P1 P2 P3 P4 P5 P6 P7 P8

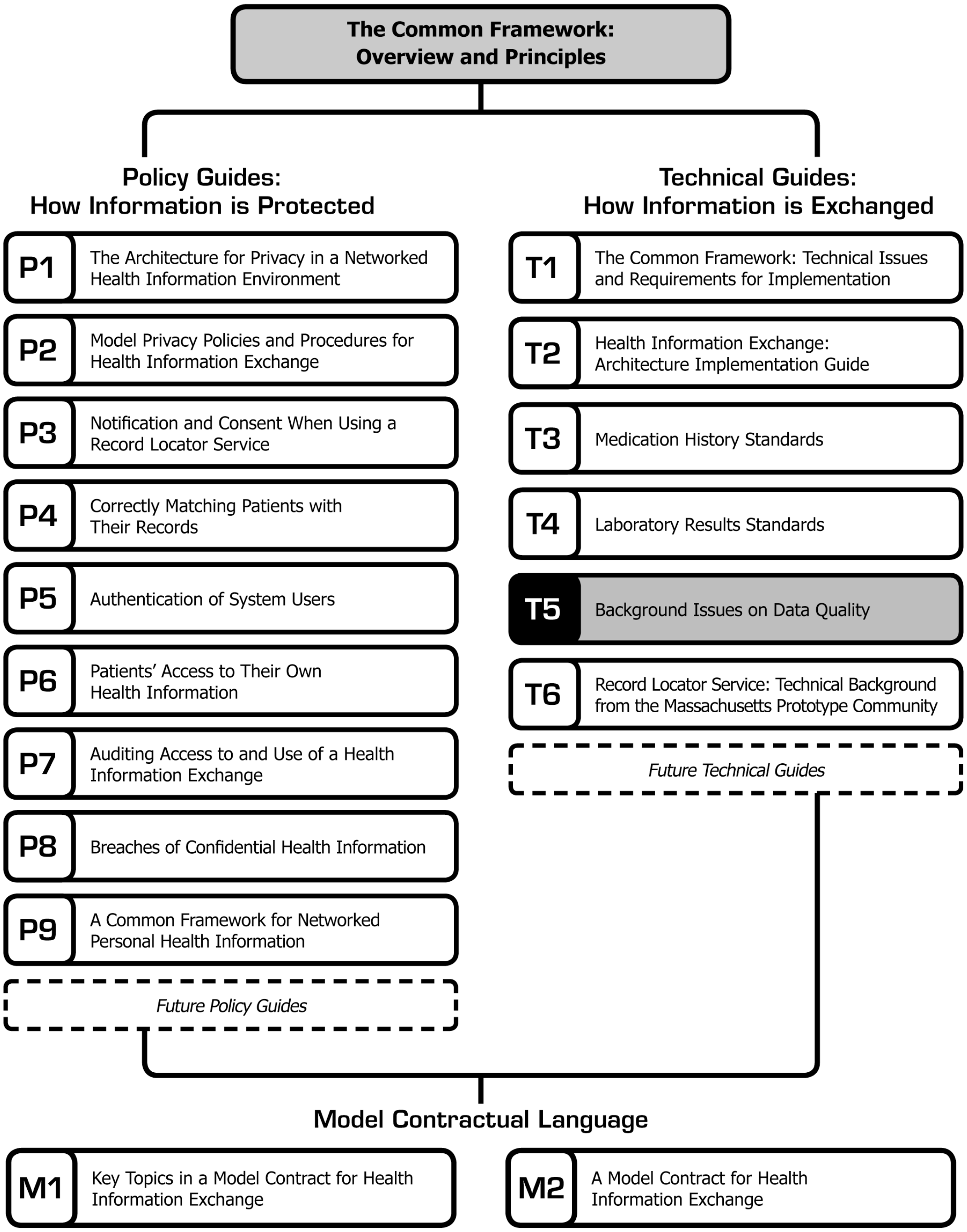
T1 T2 T3 T4 T5 T6 M1 M2

Background Issues on Data Quality

THE CONNECTING FOR HEALTH COMMON FRAMEWORK

Background Issues on Data Quality

The document you are reading is part of *The **Connecting for Health** Common Framework,* which is available in full and in its most current version at: http://www.connectingforhealth.org/. The Common Framework will be revised and expanded over time. As of October 2006, the Common Framework included the following published components:



The Common Framework: Overview and Principles

Policy Guides: How Information is Protected

- **P1** The Architecture for Privacy in a Networked Health Information Environment
- **P2** Model Privacy Policies and Procedures for Health Information Exchange
- **P3** Notification and Consent When Using a Record Locator Service
- **P4** Correctly Matching Patients with Their Records
- **P5** Authentication of System Users
- **P6** Patients' Access to Their Own Health Information
- **P7** Auditing Access to and Use of a Health Information Exchange
- **P8** Breaches of Confidential Health Information
- **P9** A Common Framework for Networked Personal Health Information
- *Future Policy Guides*

Technical Guides: How Information is Exchanged

- **T1** The Common Framework: Technical Issues and Requirements for Implementation
- **T2** Health Information Exchange: Architecture Implementation Guide
- **T3** Medication History Standards
- **T4** Laboratory Results Standards
- **T5** Background Issues on Data Quality
- **T6** Record Locator Service: Technical Background from the Massachusetts Prototype Community
- *Future Technical Guides*

Model Contractual Language

- **M1** Key Topics in a Model Contract for Health Information Exchange
- **M2** A Model Contract for Health Information Exchange

Background Issues on Data Quality*

Introduction

We live in an era of unprecedented data abundance and aggregation. The sheer variety of new information available on the Internet, in databases, and from other sources has changed the way we conduct business, undertake research, and communicate. Most of the changes are positive. Yet, increased reliance upon networked data has also introduced new challenges. One serious problem we need to address is that of "dirty data"—missing or inaccurate information that resides in (and, indeed, frequently results from) the abundance and aggregation of data in our lives today.

Dirty data can have several pernicious effects. In particular, it:

- Impacts the quality of care;
- Introduces privacy and other civil liberty concerns;
- Increases costs and inefficiencies;
- Creates liability risks; and
- Undermines the reliability and benefits of information technology (IT) investments, including the potential to streamline service delivery, accounting, and billing.

These concerns are particularly important in the medical field, where data problems represent the dark side of the tremendous potential offered by the adoption of health IT systems. In a "networked" medical setting, dirty data not only introduces economic inefficiencies; it may also cost lives. In addition, the lack of a data quality culture may be a core deterrent for many users in adopting and using health IT today.

As various regional and affinity-based information exchange networks around the country are developing and implementing strategies and architectures to link and share patients' data, the issue of dirty data will have to be addressed. Inaccurate patient data, especially if it affects the data fields used to establish individual patient identity through a Record Locator Service[1] (RLS), may be harmful if not mitigated from the outset. Dirty patient data has, for instance, the potential to undermine the matching capabilities of an RLS or to provide for an unacceptable level of false negatives. This document considers the growing need to develop a "data quality culture" at the network level and lists possible issues and options to consider.

* **Connecting for Health** thanks Stefaan Verhulst, Chief of Research, Markle Foundation, for drafting this paper.

[1] For a definition and description of a Record Locator Service (RLS), see http://www.connectingforhealth.org/assets/reports/linking_report_2_2005.pdf.

I. The Problem

By some estimates, the problem of dirty data in industry has reached epidemic proportions.[2] The problem is equally prevalent and potentially even more alarming in health care.[3]

In a medical setting, dirty data has several consequences:

First and foremost, it can lead to medical errors, which can kill or cause long-term damage to the health of patients. A widely noted 2000 Institute of Medicine report[4] estimates, for example, that between 44,000 and 98,000 lives are lost every year due to medical errors in hospitals alone, and that such errors result in an additional $17 to $29 billion in annual healthcare costs. Although not all these errors can be attributed to inaccurate data, a number of studies[5] have shown a link between poor quality data (in databases) and medical errors and subsequent poor quality of care. Further, in a "networked" health care setting, the challenge of data accuracy becomes even more critical because a health professional *immediately* uses the information accessible, especially in the case of an acute illness or emergency intervention, without any built-in step or potential to review its accuracy.[6]

Conversely, improving data quality can increase the quality of care by initiating a positive chain reaction—improving the data that clinicians see when the patient is admitted can validate the need for services to the patient, and if followed up with the provision of those identified services, may provide for better outcomes. A study on child mental health services, for instance, showed that 58 percent of the patients had improved outcomes after a data quality improvement project was instituted.[7]

Poor data quality can also reduce the accuracy of insurance bills. A study analyzing Medicare data found that 2.7 percent of the nearly 11.9 million records in the database, approximately 321,300 records, contained coding errors.[8] Such errors can impact the clinician's and/or the patient's insurance reimbursement and/or cause additional time to be spent correcting the errors. The study also identified the immediate benefits of addressing the errors. According to the Medicare study, the top 10 coding errors accounted for 70 percent of the total errors. By focusing on those 10 coding errors a high percentage of the problem can be addressed instantly, saving time and money.

Dirty data can also have serious consequences for patient privacy, especially in a networked environment. A single—and originally isolated—error in a data set can be magnified (and thus pose a

[2] Estimating the precise cost to business is difficult. According to Gartner, a consultancy, at least 25 percent of critical data stored by Fortune 1000 companies is inaccurate, and will continue to be so at least through 2007. In addition, The Data Warehousing Institute calculates that poor data quality costs US businesses more than $600 billion a year. See Leitheiser, Robert, "Data Quality in Health Care Data Warehouse Environments," Proceedings of the 34th Hawaii International Conference on System Sciences - 2001. p. 3. Available at: http://csdl2.computer.org/comp/proceedings/hicss/2001/0981/06/09816025.pdf. These are but a few of the figures to suggest the scale of the problem.

[3] See, for instance, Leitheiser, Robert, "Data Quality in Health Care Data Warehouse Environments," Proceedings of the 34th Hawaii International Conference on System Sciences - 2001. p. 3. Available at: http://csdl2.computer.org/comp/proceedings/hicss/2001/0981/06/09816025.pdf; Aronsky, Dominik, and Petr Haug, "Assessing the Quality of Clinical Data in a Computer-based Record for Calculating the Pneumonia Severity Index," *JAMIA*, 7:55-65; Seddon and Williams, "Data Quality in the Population-Based Cancer Registration: An Assessment of the Merseyside and Cheshire Cancer Registry," *Brit J Cancer*, 1997 76 (5):667–74; Barrie and Marsh, "Quality of Data in the Manchester Orthopaedic Database," *Br Med J*, 1992 304:159–62; Horbar and Leahy, "An Assessment of Data Quality in the Vermont-Oxford Trials Network Database," *Control Clin Trials*, 1995 16(1):51-61. [4] See IOM "To Err is Human: Building a Safer Health System," IOM Quality of Healthcare in America Committee 2000.

[5] See McDonald, Darryl, "Data Quality Management: Oft-Overlooked Key to Affordable, High Quality Patient Care," Whitepaper (7/17/2004) at HCT Project, Volume 2, p. 1. Available at: http://www.hctproject.com/documents.asp?grID=376&d_ID=2711; Smith, Peter et al., "Missing Clinical Information During Primary Care Visits," *JAMA*, Feb. 2 2005 (293:5), 565-571; Dovey, S. M. et al., "A Preliminary Taxonomy of Medical Errors in Family Practice," *Quality and Safety in Health Care* 2002 (11), 233-238.

[6] It is important to note that most clinicians are trained to expect errors in the data (and in the identity of the patient to whom the data belongs) and sometimes demand new tests or order additional tests to confirm a conclusion that might affect a significant clinical decision.

[7] The study also notes that clinicians saved 30 minutes a month immediately after a data quality improvement program was instituted, with trends showing that more time could be saved as time went on. See Nicholson, R. E. and Penney, D. R., "Quality Data Critical to Healthcare Decision Making," Presentation at the 2004 International Federation of Health Records Organizations (IFHRO) and Annual AHIMA Convention, 2004 AHIMA Convention Proceedings.

[8] Cottrell, Carl, "Medicare Data Study Spotlights Coding Errors," *Journal of AHIMA* 71, no. 8 (2000): 58-59.

more serious privacy risk) as it is "propagated" into various other data sets, systems and warehouses, while decreasing at each step the potential to redress the error.[9] On the other hand, a networked and aggregated data environment obviously undermines the "privacy by obscurity" paradigm that was often the sole privacy protection available in an off-line world.

