. This additional ability to easily share and pool data from multiple sites would increase the value of all of the data and lower the barriers to designing and implementing prospective multi-site studies.

Prominent examples of longitudinal health and population projects in the developing world that could benefit from standard temporal data-management tools are the demographic surveillance system (DSS) sites that collect, store, and manipulate large quantities of temporal data. Most of these sites are members of the INDEPTH Network [1] – an organization that is actively supporting a number of multi-site public health research initiatives, including the development of standard data management tools for longitudinal health and population surveillance systems in the developing world. DSS data are *temporal* because they describe the interrelated *histories* of the people, relationships and unions, households, villages and other entities that comprise the population being studied, and consequently time is an inherent dimension of the data. This fact makes the data complex and a challenge to manage because they must record the dynamic states of the entities they describe and all of the connections and relationships that are formed and broken between those entities. For example, an individual may be married three times during their lifetime and have a number of children within each of those unions; the index individual, their spouses, and all the children are related to each other in a complex set of formal and informal ties that are set up and broken by various events as time progresses. Many DSS sites faithfully capture and record the data that describe a set of interrelated histories like this, and consequently they must also store and manipulate those data and make them widely available for analysis in useful formats.

Significant effort has been invested to address the challenges posed by DSS data. The reference data model (RDM) [2] and the household registration system (HRS) [3–6] are the most prominent results of these efforts. The RDM is a relational database design that defines the various entities needed to represent a population of human beings, *specific* 'episodes' (durations when something is true), and *specific* 'events' that begin and end those episodes. The HRS is a DSS data management system built around the RDM and implemented in the Foxpro relational database management system. The RDM takes into account the temporal nature of DSS data by defining episodes and events and provides useful, consistent means through which to store and manipulate the temporal data collected by a DSS. Furthermore, because the HRS has been adopted by a number of DSS sites, the HRS and hence the RDM have become the de facto standard tools for storing and manipulating DSS data. This has advanced the situation significantly and made it easier to work with DSS data.

An important limitation of the RDM and HRS is the fact that the RDM defines *specific* entities, episodes and events, and *specific* relationships (connections) between episodes and events. This limits the information that the HRS can store and manipulate to exactly what is pre-defined by the RDM. When the requirements of a DSS differ from the pre-defined specification of the RDM, additions and modifications to the RDM and HRS are necessary. Implementing these changes can be both time consuming and costly and always leads to a slightly (or sometimes significantly) different DSS data-management system. The practical result is a collection of RDM/HRS-based DSS data-management systems that are in fact quite different and cannot easily share or compare the data that they manage, and this increases the effort necessary to conduct multi-DSS research projects that must pool and jointly analyze data. Ideas presented here begin to address the limitations of the RDM by providing a standard way of defining DSS data and a flexible, standard database design for DSS.

The general temporal data model (GTDM) is an explicitly temporal model of reality that is able to model both the complex structure and the temporal characteristics of data that describe long-lasting entities such as people, unions, or households. The GTDM is composed of three abstract components that work together in a general way to model the complex interrelated histories of real-world entities. *States* represent all that is constant and unchanging, *events* represent change and in particular the beginning and ending of all states, and *influences* form many-to-many connections between states and the events that influence them. When an event affects more than one state, the influences that represent those effects link all the affected states to the same event and thereby form an indirect connection between the affected states. In this way it is possible to represent the formation and dissolution of relationships between states, and the times when those changes took place are an inherent part of the representation of the relationships themselves. This three-part abstraction is able simultaneously to represent both the temporal aspects of reality and the dynamic structure of the connections that form and break between long-lasting entities as time progresses.

For example in the context of DSS, people, unions, households, villages and places are states because they persist through time with a constant manner of existing that both begins and ends; births, deaths, in-migrations, out-migrations, weddings and divorces are events because they all bring about some change and often initiate or terminate a state; and finally the various ways in which these events affect the states are the influences. A birth influences both the parents, the child itself (as the beginning of its life), the place where the birth takes place, the union in which the birth takes place, and potentially other family members such as already-living siblings. All of these influences and potentially more (or fewer) may be represented in a GTDM framework.

