PaTH: towards a learning health system in the Mid-Atlantic region

Waqas Amin,¹ Fuchiang (Rich) Tsui,¹ Charles Borromeo,¹ Cynthia H Chuang,³ Jeremy U Espino,¹ Daniel Ford,⁴ Wenke Hwang,⁷ Wishwa Kapoor,² Harold Lehmann,⁴ G Daniel Martich,³ Sally Morton,⁸ Anuradha Paranjape,⁶ William Shirey,¹ Aaron Sorensen,⁶ Michael J Becich,¹ Rachel Hess,² the PaTH network team

ABSTRACT
The PaTH (University of Pittsburgh/UPMC, Penn State College of Medicine, Temple University Hospital, and Johns Hopkins University) clinical data research network initiative is a collaborative effort among four academic health centers in the Mid-Atlantic region. PaTH will provide robust infrastructure to conduct research, explore clinical outcomes, link with biospecimens, and improve methods for sharing and analyzing data across our diverse populations. Our disease foci are idiopathic pulmonary fibrosis, atrial fibrillation, and obesity. The four network sites have extensive experience in using data from electronic health records and have devised robust methods for patient outreach and recruitment. The network will adopt best practices by using the open-source data-sharing tool, Informatics for Integrating Biology and the Bedside (i2b2), at each site to enhance data sharing using centrally defined common data elements, and will use the Shared Health Research Information Network (SHRINE) for distributed queries across the network.

INTRODUCTION
Patient centered outcome and comparative effectiveness research has the capacity to transform the healthcare delivery system by identifying the most effective treatment, diagnostic techniques, and preventive measures. In an effort to achieve this goal, the Patient Centered Outcomes Research Institute (PCORI) developed PCORnet, a national patient centered outcomes research network that functions as a national network for conducting clinical research. To develop the key components of PCORnet, PCORI has approved awards to 29 health data networks and a coordinating center. PaTH (University of Pittsburgh/University of Pittsburgh Medical Center (UPMC), Penn State College of Medicine (PSCoM), Temple University Hospital (TUH), and Johns Hopkins University (JHU)) is one of the 11 clinical data research networks (CDRNs) that is participating in PCORnet.³

The CDRN effort employs informatics solutions to provide access to electronic health record (EHR) data and enable platforms for sharing data across multiple institutions. It allows the aggregation and analysis of distributed data, and facilitates patient centered, comparative effectiveness research. The PaTH network collaborative sites include community hospitals, academic institutes, and outpatient practices that provide healthcare to a diverse population of 2.5 million patients across the Mid-Atlantic region (figure 1).

PaTH will provide an informatics supported infrastructure for cohort identification and data sharing within the network of three targeted conditions: idiopathic pulmonary fibrosis (IPF), atrial fibrillation (AF), and obesity. PaTH will use the semantic standards that are already developed at each network site to meet federal mandates, such as Meaningful Use requirements. These include standardized vocabularies (SNOMED CT, LOINC, RxNORM, ICD9/10).2,4 To support the broadest possible data and cohort sharing, we chose the open-source i2b2 (Informatics for Integrating Biology and the Bedside)⁵ tool for syntactic interoperability. We will enhance SHRINE (Shared Health Research Information Network)⁶ to allow it to return de-identified patient level data in addition to cohort counts and to perform queries across the network.

PATH RESEARCH INFRASTRUCTURE
Network governance
The PaTH governance structure includes the PaTH Steering Committee (SC), three advisory committees (the health System Advisory Committee (HAC), the Patient and Community Advisory Committee (PAC), and the Clinician Advisory Committee (CAC)), and four working groups (the Research Question Group (RQG), Information Technology Group (ITG), Study Design and Methodology Group (SDMG), and Regulatory and Contracting Group (RCG)).

The SC is chaired by the network principal investigator and includes two representatives from each site, as well as the PaTH informatics lead. The SC monitors the overall operation of the network and, in consultation with the advisory committees and working groups, sets network policy.

Each of the four working groups has a SC liaison and representatives from the three advisory committees. The RQG develops and vets clinical research questions for PaTH CDRN. The ITG includes informatics leader from each network site and ensures that data mapping and integration support CDRN research. The SDMG includes two statisticians and methodology experts from each site and is responsible for advising and implementing best practices for the conduct of research within PaTH. The RCG will develop and implement the institutional review board pre-review and streamlined contracting processes.
Cohort identification
The network will characterize two populations in each of the three conditions. The first is the inclusive population of all identifiable patients with the disease condition in the EHR. The second is the consented cohort, which includes a smaller number of surveyed patients (1000 for each of AF and obesity, and 80% of the inclusive population for IPF; table 1).