While poor quality data can erode privacy, strong privacy protections can enhance the quality of data and subsequent health care, for example, by increasing trust and therefore increasing the amount of data that patients are willing to share with medical providers.[10] "Data accuracy" is therefore one of the nine principles underpinning "The **Connecting for Health** Architecture for Privacy in a Networked Health Information Environment."[11]

Despite the severity of the problem, the risks posed by dirty data often go unrecognized; in many ways, the problem of inaccurate data remains a low priority for companies and organizations.[12] It is critical to understand the problem and to develop strategies for minimizing data inaccuracies and the potential harm they cause.

II. Understanding Dirty Data: Definitions, Causes, and Locations

Data quality is broadly defined as "the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data."[13] Data accuracy is one of the "foundational features" that contribute to data quality[14] (along with other attributes such as timeliness, relevancy, representation, and accessibility[15]). In addition, data quality has two essential components: content (i.e., the information must be accurate), and form (i.e., the data must be stored and presented in a manner that makes it usable). These definitions are important to keep in mind when considering ways to minimize data inaccuracies, as they illustrate why the task of fixing dirty data requires more than merely providing "right" information.

Equally important when developing a strategy to increase data quality is identification of the underlying causes of "dirty data." Two broad categories of errors can be distinguished: systematic and random. Among the sources of systematic errors are: programming mistakes; bad definitions for data types or models; violations of rules established for data collection; poorly defined rules; and poor training. Random errors can be caused by: keying errors; data transcription problems; illegible handwriting; hardware failure (e.g., breakdown or corruption); and mistakes or deliberately misleading statements on the part of patients (or others) providing primary data. This is obviously not an exhaustive list, but a few examples of the types of errors that may occur. It is worth noting that according to the Data Warehousing Institute, 76 percent of all errors, across sectors and setting, result from "data entry." This suggests the critical role played by human error; many of the strategies proposed below, therefore, focus on reducing the likelihood of human error.

III. Strategies to Address Dirty Data: Towards a Data Quality Culture

To establish data quality within a health care setting and to prevent data quality errors in the system and limit their consequences, health care organizations should develop comprehensive strategies to establish

[9] See, for instance, Gibbs, Martin et al., "Data Quality, Database Fragmentation and Information Privacy," *Surveillance and Society*, 3(1): 45-58. Available at: http://www.surveillance-and-society.org/ Articles3(1)/data.pdf.

[10] Privacy concerns with regard to medical data often lead to "privacy protective behavior" that ranges from disclosing wrong or no information, to not seeking health care altogether. See **Connecting for Health**, "The **Connecting for Health** Architecture for Privacy in a Networked Health Information Environment."

[11] See **Connecting for Health**, "The **Connecting for Health** Architecture for Privacy in a Networked Health Information Environment."

[12] Multiple reasons why data quality problems are not addressed can be given. These range from "low awareness of the cost of data quality problems; tolerance for errors; to skepticism over ability to improve things and get returns." See, for instance, Olsen, Jack, *Data Quality: the Accuracy Dimension*. Morgan Kaufman Publishers, 2003. Page 13.

[13] Arts et al., op cit, p. 602. A similar definition is provided by Juran, who defines "data to be of high quality if they are fit for their intended uses in operations, decision making and planning" (Cited in Redman, *DM Review*, p. 2).

[14] Olsen, Jack, *Data Quality: the Accuracy Dimension*. Morgan Kaufman Publishers, 2003. Page 24

[15] Cited in Gendron, Michael et al., "Data Quality in the Healthcare Industry," *Data Quality*, September 2001 7:1, p. 1.

a data quality culture. Ideally, such strategies should be developed from the outset and be embedded in the design of any networked health information exchange system.

Organizations can use a variety of tools and techniques to increase the cleanliness of data, both at the time of collection and during subsequent processing.

For the purposes of **Connecting for Health**, data cleanliness efforts should be concentrated on those data elements required by the RLS. As the US moves towards widespread data standardization,[16] data input quality control can improve the usability and quality of data outputs. It should be noted that the documentation of a clinician *cannot*, by law, be changed retroactively, as this constitutes a change to the documented medical record of an individual; adding corrected information is allowed.

For cases in which data cleansing techniques[17] are applicable in health care, for example, *detection* (not resolution) of a single patient with two records, these techniques can be automated (e.g., in the form of software packages) or involve a human component (e.g., monitoring and training).

Ultimately, a well-thought-out and comprehensive data quality program should include both automated and human strategies, such as:

- *Standardize* data entry fields and processes for entering data[18];
- *Institute real-time quality checking*, including the use of validation and feedback loops[19];
- *Design data element to avoid errors* (for example, through the use of check digits and checking algorithms on numeric identifiers where human entry is involved and the use of well-designed user interfaces)[20];
- *Develop and adhere to guidelines* for documenting the care that was provided to the patient [21];
- *Review automated billing software;*
- *Build human capacity*, including training, awareness-building, and organizational change.

Each of these strategies will incur certain costs, but they are likely to be less expensive than addressing errors resulting from a system designed without data quality features. The US health care system has a unique window of opportunity to establish such an internal data quality culture when considering how to adopt health IT systems in the near future.

These "organizational strategies" should be complemented by external strategies, especially redress mechanisms, which encourage identification and correction of errors. Redress mechanisms are frequently built into laws and regulations, which, among other things, allow consumers to access and correct errors in personal information.

In the United States, legal systems for redress date back at least to the Fair Credit Reporting Act of 1970. In addition, redress is built into the Privacy Act of 1974, and the Health Insurance Portability and Accountability Act of 1996.

Common redress strategies include:

- Notice of a possible adverse decision using inaccurate data and the procedure for challenging it;

[16] For an interesting analysis of some attempts towards clinical data standards and the challenges for adoption see Kim, K., "Clinical Data Standards in Healthcare: Five Case Studies." Prepared for California HealthCare Foundation, July 2005. Available at: http://www.chcf.org/documents/ihealth/ClinicalDataStandardsInHealthCare.pdf.

[17] For a discussion, see Arts et al., op cit, and Leitheiser, op cit.

[18] See, for example, Teperi, J., "Multi Method Approach to the Assessment of Data Quality in the Finnish Medical Birth Registry," *J Epidemiol Community Health*, 1993, 47(3), p. 242–7; Gissler M, et al., "Data Quality After Restructuring a National Medical Registry," *Scand J Soc Med*. 1995, 23(1), p. 75-80.

[19] See, for example, de Lusignan, Simon, "Does Feedback Improve the Quality of Computerized Medical Records in Primary Care?" *JAMIA*, 2002, 9, p. 395-401; Porcheret, Mark, "Data Quality of General Practice Electronic Health Records: The Impact of a Program of Assessments, Feedback, and Training," *JAMIA*, 2004, 11, p. 78-86.

[20] For additional insights in the importance of data element design to prevent errors see Koppel, R., et al., "Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors." *JAMA*. 2005;293:1197-1203.

[21] AHIMA Coding Products and Services Team. "Managing and Improving Data Quality (Updated) (AHIMA Practice Brief)." *Journal of AHIMA* 74, no.7 (July/August 2003): 64A-C.

- Access to the information on which the decision is based, which is premised on the ability to trace information to its source for verification;
- Opportunity to correct erroneous information and an obligation to correct or delete information that is erroneous;
- Procedures for ensuring that erroneous information does not re-enter the system;
- Obligations on data furnishers to respond to requests for reconsideration of data and to take corrective action when justified; and
- Independent administrative or judicial review and enforcement.

IV. Creating a "Data Quality Culture": Implementation Issues to Consider

Implementing a data quality culture, as suggested above, poses various challenges. Without specifying the operational procedures that may be unique to each network design and RLS implementation, the following set of questions will need to be addressed:

Record Locator Service

- How do data quality concerns affect the RLS and clinical data exchange? What are the particular data quality problems likely to afflict the RLS, requiring RLS-specific interventions?
- How does the network deal with the integrity of data in the RLS itself? Who is responsible for these cleaning functions in the network?

Network versus Participants

- What are the expectations or requirements for each Participant vis-à-vis the Network with regard to sustaining a data quality culture?
- Are patients' rights to access their records, as they pertain to the RLS, provided by the Network centrally, or does each Network participant offer such a policy individually?
- Should the Network provide universal data audits on RLS fields across all participants, flag conflicts, and resolve them?
- Should Networks use common training modules and protocols across all participants to address human errors?
- What are the roles of the Network and its participants with regard to cleaning clinical data that is exchanged among participants?

Patient Empowerment

- How do patients communicate corrections through the entire system, rather than just to the first place they might identify dirty data?
- Are patients or Network participants allowed to complete partial matches by using the RLS to search across the system (on the patient's behalf) in the interest of improving data quality?
- What are reasonable procedures for rollover and rollback corrections across the system?

Acknowledgements

The members of the **Connecting for Health** Policy Subcommittee have accomplished an extraordinary task in less than a year's time—the development of an evolving piece of work that can serve as the core of nationwide health information exchange—the policy components of **The Common Framework**. During this time, we have been fortunate to work with respected experts in the fields of health, information technology, and privacy law, all of whom have contributed their time, energy, and expertise to a daunting enterprise. Our consultants and volunteers have worked long hours in meetings and conference calls to negotiate the intricacies of such issues as privacy, security, authentication, notification, and consent in health information exchange.We offer them our heartfelt thanks for taking on this journey with us, and look forward to the remaining work ahead.

In addition, we would like to offer special thanks to the volunteers and consultants who authored the initial drafts of this body of work—their hard work created a strong foundation upon which to focus the Subcommittee's deliberations:Stefaan Verhulst, Clay Shirky, Peter Swire, Gerry Hinkley, Allen Briskin, Marcy Wilder, William Braithwaite, and Janlori Goldman.

Finally, we must note that none of this work would have been possible without the leadership and inspiration of our co-chairs, William Braithwaite and Mark Frisse.They have led us with steady hands and determination of spirit.

Connecting for Health Policy Subcommittee

William Braithwaite, MD, eHealth Initiative, (Co-Chair)

Mark Frisse, MD, MBA, MSc, Vanderbilt Center for Better Health, (Co-Chair)

Laura Adams, Rhode Island Quality Institute

Phyllis Borzi, JD, George Washington University Medical Center

Susan Christensen*, JD, Agency for Healthcare Research and Quality, United States Department of Health and Human Services

Art Davidson, MD, MSHP, Denver Public Health

Mary Jo Deering*, PhD, National Cancer Institute/National Institutes of Health, United States Department of Health and Human Services

Jim Dempsey, JD, Center for Democracy and Technology

Hank Fanberg, Christus Health

Linda Fischetti*, RN, MS, Veterans Health Administration

Seth Foldy, MD, City of Milwaukee Health Department

Janlori Goldman, JD, Columbia College of Physicians and Surgeons

Ken Goodman, PhD, University of Miami

John Halamka, MD, CareGroup Healthcare System

Joseph Heyman, MD, American Medical Association

Gerry Hinkley, JD, Davis, Wright, Tremaine LLP

Charles Jaffe, MD, PhD, Intel Corporation

Jim Keese, Eastman Kodak Company

Linda Kloss, RHIA, CAE, American Health Information Management Association

Gil Kuperman, MD, PhD, New York-Presbyterian Hospital

Ned McCulloch, JD, IBM Corporation

Patrick McMahon, Microsoft Corporation

Omid Moghadam, Intel Corporation

Joyce Niland, PhD, City of Hope National Medical Center

Louise Novotny, Communication Workers of America

Michele O'Connor, MPA, RHIA, MPI Services Initiate

Victoria Prescott, JD, Regenstrief Institute for Healthcare

Marc A. Rodwin, JD, PhD, Suffolk University Law School

Kristen B. Rosati, JD, Coppersmith Gordon Schermer Owens & Nelson PLC

Sara Rosenbaum, JD, George Washington University Medical Center

David A. Ross, ScD, Public Health Informatics Institute

Clay Shirky, New York University (Chair, Technical Subcommittee)

Don Simborg, MD, American Medical Informatics Association

Michael Skinner, Santa Barbara Care Data Exchange

Joel Slackman, BlueCross/BlueShield Association

Peter P. Swire, JD, Moritz College of Law, Ohio State University

Paul Tang, MD, Palo Alto Medical Foundation

Micky Tripathi, Massachusetts eHealth Collaborative

Cynthia Wark*, CAPT, United States Public Health Service Commissioned Corps, Centers for Medicare and Medicaid Services, United States Department of Health and Human Services

John C. Wiesendanger, MHS, West Virginia Medical Institute/Quality Insights of Delaware/Quality Insights of Pennsylvania

Marcy Wilder, JD, Hogan & Hartson LLP

Scott Williams, MD, MPH, HealthInsight

Robert B. Williams, MD, MIS, Deloitte

Joy Wilson, National Conference of State Legislatures

Rochelle Woolley, RxHub

Amy Zimmerman-Levitan, MPH, Rhode Island State Department of Health

**Note: Federal employees participate in the Subcommittee but make no endorsement*