A centralized research data repository enhances retrospective outcomes research capacity: a case report

Gregory William Hruby,1,2,3 James McKiernan,1 Suzanne Bakken,2,3,4 Chunhua Weng2,3

ABSTRACT
This paper describes our considerations and methods for implementing an open-source centralized research data repository (CRDR) and reports its impact on retrospective outcomes research capacity in the urology department at Columbia University. We performed retrospective pretest and post-test analyses of user acceptance, workflow efficiency, and publication quantity and quality (measured by journal impact factor) before and after the implementation. The CRDR transformed the research workflow and enabled a new research model. During the pre- and post-test periods, the department’s average annual retrospective study publication rate was 11.5 and 25.6, respectively; the average publication impact score was 1.7 and 3.1, respectively. The new model was adopted by 62.5% (5/8) of the clinical scientists within the department. Additionally, four basic science researchers outside the department took advantage of the implemented model. The average proximate time required to complete a retrospective study decreased from 12 months before the implementation to <6 months after the implementation. Implementing a CRDR appears to be effective in enhancing the outcomes research capacity for one academic department.

INTRODUCTION
Retrospective outcomes research produces level III and IV evidence owing to its inherent limitations.1 Nevertheless, retrospective studies enable cost-effective hypothesis generation and play an important role in translational and clinical discovery2 3 and comparative effectiveness research4 5 by providing clinicians with health outcomes patterns.6 They have been used in various medical specialties, such as new-born health assessment,7 perioperative cardiac risk assessment,8 and prostate cancer risk stratification.9 10 By reusing rich clinical data from electronic health records (EHRs) and other databases, retrospective studies require no patient intervention defined a priori but reflect real-world physician decision-making scenarios and are less expensive than prospective randomized controlled clinical trials. With these advantages, retrospective studies have increasingly become a major source of contributions to the biomedical science literature.

Data consistency and completeness are critical to the scientific rigor of retrospective studies. Thus, it is important to employ a data management process that promotes data integrity.11 However, existing biomedical data management strategies in academic medicine are often non-intuitive to clinicians and scientifically prohibitive, expensive, and lack consistent institutional guidance.12 Underestimating the importance of data management can hinder data quality, impair research results, misinform clinical practice,11 or generate invalid hypotheses for new clinical trial designs.13

Retrospective outcomes research is an integral part of the academic mission of the department of urology at Columbia University Medical Center. For example, the PubMed query ‘2010 (dp) AND Columbia University (ad) AND urology’ returns a total of 50 articles: 28 retrospective studies (56%), nine literature reviews (18%), eight basic science studies (16%), and five clinical trials (10%). Since 2005, with the increasing availability of EHRs, retrospective studies have been rising steadily as a significant contributor to the departmental scholarly publications. This paper reports our experience in implementing and evaluating an open-source centralized research data repository (CRDR) for the urology department. The CRDR increases data access, user capacity, research productivity and quality, and enables a new research model and best practices for using electronic data for retrospective outcome research.

CRDR SELECTION AND IMPLEMENTATION PROCESSES
The need for a CRDR
In 2001, the department of urology at Columbia University established an initial simple research data management strategy to support outcomes research. It consisted of flat tables representing three areas of active research: prostatectomy, nephrectomy, and cystectomy procedures. Data extraction was distributed among project managers, who were medical students, residents and fellows from different research laboratories, and one shared data manager. This strategy proved inadequate to respond to the majority of research requests for patient cohort selection and dataset creation. Data redundancy was common across research datasets since many research participants had multiple comorbidities and hence were identified repeatedly by researchers studying different diseases.

Additionally, when project managers left the department, many ad hoc datasets were either physically lost or the user-defined data structure was incomprehensible to the new project manager. Furthermore, the ad hoc flat table design did not effectively represent the relationships of complex disease variables. For example, superficial bladder cancer consists of a convoluted treatment and surveillance cycle encompassing multiple clinical and pathologic staging events, surveillance and imaging procedures, and varying medical treatment algorithms. The tabular data solution was also not scalable to respond to the rapidly increasing data volume, variety, and velocity (ie, frequency of data generation and delivery).
Additionally, although the institution maintains an extensive central data warehouse (CDW), the CDW represents an aggregated clinical repository for the entire hospital. For our purpose, we needed to be able to make significant annotations to the clinical record for research uses, which is not supported by the institutional CDW. Even with the addition of business intelligence reporting tools, the CDW would not meet this need. The CRDR allowed the creation of structured, research variables to facilitate the annotation of unstructured clinical data from the CDW or the patient paper chart.