To illustrate, imagine we are interested in recording the vital events (births and deaths), marital histories, and migratory behavior of a small number of people living in Durban and Nairobi. We initiate a small study at time 0.5 and enroll the two locations, Durban and Nairobi; two people, Jabulani and Thembi; and because Jabulani and Thembi live in Durban, we initiate a residence for each at Durban. At time 3.0 a wedding occurs in Durban that joins Jabulani and Thembi as a couple and initiates their marital union. At time 6.5 in Durban Thembi gives birth to Zabela, and at time 9.0 we visit our "sites" and make an observation of all the existing entities enrolled in our study. At time 12.5 in Durban Jabulani dies and as a result of Jabulani's death Thembi and Zabela move from Durban to Nairobi at time 15.5. At time 17.8 we visit our "sites" again and make another observation of all the existing entities enrolled in our study, and the study continues to the present time, 19.0. Throughout the study we organize and record the information as states, events, and influences.

Figure 1 displays a diagram of the information we collect. Time is recorded on the horizontal axis and is marked with equidistant positional markers from 0 to 19. Horizontal (grey) lines represent states, vertical (black) lines represent events, and the shaded circles at the intersection of horizontal and vertical lines represent (potentially multiple) influences. States and

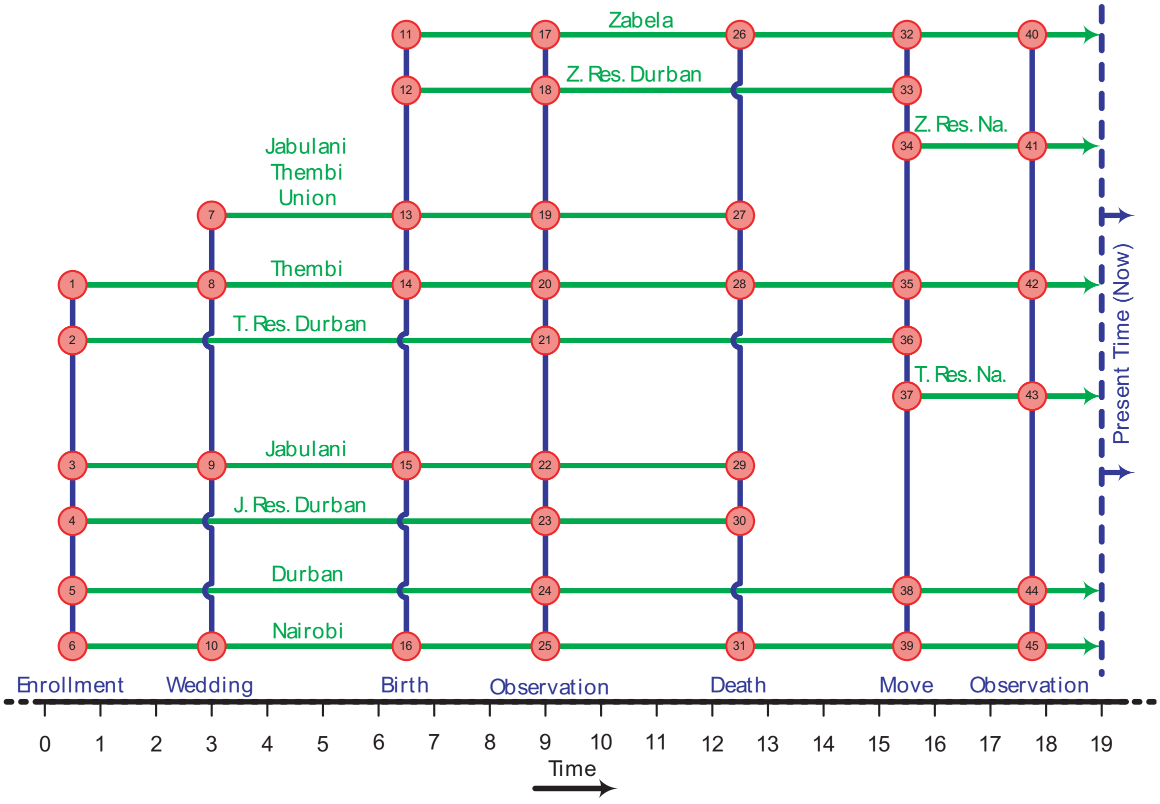

Figure 1. Diagram of Event ⇔ Influence ⇔ State example.

events are labeled with descriptive labels while intersections between states and events where influences occur are numbered. Sometimes there is no relevant connection between a state and an event, and in that case the lack of an influence is indicated with a "jump" (semi-circular intersection symbol) instead of a shaded circle.

The structured population event history register (SPEHR) is a relational database design (schema) that is based on the GTDM and contains three main tables: one containing all states, one containing all events, and one containing all influences. To implement the GTDM two additional components are necessary: (1) because many different *types* of state, event, and influence are stored in the three main tables, there must be a way of differentiating the different types of states, events, and influences stored in each of those tables; and (2) there must be a way of attaching type-specific descriptive information (attributes) to specific states and events stored in the states and events tables. Descriptions of the specific types of states and events that a SPEHR database can represent are stored in database tables that are separate from the main tables. This database structure allows the database to represent and store information about *new* types of things – states, events, and influences – without having to change the design or structure of the database. Adding a new type of state, for example a residence episode, involves entering its description in the table that holds the descriptions of the state types that the database recognizes. A similar process is necessary to add a new event. The state and event-specific "attributes" (descriptors) are stored in a separate set of tables that work in much the same way; there are tables to hold the actual attribute values themselves and a separate set of tables that hold descriptions of the possible attributes of different types of states and events. Each stored attribute has an associated attribute type, and if new attributes are required their descriptions are simply added to the attribute types tables, thereby allowing the database to represent and store new types of attributes. The data that describe the different types of states, events, influences, and attributes function in another important role in that they comprise a complete description of the contents of the database – a *data dictionary*. In this capacity they can be used to document and facilitate the sharing and comparing of the primary data that they describe.

In addition to having a complete built-in data dictionary, other significant benefits result from this abstract way of conceptualizing and storing longitudinal data. Most importantly, the structure and organization of the database do not need to be changed in order to allow it to store more information on entities that it already understands or to store information describing entirely new entities. This allows the database to be flexible and grow and customize itself to the current needs of the project as time goes on, all *without* having to make costly, complex, and time-consuming fundamental changes and additions to the design and implementation of the database. Building on the fact that SPEHR databases all share a standard structure and conceptual design, and the fact that each individual SPEHR database contains a full data dictionary in a standardized format that can be readily understood by other SPEHR databases, it is straightforward to pool, share, and compare data stored in SPEHR databases. This could potentially lower the barrier to conducting both retrospective and prospective multi-site investigations that must make extensive use of longitudinal data collected at a large number of different research sites.

Both the GTDM and SPEHR are described in detail elsewhere [7,8], and work in progress aims to develop a useful, working version of SPEHR. The Agincourt DSS site in South Africa is currently hosting and serving as test site for an effort aimed at developing and testing the software necessary to realize SPEHR. Toward the end of the software development effort a workshop is planned to present prototypes, gather feedback, and assess the need for training and capacity-building that may be necessary for other DSS sites to adopt and fully utilize SPEHR. In order to fully harness the benefits of SPEHR, including the ability to design and implement multi-site studies, it will be necessary for SPEHR to be widely used and for its users to be well trained, and this will require concerted efforts to both disseminate and install SPEHR and to train its users.

References

[1] INDEPTH Network. An International Network of Field Sites with Continuous Demographic Evaluation of Populations and Their Health in Developing Countries – INDEPTH. 2006. Available at: http://www.indepth-network.net; http://www.indepth-network.org (accessed 12 January 2006).

[2] Benzler J, Herbst K, MacLeod B, A Data Model for Demographic Surveillance Systems. 1998. Available at: http://www.indepth-network.org/publications/indepth_publications.htm (accessed 12 January 2006).

[3] MacLeod B, Computer Program: The Household Registration System. 2003. Available at: http://www.popcouncil.org/hrs/hrs.html.

[4] MacLeod BB, Phillips JF, Binka FN. Sustainable software technology transfer: The Household Registration System. In: Kent A, editor. Encyclopedia of library and information

science, vol. 58. New York: Marcel Dekker; 1996, pp. 302–10.

[5] Outlaw JAM. 2000. The Household Registration System: A point and click revolution in health and demographic research. Available at: http://www.popcouncil.org/pdfs/hrs_report.pdf (accessed 12 January 2006).

[6] Phillips JF, MacLeod B, Pence B. 2000. The Household Registration System: Computer software for rapid dissemination of demographic surveillance systems. Demographic Research, 2. Available at: http://www.demographic-research.org/Volumes/Vol2/6.

[7] Clark SJ, SPEHR: The Structured Population Event History Register. 2007. Available at: http://www.samclark.net/spehr/ (accessed May 5 2007).

[8] Clark SJ. A General Temporal Data Model and the Structured Population Event History Register. Demographic Research 2006;15(7):181–252. Available at: http://www.demographic-research.org/Volumes/Vol15/7/default.htm.