Data collection
PaTH will acquire a complete set of longitudinal data from inpatient and outpatient settings from our EHR system including patient’s enrollment, demographics, diagnosis, procedure, laboratory, radiology, and ordering/dispensing medication. PaTH CDRN experts will define clinically relevant questions for three disease cohorts and identify the data needed to answer these questions. These data elements are vetted with our health systems’ informatics groups, as well as the PaTH core informatics group, in order to ensure that these elements are currently available, or will be available, for use in the PaTH network (table 2).

Engagement of patients and clinicians in PaTH
PaTH will implement PCORI methodology standards to approach patients, clinicians, and stakeholders. We will work with national organizations with an interest in our disease states both as community stakeholders and to help identify potential individual patient stakeholders. We will collaborate with our community/patient advisory boards to create educational programs for clinicians and patients and allow them to be active, assertive advocates on the team, and understand the privacy and confidentiality issues associated with research.

Collection of patient reported outcome information to support clinical trials
The PaTH network sites have extensive experience in collecting patient-reported outcomes (PROs) within the EHR, including health behaviors, symptoms, and functional status and using these data for clinical care. For example, UPMC uses Epic’s MyChart and Welcome questionnaires to collect PROs within several clinical settings. All PROs are recorded in the EHR during clinical care episodes and are also available for research.

Table 1  Estimated cohort size of three targeted conditions at PaTH network sites

<table>
<thead>
<tr>
<th>Prevalence of Idiopathic Pulmonary Fibrosis, Atrial Fibrillation and Obesity at PaTH</th>
<th>IPF</th>
<th>AF</th>
<th>Obesity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitt/UPMC</td>
<td>367</td>
<td>27,743</td>
<td>314,147</td>
</tr>
<tr>
<td>UPMC-HP</td>
<td>106</td>
<td>12,793</td>
<td>4,268</td>
</tr>
<tr>
<td>PSCom/HMC</td>
<td>70</td>
<td>&gt;10,000</td>
<td>69,804</td>
</tr>
<tr>
<td>TUH</td>
<td>71</td>
<td>13,537</td>
<td>22,447</td>
</tr>
<tr>
<td>JHU</td>
<td>&gt;200</td>
<td>58,104</td>
<td>104,799</td>
</tr>
</tbody>
</table>

Table 2  Data categories that are vetted by PaTH core informatics group, to ensure that these elements are currently available, or will be available, for use in PaTH, either through the Electronic Medical Record (EMR) or other data source

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Cohort</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IPF</td>
<td>AF</td>
</tr>
<tr>
<td>Patient-reported outcomes</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Demographic data</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Body mass index</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Prior hospitalization</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Medications</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Surgeries</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Other treatment</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Family history</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Social history</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Laboratory tests (e.g., Pulmonary function tests, ECG, echocardiogram, stress test, sleep study)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Radiology reports</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
use, JHU developed ‘Patient Viewpoint’ to collect PROs linked with the EHR from patients in oncology. Pitt/UPMC and JHU will serve as expert resources for the growth of these areas at the PSCoM/Hershey Medical Center (HMC) and TUGH sites, as well as participate actively in the PCORNet PRO effort.

Supporting large scale comparative effectiveness randomized trials

The PaTH network will compile relevant comparative effectiveness research focused data and outcomes which will be progressively embedded within clinical care pathways occurring at each of our institutions to create learning health systems. To build on the knowledge of the PaPATH investigators, interviews with principal investigators, project coordinators, clinical partners, and patient participants will be conducted to understand the facilitators and barriers to both successful and less successful research projects. Areas of particular interest include effective patient engagement, recruitment, and impact on clinical flow. We will compare the facilitators and barriers identified by our evaluation to those identified in the literature and compile them into a manual of best practices for use in PaPATH. This manual will be shared across the CDRNs and available on the PaPATH website. Implementation of these best practices will improve the capacity of PaPATH and other institutions to conduct large-scale comparative effectiveness randomized trials that are embedded within clinical care.

PATH INFORMATICS INFRASTRUCTURE

Design objective

The PaPATH informatics design is guided by three major principles: (1) the ability to perform exploratory cohort search across PaPATH network sites; (2) the use of de-identified data to analyze cohorts that meet a case definition; and (3) the capability to re-identify a patient enrolled in a randomized controlled trial or observational cohort for whom additional data are needed. To implement these principles, PaPATH will develop and/or implement the following key components of informatics infrastructure discussed below.

Standardized data elements and vocabularies

The ITG at Pitt under the guidance of PaPATH RQG and the SC will lead development of common data elements (CDEs), which will serve as a common information model for PaPATH. The standards (ISO/IEC—International Standards Organization/the International Electro-technical Commission) and vocabularies (SNOMED, RxNORM, ICD9/10, and LOINC) will then be used to code the CDE derived value sets. The PaPATH health systems have standardized their electronic medical record and ancillary systems on HL7 for healthcare messaging like LOINC for encoding laboratory tests, SNOMED, CPT, and ICD9/10 for encoding laboratory results, problems, diagnoses, and procedures, RxNORM for encoding medications, and Continuity of Care Document (CCD) for patient record export. The CDE specification provides metadata or data descriptors about the content, quality, condition, and other characteristics of the data and standard vocabularies to inform a shared information model and facilitate data sharing with the entire network. Each CDE is associated with an object or concept, attribute, and valid value and the representation.

Informatics for integrating biology and the bedside (i2b2)

To understand the heterogeneity of data in EHRs, the ITG adopted an existing solution, i2b2, which is already deployed at two network sites (PSCoM and JHU). i2b2 is an NIH-funded National Center for Biomedical Computing based at Harvard and it is widely adopted at Clinical and Translational Science Awardes (CTSAs), academic health centers, and industry. The i2b2 functions as an expandable clinical research database that provides a standard format (syntax) in which to represent clinical data, provides mechanisms for using standard vocabularies (semantics), and has existing tools (SHRINE) to perform centralized queries across multiple sites.

Shared Health Research Information Network

The PaPATH ITG will develop a modified version of SHRINE called SHRINE+ with the use of the well-accepted Agile software development method known as Scrum as well as industry standard development tools that include user stories, bug trackers, source-code management, automated builds, unit testing, continuous integration, and measurement of code coverage. SHRINE+ will incorporate authorization and audit mechanisms to ensure that each site retains adequate control and logs of their data. SHRINE+ will allow for the extraction of de-identified individual, patient-level data, in addition to cohort counts. The aggregate results of a SHRINE+ query will reside at the University of Pittsburgh’s Comparative Effectiveness Research Core Data Center (CERC-DC), which is specifically designed for multisite research collaborations that involve large clinical and administrative datasets such as those proposed in PaPATH.

Evaluation of PaPATH network informatics implementation

During the 18-month project period, and beyond, we will ensure adherence to the standards in PaPATH through manual quality assurance and automated rule-based validation. We will employ full-time central ETL and data quality engineers to work with personnel at each PaPATH site and verify that the syntax of data provided by the local i2b2 system conforms to PaPATH CDE specifications before these systems go live. In addition, we will employ routine, automated checks on all data imported to our centralized research data center to ensure that they conform to the CDE specification.

SUMMARY

The PaPATH network will adhere to best practices by using as its backbone open source tools (i2b2 and SHRINE) to aggregate data using standard vocabularies and provide distributed, de-identified cohort queries. PaPATH will test these systems in three targeted disease conditions. PaPATH will provide a robust informatics supported platform to facilitate comparative effectiveness research, support the conduct of clinical trials, and improve the decision making capability of both patients and physicians through a collaborative process that brings each partner closer to the ideals of a learning health system.

Author affiliations

1Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
Figure 2. Extraction, transformation, and loading (ETL) process within each site. The ETL process first extracts data based on common data elements (CDE) from source data. Then it de-identifies patient sensitive information and maps local codes to standard codes based on standard terminology systems—RxNorm, LOINC, SNOMED. Finally the ETL process loads the mapped data into the i2b2 data warehouse, where users can perform cohort queries through the i2b2 workbench.

Acknowledgements. We acknowledge all the PaTH network collaborative sites leadership, advisory committee, and working groups.


Contributors. WA prepared the first draft of this manuscript. RH, MJB, JE, and FT reviewed and revised the manuscript. All authors read and approved the final manuscript.

Funding. Patient Centered Outcome Research Institute (PCORI) supports this work; contract number CDRN-1306-04912.

Competing interests. None.

Provenance and peer review. Commissioned; externally peer reviewed.

Open Access. This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/.

REFERENCES.