**CRDR selection**

We performed an initial investigation into potential data management strategies, including Microsoft Office Access and the broadly adopted REDCap (research electronic data capture) system. 14 Microsoft Access is a standard relational data management system but has limited scalability and capacity. REDCap offers flexibility in creating datasets and data form templates and centralizes the management of protected health information (PHI) on a secure platform that ensures data confidentiality and patient privacy. However, REDCap did not satisfy our needs for promoting data reuse and tracking data redundancy across overlapping study-specific datasets, and support for standardized terminology. Therefore, we decided that we needed a research data management system to facilitate standards-based data exchange with clinical data repositories.15 We sought a platform that offered web-based user access to patient research records that would comply with the Health Insurance Portability and Accountability Act (HIPAA)16 and meet the following requirements: be able to integrate into a federated data network, minimize data redundancy, be flexible to accommodate unique data points, allow users to refine patient research data based on inferences made on the clinical record, and deal with physician concerns over data ownership with a new model for data governance, including data rights and responsibilities.

CAISIS, an open-source, web-based data management system that integrates research data with patient care data,17 18 satisfied all these requirements. CAISIS provides a scalable infrastructure allowing flexibility as new user needs emerge. Compared with Access and REDCap, the CAISIS data structure offers a more complete integration with clinical information systems, thereby reducing data redundancy and improving data acquisition efficiency. These features, coupled with a web–user interface, provide an ideal solution of our needs. Therefore, CAISIS was implemented as the primary departmental research data management system and provided exclusive support for clinical research. Table 1 compares the three solutions specific to features we deemed ideal for our tasks.

**CRDR implementation**

Before we uploaded any patient data, the department consulted the Columbia University Medical Center information technology department to ensure compliance with HIPAA and institutional standards for the CAISIS application and hardware infrastructure. This process was completed through an initial assessment, followed by an evaluation with the chief security officer and a major university audit of all applications containing PHI.

The CRDR was populated with two sources of information, electronic billing records and the pre-existing research datasets. First, we prepared for uploading patient data billing records from January 2001 to June 2008, including the Current Procedural Terminology and International Classification of Diseases, ninth revision (ICD-9), clinical modifications codes. Specifically, we uploaded data for patient demographics, diagnosis, procedure, medical treatment, encounter, and comorbidities. Second, the research datasets were cross-referenced to identify and eliminate redundancy resulting from the previous individual data collection processes on a study-by-study basis, including many obscure variables that were not collected systematically and individual research records with an asymmetrical list of variables. Extensive effort was spent on mapping these datasets to the CRDR data structure.

We further established monthly-cached data updates to incorporate data from our faculty billing and scheduling system, GE IDX,19 to the CRDR. This data feed served as the baseline patient registry. Only finalized accounts from the previous month were imported. Open accounts were not imported until they were closed. We also obtained access to the institutional inpatient CDW and developed SQL import scripts for patient demographics, laboratory tests, and procedural, pathologic, and radiologic data. The CRDR was cross-referenced with the CDW to import any information not previously obtained. This data flow structure minimizes manual data entry and updates. Additionally, we requested access to outpatient EHR data, which are not integrated with the inpatient CDW, to further minimize manual data updates to the CRDR.

We implemented a strict access management policy by granting only one member of each outcomes research team access to the CRDR database for aggregated population-level queries. We deemed this necessary to better guard access to easily extract large amounts of PHI in a single setting. However, web-based patient-level browsing is available to all researchers through a separately controlled account activation/deactivation process, which employs a standard form requiring department approval before access to patient-level records is granted or removed.

**RESULTS**

We performed a retrospective analysis of the CRDR’s impact by comparing the research capacity of the department of urology during a pre-CRDR period (2005–8) and a post-CRDR period (2009–11). Our measurements included (1) user satisfaction and adoption; (2) workflow efficiency; (3) publication quantity; and (4) publication quality. A descriptive statistical analysis was performed on PubMed and the research request logs.

There was no campaign to require using the CRDR, but it gained great acceptance within the department. Currently, five out of eight (62.5%) clinician researchers and all four basic science researchers in the department use the CRDR. The three

<table>
<thead>
<tr>
<th>Desired features for supporting research</th>
<th>MS Access</th>
<th>REDCap</th>
<th>CAISIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provides structured data</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can add new variables</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Provides web interface for researchers</td>
<td>Yes, forms</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Generates chronological patient views</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Supports predefined data model</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Mirrors representation of clinical record</td>
<td>Yes, limited</td>
<td>No, snapshot study views</td>
<td>Yes</td>
</tr>
<tr>
<td>Limits data redundancy</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Supports standards-based data exchange</td>
<td>No</td>
<td>Yes, limited</td>
<td>Yes</td>
</tr>
</tbody>
</table>
clinical researchers not using this system investigate benign urologic disorders. These research clinicians have not rejected this model for retrospective research but have yet to be trained to use it since an adoption campaign had not been employed to introduce the CRDR to them.

The CRDR transformed research methodology within the department. With its implementation, data quality lays the foundation for outcomes research, and a more complete representation of every patient minimizes information bias. In contrast to the legacy method described earlier, it relies on automated prospective data collection from multiple institutional clinical information systems with minimized manual data entry and permits quality control at key research steps (ie, idea generation, feasibility assessment, statistics consultation, dataset generation, data analysis, team review, and manuscript preparation) through close collaboration among key personnel, most notably a principal investigator, an informatician, and a statistician. In this way, the research workflow has evolved from an ad hoc ‘analyze now, ask questions later’ process to a systematic one (figure 1).

Before the CRDR and the new research model were implemented, research capacity was limited to the number of researcher assistants available to the clinical researcher, and was indirectly related to data access in the old system. For example, only one assistant could access one dataset at a time, which formed a significant bottleneck for data access. The time to complete a project was 12 months. After the new model and CRDR had been implemented, five research assistants were hired and could access the datasets simultaneously, with more room remaining for additional personnel. The time to project completion also decreased to <6 months.

Figure 2 shows the distribution of the department’s research productivity by publication type between 2005 and 2011. Figure 3 depicts the department’s retrospective study output for the same time period. During the pre- and post-implementation periods, the department’s average annual retrospective study publication rate was 11.5 and 25.6 publications, respectively. The number of retrospective studies published using the CRDR exceeded that of the traditional methods. Figure 4 contrasts the publication impact factors of studies using and not using the CRDR. From 2005 to 2008, the department’s average journal impact rating steadily declined. In contrast, after the implementation of the CRDR (from 2009 to 2011), the average impact score of the papers published increased steadily from 1.7 to 3.1.

DISCUSSION
The implementation of the CRDR correlated with an improvement in both the quality and quantity of departmental publications on retrospective studies. We believe that the improvement resulted collectively from the quality research data in the CRDR, the new, efficient research workflow model enabled by the CRDR, and the expanded access to comprehensive research data for researchers. Through the CRDR, the department was
able to increase direct support for more researchers and standardize the retrospective research process using a consistent protocol. These contributions from the CRDR indirectly led to an increase in quality and quantity of publications in the department. As seen in figure 1, the CRDR also enhanced the communication and closed the feedback loop for all team members, thereby minimizing potential biases that might affect any retrospective study. The CRDR also minimized data collection bias through its interoperability with clinical systems and allowed researchers to focus their chart review efforts on more specific data elements. This enabled researchers to obtain a more complete and representative picture of the patient study population and allowed for the derivation of a more informed conclusion about this population.

Retrospective reviews and our observational study of the impact of a CRDR and the resulting research model have many of the same limitations. The successes rather than the failures of this implementation may be over-represented in the data presented. We assumed that the CRDR data are more accurate than the classic datasets that were in use. Studies have shown that research data can contain errors and misrepresentations at various levels. As such, this concern is valid but the new research model with the iterative, fine-grained data enrichment and validation process can mitigate the problem. Another possible source of bias was personnel changes over time during this study. Significant turnover was experienced in both the medical student and resident research rotations. It is true that the predisposing factors relative to the conduct of research among these cohorts varied. However, we did not observe random variability but a linear growth of research productivity. Finally, we primarily attribute the CRDR's indirect effect on academic productivity and quality to the infrastructural support provided for the research workflow process.

CONCLUSION

In this case report, we contribute empirical data showing how a CRDR increased a department's research capacity. The most rewarding outcome from the CRDR implementation was the emergence of a new research model that transformed the research workflow and improved research efficiency and quality on multiple levels. We hope our experience can be generalized to other academic medical departments that are interested in strengthening their retrospective outcomes research capacity.

Funding GWH is sponsored by a training grant from The National Library of Medicine 5T15LM007079. CW is sponsored by a CTSA award UL1 TR000040 from the National Center for Advancing Translational Sciences (NCATS).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES


