Three decades of utilising an integrated information infrastructure for outcomes research: a systematic review of studies conducted at the Regenstrief Institute

Brian E. Dixon*,†, Elizabeth C. Whipple§, John M. Lajiness*,# & Michael D. Murray||
*Richard M. Fairbanks School of Public Health at IUPUI, Indianapolis, IN, USA, †Regenstrief Institute, Inc., Indianapolis, IN, USA, ‡Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA Medical Center, Indianapolis, IN, USA, §Ruth Lilly Medical Library, Indiana University School of Medicine, Indianapolis, IN, USA, #Eli Lilly and Company, Indianapolis, IN, USA, and ||Regenstrief Institute and Purdue University, Indianapolis, IN, USA

Abstract

Objective: To explore the ability of an integrated health information infrastructure to support outcomes research.

Methods: A systematic review of articles published from 1983 to 2012 by Regenstrief Institute investigators using data from an integrated electronic health record infrastructure involving multiple provider organisations was performed. Articles were independently assessed and classified by study design, disease and other metadata including bibliometrics.

Results: A total of 190 articles were identified. Diseases included cognitive, (16) cardiovascular, (16) infectious, (15) chronic illness (14) and cancer (12). Publications grew steadily (26 in the first decade vs. 100 in the last) as did the number of investigators (from 15 in 1983 to 62 in 2012). The proportion of articles involving non-Regenstrief authors also expanded from 54% in the first decade to 72% in the last decade. During this period, the infrastructure grew from a single health system into a health information exchange network covering more than 6 million patients. Analysis of journal and article metrics reveals high impact for clinical trials and comparative effectiveness research studies that utilised data available in the integrated infrastructure.

Discussion: Integrated information infrastructures support growth in high quality observational studies and diverse collaboration consistent with the goals for the learning health system. More recent publications demonstrate growing external collaborations facilitated by greater access to the infrastructure and improved opportunities to study broader disease and health outcomes.

Conclusions: Integrated information infrastructures can stimulate learning from electronic data captured during routine clinical care but require time and collaboration to reach full potential.

Keywords: bibliometrics, comparative study; effectiveness, informatics, medical, information systems; research networks
Key messages

- Integrated information infrastructures can support robust, high quality observational data sets suitable for research and measurement of evidence-based practice.
- Integrated information infrastructures support the goals of a flexible, learning health system capable of continuous quality improvement.
- Health information professionals play critical roles in developing and improving the performance of integrated information infrastructures.
- Funding agencies that support the development of integrated information infrastructures need to continue investment in research and evaluation of these infrastructures impact on both knowledge generation and population health outcomes.

Background and significance

The term ‘big data’ in health care is concerned with analysing clinically relevant data sets that are too big (e.g. numerous in size), too fast (e.g. exponential growth) and too complex (e.g. heterogeneous, highly variable data of poor quality) for health care providers to process and interpret with existing tools. Numerous efforts seek to leverage ‘big data’ to improve the quality, safety and efficiency of the health care system. Chief among these is the learning health system (LHS) from the Institute of Medicine in the United States. In an LHS, both patients and providers have access to timely, accurate and comprehensive health information to support effective and efficient delivery of health services and achievement of health outcomes. Recent growth in the adoption of electronic health record (EHR) systems lays the foundation for an information infrastructure capable of supporting the LHS.

A robust information infrastructure to support an LHS requires the presence of (i) people, a skilled workforce to create and manage the infrastructure; (ii) organisations, entities that create and enforce policies governing the infrastructure; and (iii) information and communications technologies to capture, store, manage, analyse and exchange data in support of the learning health system. A number of organisations and professionals in the health care and informatics communities are contributing to the development of an information infrastructure to support the LHS in the United States. These include the U.S. Centers for Disease Control and Prevention’s (CDC) Vaccine Safety Datalink, the Food and Drug Administration’s Mini-Sentinel, the National Cancer Institute’s Cancer Research Network, the HMO Research Network, the 60 National Center for Advancing Translational Sciences’ Clinical and Translational Science Awards and the Patient-Centered Outcomes Research Institute’s National Patient-Centered Clinical Research Network. The list further includes the 11 Agency for Health care Research and Quality’s (AHRQ) Infrastructure Projects, such as the Scalable Architecture for Federated Translational Inquiries Network, the Comparative Effectiveness Research and Translation Network and the SUrveillance, PREvention, and ManagEment of Diabetes Mellitus. Similar efforts exist outside the United States, such as the Electronic Health Records for Clinical Research project in Europe.

While all of these efforts are making progress, the road to achieving the vision of a LHS is long. Many of the initiatives are less than a decade old with early evaluations of these infrastructures just now beginning to emerge. For example, at Columbia University, Hruby et al. describe the initial evaluation of a centralised data repository for supporting retrospective data analysis, which is part of its CTSA programme. After 3 years, the authors noted an increase in the urology department’s capacity to conduct retrospective data analysis based on the number and quality of publications produced by department investigators. Such results are promising, but 3 years is a short time frame. It is unclear whether an information...
infrastructure is capable of sustaining growth in retrospective analyses, such as comparative effectiveness research (CER), over an extended period of time.

For over four decades, the Regenstrief Institute (RI), based in the United States, has conducted health services and biomedical informatics research to improve health care delivery and outcomes. Viewed as a biomedical informatics pioneer,25 Regenstrief initiated the development of its information infrastructure in 1972 with a computerised patient management system for outpatient diabetes care.26 The system in the diabetes clinic quickly expanded to other parts of Wishard Memorial Hospital (renamed Eskenazi Hospital in 2013), integrating data from the laboratory, radiology and various other information systems. The integrated system became known as the Regenstrief Medical Record System (RMRS) and has served as the foundation for numerous informatics innovations, including the Medical Gopher order entry system.26–28 In addition, the RMRS served as an information infrastructure to conduct observational studies of health services29 as well as clinical trials.30–32 In the 1990s, the RMRS was expanded into the Indiana Network for Patient Care (INPC), which integrates data from large health systems, community hospitals, ambulatory practices, radiology centres and long term post-acute care facilities.33–35 More than 80 research scientists affiliated with RI and the Indiana University School of Medicine have utilised the RMRS and/or INPC to conduct both retrospective and prospective studies on health services and population outcomes during Regenstrief’s first 45 years.

To better understand the evolution and utilisation of the RMRS-now-INPC information infrastructure to support the kind of data analysis envisioned in an LHS over multiple decades, we conducted a systematic review of studies involving data from either RMRS or INPC in which health services or outcomes were evaluated. In this article, we describe the methods used to select and categorise the literature describing such studies, the patterns and trends in studies utilising Regenstrief’s infrastructure and the impact of papers and authors on their respective fields. We further discuss the potential of such integrated infrastructures, like the INPC, to achieve the LHS vision of enabling highly collaborative teams to rapidly study and improve health care delivery.

Materials and methods

Study design

We sought to identify peer-reviewed articles that document utilisation of data from the Regenstrief Institute’s information infrastructure to examine health services or outcomes. Independent searches by BED and ECW were performed in July 2013 of the English-language indexed and non-indexed literature in MEDLINE (1983 to June 2013) using a discrete set of terms given a fairly narrow target. Keywords included, but were not limited to, the following: Regenstrief, Regenstrief Institute, Regenstrief Medical Records System (RMRS) and Indiana Network for Patient Care (INPC). All fields were searched, including institution and abstract, to broadly identify articles published by Regenstrief investigators who are the principal users of the repository. Searches were supplemented by reviewing available curricula vitae and soliciting self-nominations from past and present Regenstrief scientists.

Eligibility criteria were used to identify articles that analysed health outcomes as part of a research study actually using data from the repository, as opposed to articles that principally described technical infrastructure or enhancements to the repository or simply used the repository to identify a study cohort. Therefore, the main inclusion criteria focused on the Methods section of the articles. Reviewers examined whether and how the data from either the RMRS or INPC were utilised by the authors. System development and other technical papers, such as how patient records are matched, were excluded. Papers pertaining to the AMPATH Medical Record System (AMRS) and Open Medical Record System (OpenMRS), technical infrastructures modelled on the RMRS but deployed outside Indiana, were also excluded.

Methods for data acquisition and measurement

Eligibility assessment was performed independently using an unblinded, standardised
approach by two reviewers (ET and WH; see Acknowledgements). Disagreements between reviewers were resolved by consensus involving discussion with the lead author (BED). Reviewers screened titles, abstracts and the full text of candidate articles. Full text review was necessary for all articles because reviewers were unable to confirm data from the RMRS or INPC were used for analysis and not simply for cohort selection given the information in the abstract. A third independent reviewer checked a randomised subset of candidate articles (20%) following initial classification for quality assurance.

Reviewers classified papers along several dimensions, including study design, cohort size, the diseases of interest and funding sources. Authors and their affiliations were also categorised. Reviewers were asked to classify papers based on whether the studies involved non-Regenstrief collaborators as well as collaborators outside the United States. These classifications enable analysis of data use over time beyond the health services research (HSR) laboratories within the Regenstrief and IU School of Medicine facilities. A full list of the classifications and discrete values available to reviewers is included as Appendix A.

Methods for data analysis

First, we examined study characteristics as a whole over the three decades, including total number of relevant articles, cohort size and publication type. We also examined the set of articles for patterns based on disease, funding source and non-RI collaborator involvement. We further used a variety of bibliometric analyses to examine article and author influence on various bodies of knowledge outlined below (e.g. geriatrics, medical informatics) given the areas of focus within RI’s centres of excellence.

Our bibliometric analysis examined journal impact factors (IFs), author impact ($h$-index) and article citation patterns. The journal impact factor (IF) approximates the average number of citations to recent articles published in a given journal. Journal IFs are frequently used as a proxy measure for the relative importance of a journal within its field. We extracted IF data from the 2012 edition of the Journal Citations Report (JCR) published by Thomson Reuters. The $h$-index attempts to measure both the productivity and impact of a given scholar. For our analysis, we focused on first and last authors, author positions that carry special weight in biomedical publications, and accessed $h$-index values from Scopus. Finally, we examined the number of citations (until July 2013) to each article using the union of citation data from both Thomson Reuters’ Web of Science and Google Scholar.

Results

Study selection

We identified 692 candidate articles by searching MEDLINE. Vita and self-nominations from Regenstrief scientists yielded 106 candidate articles. After combining data sets and removing duplicates, we identified a total of 666 candidate articles. Screeners then reviewed the candidate articles, applying eligibility and exclusion criteria. This process narrowed the results to 190 final articles. The article selection process is summarised in Figure 1. The full list of articles can be found in Appendix B.

Study characteristics

The final set of articles represents a heterogeneous group of studies involving Regenstrief investigators over three decades. Cohort sizes ranged from one to 800,000. Methodology ranged from case studies involving detailed review of EHR data to randomised control trials (RCTs) to comparative effectiveness research (CER) studies. The number of articles per decade increased, doubling between the first and second decade and nearly doubling again in the third decade. The diseases of interest varied yet clustered on four categories: cognitive, cardiovascular, infectious and chronic conditions. Acknowledgement sections in each paper highlight a mix of U.S. Government agencies along with grants and contracts from foundations and industry. Table 1 summarises article characteristics.
Temporal analysis

The number of Regenstrief publications analysing data from the RMRS or INPC steadily grows over time. Figure 2 summarises the number of selected publications each year during the three decades and displays the ratio of selected publications divided by the total number of Regenstrief publications in a given year. This ratio ranges from 20% to 100% with an average of 33.8%, meaning that on average about one-third of all Regenstrief publications across the three decades utilise electronic health record data from either the RMRS or INPC. Initially, in the first two decades, this ratio was around 40%. In the final decade, the ratio falls to around 25%.

Impact factor analysis

Impact factors were available for journals associated with 166 (87.4%) selected articles. The mean IF across these journals was 6.069; the median IF was 3.278. Comparison of journal IFs is most prudent within a subject category. We observed four main subject categories across the 166 journals: Medicine, General & Internal; Health Care Sciences & Services; Geriatrics & Gerontology; and Medical Informatics. The number of articles in journals matching these categories is summarised in Table 2. The table further provides the mean IF across each category for the selected articles matching that category, as well as the aggregate IF calculated by JCR across all journals in the subject category for reference. On average, articles utilising RMRS or INPC data for outcomes analysis were published in journals with IF values above the category aggregate IF.

Citation analysis

At least one citation existed for 186 (97.9%) of the selected articles. Citation counts ranged from 1 to 549 with an average citation count of 58.4, median citation count of 25 and a standard deviation of 84.4. There were 32 (17.2%) articles with a citation count greater than 100. These articles appeared in journals with high IFs across medicine and medical subspecialties, such as

© 2015 Health Libraries Group
Health Information & Libraries Journal
The article with the highest citation is a CER study comparing two screening methods for colorectal cancer. The remaining articles in this top group were generally either CER studies comparing alternative treatments or effectiveness RCTs across a broad range of medical subspecialties (e.g. endocrinology, rheumatology, gerontology, informatics) and health services domains (e.g. prevention, medication adherence, diagnostic screening).

Author analysis

Of the total articles examined, 122 (64.2%) included at least one non-Regenstrief author. During the first decade (1983–1992), 14 (53.8%) articles included at least one non-Regenstrief author. In the second decade (1993–2002), 36 (56.3%) articles involved non-Regenstrief authors. In the final decade (2003–2012), 72 (72%) articles included at least one non-Regenstrief author. Articles with only Regenstrief authors were cited 44.75 times on average, whereas those articles with non-Regenstrief authors were cited 64.13 times on average.

We further analysed first and last authors, author positions that carry special weight in biomedical publications, across all articles. Of all unique first and last authors (N = 145), 87 (60.0%) authors had been or currently serve as a research scientist at the Regenstrief Institute. The average h-index across first and last authors was 17 with a median of 13 and standard deviation of 16. Similar values exist when examining just first authors. H-index values for senior authors were higher, with the average h-index across just last authors (N = 78) equal to 20 with a median of 17 and standard deviation of 17.

Discussion

We performed a systematic review of articles over three decades (1983–2012) published by Regenstrief scientists and their collaborators in which data from either the RMRS or INPC were utilised to evaluate health services, patient outcomes or population outcomes. Of a total of 666 articles involving Regenstrief scientists during the three decades, 190 (28.5%) articles met the inclusion criteria. As anticipated, we observed growth in publication volume in almost every year along with diversity among study designs, funding sources, diseases of interest and authors. Articles were published in journals belonging to four

| Table 1 | Article characteristics including decade of publication, disease focus and funding sources for studies using data from the Regenstrief Institute between 1983 and 2012 |
|---------------------------------------------------------------|
| Characteristics | # Articles |
| Decade of Publication | |
| 1983–1992 | 26 |
| 1993–2002 | 64 |
| 2003–2012 | 100 |
| Disease Focus | |
| Cognitive Diseases: Alzheimer’s disease, dementia, delirium | 16 |
| Cardiovascular Diseases: heart disease, stroke, hypertension | 16 |
| Infectious Diseases: chlamydia, influenza, syphilis, MRSA | 15 |
| Chronic Diseases: asthma, diabetes, obesity | 14 |
| Cancer | 12 |
| Other: anaemia, spinal stenosis, osteoporosis, hyperkalaemia | 10 |
| Mental Health: depression, schizophrenia, affective disorders | 8 |
| Renal Diseases: renal failure, renal insufficiency | 5 |
| Rheumatoid Diseases: arthritis, osteoarthritis | 5 |
| Funding Sources | |
| U.S. National Institute for Aging (NIA) | 36 |
| U.S. Agency for Health care Research and Quality (AHRQ) | 31 |
| U.S. National Institutes of Health (NIH) excluding NIA and NLM | 24 |
| Foundations | 23 |
| Industry | 14 |
| U.S. National Library of Medicine (NLM) | 12 |
| U.S. Department of Veterans Affairs (VA) | 10 |
| U.S. Health Resources and Services Administration (HRSA) | 7 |
| U.S. Centers for Disease Control and Prevention (CDC) | 6 |
| U.S. Centers for Medicare and Medicaid Services (CMS) | 2 |
| U.S. Health and Human Services Agency (HHS) | 2 |
| State Government | 1 |

JAMA (N = 6; IF = 29.978), NEJM (N = 4; IF = 51.658), Annals of Internal Medicine (N = 4; IF = 13.976), Journal of the American Geriatrics Society (N = 3; IF = 3.978) and JAMIA (N = 1; IF = 3.571). The article with the highest citation is a CER study comparing two screening methods for colorectal cancer. The remaining articles in this top group were generally either CER studies comparing alternative treatments or effectiveness

© 2015 Health Libraries Group
Health Information & Libraries Journal
primary classifications from the JCR – MEDICINE, GENERAL & INTERNAL; HEALTH CARE SCIENCES & SERVICES; GERIATRICS & GERONTOLOGY; and MEDICAL INFORMATICS – reflecting the major centres of excellence housed within the Regenstrief Institute. Journal as well as article impact varied, with the highly cited articles presenting the results of primarily CER or RCT designs.

Patterns observed in this set of articles demonstrate that an information infrastructure like the INPC can, over time, support both a wide range of studies and external collaborators. Furthermore, integrated information infrastructures move us closer to achieving the goals of a learning health system. Traditionally, HSR as well as public health researchers have principally used administrative data sets (e.g. insurance claims) to examine process measures indicative of health care delivery and outcomes. Integrated data resources such as the INPC enable examination of both administrative and clinical data for a wide variety of patient populations. Such resources are necessary not only for CER in which one desires to examine two or more interventions,37 but they are a prerequisite for post-marketing drug surveillance, policy analysis and understanding of the social determinants of health.12,38 In the remainder of this section, we discuss the results in the context of Regenstrief and its evolution towards supporting the learning health system. We further discuss the limitations of this study, future directions and our recommendations for others seeking to understand the value of their own infrastructure or research publications.

Regarding the scope of studies, the INPC has enabled a growing community of researchers to examine the prevalence, burden and interventions designed to address multiple, vexing diseases of public health significance. Investigators in the Indiana University Center for Aging Research,
housed within the Regenstrief Institute, have utilised INPC data to measure the prevalence of cognitive disorders and evaluate interventions designed to address dementia and other cognitive impairments that affect older adults’ morbidity and mortality.39–43 Investigators in Regenstrief’s centres for HSR and biomedical informatics have leveraged INPC data to evaluate informatics as well as non-technical interventions designed to improve the diagnosis, treatment and self-management of non-communicable diseases such as diabetes and chronic heart failure,29,30,44 which account for a large proportion of the health burden in the U.S. as well as globally.45,46 Finally, public health researchers at Regenstrief have utilised INPC data to study the incidence and interventions designed to prevent or control the spread of infectious diseases such as syphilis and methicillin-resistant staphylococcus aureus.47–49 Such insight is useful in demonstrating the value of the infrastructure to health system partners, the Regenstrief Board of Directions as well as external funders of research at Regenstrief.

In addition to supporting a broad set of studies, analysis of these articles suggests that a robust information infrastructure can support broad collaboration or team science. The patterns in publications show that in each subsequent decade, an increasing number of non-Regenstrief collaborators appear as authors, including first and senior author, on published papers utilising RMRS or INPC data. For example, in Kim et al.,50 the INPC was utilised to test for associations between genetic polymorphisms of adrenergic receptors (AR) and health care utilisation in heart failure patients. In Aljadhey et al.,51 the INPC was utilised in a cohort study to compare serum potassium concentrations and electrocardiograms for patients prescribed cycloxygenase-2 (COX-2) inhibitors with those prescribed non-selective non-steroidal anti-inflammatory drugs (NSAIDs) using propensity matching. These first authors are not Regenstrief scientists, documenting productive collaborations between RI and non-RI researchers as well as staff who support the extraction and analysis of data from the INPC. Again, documentation of external collaboration provides evidence of value that is useful in discussions with funders of extramural research as well as health system partners who contribute data to the infrastructure.

Trends from RI publications provide evidence that efforts to support the creation of cross-disciplinary as well as translational mechanisms within academic medical centres through infrastructure can be successful. The CTSA programme funded by NIH has created more than 60 infrastructures where informatics is a core.17 The goal is to stimulate translational research whereby novel discoveries in bench science can be systematically evolved towards clinical application to improve health services as well as population outcomes. Regenstrief is responsible for the informatics core at the CTSA in Indiana, which serves Indiana University, Purdue University and Notre Dame University. Like Hruby et al.,24 Regenstrief has created streamlined processes to enable investigators within the CTSA to request, access and utilise data from the INPC to support translational research. Although the Institute did not explicitly prevent non-Regenstrief scientists from accessing its data in the past, accessing and manipulating data has become more efficient in the last decade. For example, previously accessing RMRS data required funding for the creation of a ‘study database’ into which data were manually extracted, linked and reviewed in combination with a chart review to ensure accuracy prior to initiation of the actual study. Furthermore, the study often required ethics and compliance review by all members of the institutional review board (IRB), which was largely unfamiliar with EHR data. Today, a collaboration to analyse RMRS data requires less ‘overhead’ as the data can be more efficiently queried from the INPC, automatically linked using probabilistic techniques, and chart review is rarely required due to advances data quality checks. Moreover, the IRB now offers a simple checkbox on expedited as well as exempt study applications for data from the INPC. Standardised Web-based submission forms on the CTSA website even allow non-Regenstrief investigators to obtain de-identified cohort information for grant applications without IRB approval. Streamlined processes at both Regenstrief and the IRB facilitate quicker and less costly methods for obtaining data prior to the start of a study facilitating both access and
collaboration. These process changes may explain, in part, the rise in more recent studies able to access and use data from the INPC infrastructure. Furthermore, expanded access and use are indicators that Regenstrief is meeting its goals as a core service of the CTSA.

Patterns of increased utilisation by larger numbers of collaborators further suggest that robust information infrastructures are likely to enable a learning health system. The articles in this analysis represent the yield from significant investment by Regenstrief and its clinical partners in creating a robust information infrastructure capable of supporting analysis of heterogeneous data collected during routine health services across a variety of clinical facilities for a wide range of purposes (e.g. quality improvement, CER, HSR). Therefore, infrastructures like the INPC and the clinical data repository in Hruby et al. can become assets that health systems utilise to improve health care delivery and population outcomes. Yet creating and sustaining such infrastructures require investment not only in technology but also in people and processes capable of streamlining data requests, data quality assurance and analysis. After all, it is people who do something with the results, whether that be to publish peer-reviewed publications or implement new techniques into routine clinical care. It is also people who troubleshoot and resolve technical issues that inevitably arise when operating an information system like the INPC.

Both the RMRS and INPC infrastructures faced significant challenges in their evolution. Many of the 83 articles that were categorised as technical papers and excluded from this review summarise the development, expansion and improvement of these infrastructures over time. For example, creating and sustaining interoperability among disparate clinical information systems is difficult. The INPC continues to struggle with ever changing source systems, system interfaces, and semantic coding that impacts the capture, storage and utilisation of routine clinical data. Our clinical decision support (CDS) rules and knowledge bases continue to need care and regular updates, and our software tools continue to evolve as users demand more graphical interfaces in place of command line prompts. Furthermore, we continue to pursue new data sources and elements to expand the INPC’s capacity to support quality improvement as well as CER for our clinical and research partners.

Future directions

Moving forward, Regenstrief will continue to evolve and expand its information infrastructure. We will continue to support the Indiana CTSA and its vast network of researchers. In recent years, we also formed new strategic partnerships with the Merck Corporation, CDC, and ConvergeHealth by Deloitte. These partnerships expand the ‘big tent’ of HSR and informatics to increase the utilisation of data from the infrastructure as well as develop novel tools and analytics techniques to enhance our capacity to perform HSR, CER and quality improvement. For example, we are actively working on projects to improve the quality of data emanating from EHR systems, further unlock the potential of free text in clinical records using natural language processing, and to develop new methods for analysing observational data. These activities not only help us sustain the infrastructure but also collectively contribute to the LHS and to meeting our organisational mission, which is to improve patient and population health services as well as outcomes.

We will further continue to engage with health information professionals for input and research on our integrated health information infrastructure. Regenstrief has historically recognised the value of making health information professionals partners in research. For over a decade, the head of our medical library served as a core investigator at Regenstrief, guiding and conducting research on both the RMRS and INPC. This study is a continuation of that culture, even though the former medical library director was not involved in it. Although this may not be the trend elsewhere, we recommend partnering with health information professionals in translational research. From the initial discussion of the project, framing the study design as well as the methods, and conducting the analysis, the health information professional co-author played a vital role in this study. We hope that others can engage their local
health information professionals to design and conduct research not only on the scholarly impact of an information infrastructure like the INPC but also on a range of other informatics tools including clinical decision support for health professionals, portals for patient access to their health information, clinical data management and population health tools for use in public health departments.

Limitations

Judgment of the reviewers and authors when reviewing candidate articles may have excluded articles considered by others within Regenstrief or the biomedical informatics community to represent studies involving RMRS or INPC data. For example, in the final decade, the number of investigators at Regenstrief more than doubled (from 26 in 2003 to 62 in 2012). This coincided with an overall increase in RMRS and INPC publications but represents a lower proportion of total publications by Regenstrief investigators. One possible reason for this may be an unclear or inadequate description of the source of data in published articles. For example, an investigator may have utilised RMRS data but neglected to mention the RMRS by name in the Methods section of the paper, either because the journal did not require a formal description of the data source or because the use of EHR data has become so ubiquitous at Regenstrief the authors felt it unnecessary. This may have impacted our ability to appropriately classify an article, illustrating that authors need to do a better job of following publication guidelines and best practices, such as those outlined in the Statement on Reporting of Evaluation Studies in Health Informatics and the STROBE (strengthening the reporting of observational studies in epidemiology) statement.

Our analysis was further limited to examination of articles as independent observations as we did not link articles published from the same study. It is therefore plausible that the 190 articles represent just 80–100 unique studies that each has multiple aims. Future examinations of articles published by a research organisation like Regenstrief might consider linking studies based on the grant numbers listed in the acknowledgements section of the articles. Such insight would provide another measure of research productivity in addition to utilisation of data from an information infrastructure.

In addition, our search was limited to MEDLINE. While it is a comprehensive source of the biomedical literature, especially outcomes research, it is possible that we did not identify all articles published by Regenstrief investigators. We attempted to compensate for this limitation by asking investigators to submit curricula vitae for comparison with search results. However, it was not possible to collect curricula vitae from all investigators as some left the institute more than a decade ago and could not be contacted.

Free tools like Google Scholar, as well as proprietary tools like Scopus and Web of Science, are only as good as the information contained in them and the methods of retrieval available. While the data in these tools overlap, there are differences in the indexing methods which yield different results for the same type of query. For example, Google Scholar indexes not only peer-reviewed literature but also book chapters, dissertations and conference proceedings. When searching Google Scholar, we excluded presentations, book chapters and dissertations, and we manually checked that the citations we included were for peer-reviewed articles. In addition, Google Scholar blocked the number of citation searches we could run in a day, making it cumbersome and time-consuming to complete this part of the analysis. This may have caused us to miss some valid citations.

Furthermore, our methods targeted only a portion of the larger LHS concept, which involves feeding knowledge back into the health system to stimulate change in behaviour as well as outcomes. We did not examine whether the articles utilising RMRS or INPC data actually influenced local care delivery, health system behaviour or health policy. These aspects would need to be examined to comprehensively evaluate the impact of an information infrastructure on the LHS. Finally, the article type categories largely reflect the interests of the cadre of RI investigators, which while diverse, still reflects a limited scope of possible uses of digital data for the LHS.
Recommendations

Those wishing to conduct similar analyses should take note of these limitations. Furthermore, we recommend exploring the use of the ‘Organizations – Enhanced List’ feature available in Web of Science. Unfortunately, this functionality did not exist when this study was conducted. The new feature allows for search of organisation preferred names and their variants, which might have enabled us to better identify Regenstrief investigators. We also recommend the development and use of methods to enable analysis of average journal impact factors, citation counts and h-indices over time in relation to the number of investigators and volume of data available in an infrastructure like the INPC. In this study, we only measured the discrete indicators over time and summarised them for a given decade. We did not use them as input to any statistical models, nor did we control for changes in the number of investigators or volume of data in the INPC.

Conclusion

An integrated, robust information infrastructure is needed to support the broad vision of an LHS. Efforts like those at the Regenstrief Institute are providing platforms that enable data captured for routine clinical purposes to be utilised for observational studies, HSR and CER. As similar efforts expand across the U.S. and globally, the informatics community should continue to evaluate successes and failures, strengthening the evidence base with respect to why and how such infrastructures contribute to local, regional and national learning health systems. Such evidence will likely be necessary to demonstrate to stakeholders that robust information infrastructures contribute to the LHS yet require continued investment to sustain activities that are complex and challenging given the nature of electronic clinical data. Importantly, our results support a strong spirit of collaboration between Regenstrief investigators and external colleagues.

Acknowledgements

The authors thank Erika Titus and Wesley Hoffmann, pharmacy interns from Purdue University, for their help in reviewing candidate articles for inclusion in the study. Dr. Dixon is a Health Research Scientist at the Richard L. Roudebush Veterans Affairs Medical Center in Indianapolis, Indiana. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Regenstrief Institute, Robert Wood Johnson Foundation, Agency for Health Care Research and Quality, Centers for Disease Control and Prevention, Department of Veterans Affairs, or the U.S. government.

Funding

BED is supported by a Mentored Research Scientist Development Award from the Robert Wood Johnson Foundation (71596) as well as awards from the U.S. Agency for Health Care Research and Quality (R01HS020209), U.S. Centers for Disease Control and Prevention (200-2011-42027 0003), the Merck-Regenstrief Program in Personalized Health Care Research and Innovation, and the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416.

Competing Interests

BED and MDM are Research Scientists with the Regenstrief Institute, Inc. Although they are not paid a salary by Regenstrief, they direct research teams at Regenstrief and receive non-monetary benefits (e.g. office space, administrative support). ECW and JML have no real or apparent conflicts of interest.

References

3 The Learning Health care System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine). Olsen,
A review of EHR-based outcomes research, Brian E. Dixon et al.


9 Etheredge, L. M. A rapid-learning health system. *Health Affairs (Project Hope)* 2007, 26, w107–w118.


Overhage, J. M., Grannis, S. & McDonald, C. J. A comparison of the completeness and timeliness of automated
A review of EHR-based outcomes research, Brian E. Dixon et al.


Dixon, B. E. & Grannis, S. Why Asking Questions about Data and Their Sources is Important to Public Health Informatics Practice and Research. Online Journal of Public Health Informatics 2011, 3, ojphi.v3i3.3792.


Received 9 December 2014; Accepted 16 October 2015

Appendix A

Metadata Extracted from MEDLINE and Article

- Title
- Authors
- Abstract
- Source
- Volume, Issue, Page Numbers
- Year
- Article Type
- MeSH Headings
- Institution(s)
- Acknowledgements
- Cohort Size

Category: Data Sources
- RMRS
- INPC
- OTHER
- UNKNOWN

Category: Disease
- Acne
- Achondroplasia
- OTHER

© 2015 Health Libraries Group
Health Information & Libraries Journal
• Affective Disorders
• AIDS
• Albinism
• Alcoholic hepatitis
• Allergy
• Alopecia
• Alzheimer’s disease
• Amblyopia
• Amoebiasis
• Anaemia
• Aneurysm
• Anorexia
• Anosmia
• Anotia
• Anthrax
• Appendicitis
• Apraxia
• Argyria
• Arthritis
• Aseptic meningitis
• Asthma
• Astigmatism
• Atherosclerosis
• Athetosis
• Atrophy
• Auditory agnosia
• Bacterial meningitis
• Behavioural disorder
• Bella killer
• Beriberi
• Black Death
• Botulism
• Breast cancer
• Bronchitis
• Brucellosis
• Bubonic plague
• Bunion
• Calculi
• Campylobacter infection
• Cancer
• Candidiasis
• Carbon monoxide poi
• Celiacs disease
• Cerebral palsy
• Chagas disease
• Chalazion
• Chancroid
• Chavia
• Cherubism
• Chickenpox
• Chlamydia
• Chlamydia trachomatis
• Cholera
• Chordoma
• Chorea
• Chronic fatigue syndrome
• Chronic Heart Disease
• Chronic Lung Disease
• Circadian rhythm sl
• Coccidioidomycosis
• Cognitive impairment
• Colitis
• Common cold
• Condyloma
• Congestive heart disease
• Coronary heart disease
• Cowpox
• Cretinism
• Crohn’s Disease
• Dehydration
• Delirium
• Dementia
• Dengue
• Depression
• Diabetes mellitus
• Diphtheria
• Ear infection
• Ebola
• Emphysema
• Encephalitis
• Epilepsy
• Erectile dysfunction
• Foodborne illness
• Fractures
• Gangrene
• Gastroenteritis
• Genital herpes
• GERD
• Goitre
• Gonorrhoea
• Heart disease
• Hepatitis A
• Hepatitis B
• Hepatitis C
• Hepatitis D
• Hepatitis E
• Histiocytosis (Child)
• histoplasmosis
• HIV
• Human papillomavirus
• Huntington’s disease
• Hypercalcaemia
• hyperkalaemia
• Hypermetropia
• Hyperopia
• Hypertension
• Hypertensive Nephropathy
• Hyperthyroidism
• Hypokalemic Nephropathy
• hyponatraemia
• Hypothermia
• Hypothyroid
• Hypotonia
• Ignious Syndrome
• Impetigo
• Infertility
• Influenza
• Interstitial cystitis
• Interstitial Nephri
• Intervertebral Disc
• Iritis
• Iron-deficiency anaemia
• Irritable bowel syndrome
• Jaundice
• Keloids
• Kuru
• Kwashiorkor
• Laryngitis
• Lead poisoning
• Legionellosis
• Leishmaniasis
• Leprosy
• Leptospirosis
• Leukaemia
• Lice
• Listeriosis
• Loiasis
• Lung cancer
• Lupus erythematosus
• Lyme disease
• Lymphogranuloma ven
• Lymphoma
• Major Depression
• Malaria
• Marburg fever
• Mattticular syndrome
• Measles
• Melanoma
• Melioidosis
• Ménière’s disease
• Meningitis
• Metastatic cancer
• Migraine
• Mononucleosis
• Morquo Syndrome
• MRSA
• Multiple myeloma
• Multiple sclerosis
• Mumps
• Muscular dystrophy
• Musculoskeletal Pain
• Myasthenia gravis
• mycobacterium
• Myelitis
• Myocardial Infarc.
• Myoclonus
• Myopia
• Myxedema
• Necrotizing Fasciitis
• Neoplasm
• Night blindness
• Non-gonococcal urethritis
• Obesity
• Obstructive Uropathy
• Osteoarthritis
• Osteoporosis
• Otitis
• Palindromic rheumatism
• Paratyphoid fever
• Parkinson’s disease
• Pelvic inflammatory
• Periodontal disease
• Peritonitis
• Pertussis
• Phenylketonuria
• Pilia
• Plague
• Pneumocystitis
• Poliomyelitis
• Porphyria
• Prerenal Azotemia
• Progeria
• Prostatitis
• Psittacosis
• Psoriasis
Pubic lice
Pulmonary embolism
Q fever
Ques fever
Rabies
Renal disease
Renal Failure
Renal Insufficiency
Renal Vascular Disease
Repetitive strain i
Rheumatic fever
Rheumatic heart
Rheumatism
Rheumatoid arthritis
Rickets
Rift Valley fever
Rocky Mountain spot
Rubella
Salmonellosis
SARS
Scabies
Scarlet fever
Schizophrenia
Sciatica
Scleroderma
Scrapie
Scurvy
Sepsis
Septicaemia
Shigellosis
Shin splints
Shingles
Sickle-cell anaemia
Siderosis
SIDS
Silicosis
Smallpox
Spinal stenosis
Stevens-Johnson syndrome
Stomach flu
Stomach ulcers
Strabismus
Strep throat
Streptococcal infection
Swine influenza
Synovitis
Syphilis
Taeniasis
Tay-Sachs disease

Tennis elbow
Teratoma
Tetanus
Thalassaemia
Thrush
Thymoma
Tinnitus
Tonsillitis
Tooth decay
Toxic shock syndrome
Trichinosis
Trichomoniasis
Trisomy
Tuberculosis
Tularaemia
Tumour
Tungiasis
Typhoid fever
Typhus
Ulcerative colitis
Ulcers
Uraemia
Urticaria
Uveitis
Varicella
Varicose veins
Vasovagal syncope
Viral fever
Viral meningitis
Vitiligo
Von Hippel-Lindau d
VRE
Warkany syndrome
Warts
Watkins
Yellow fever
Yersiniosis

Category: Funding Source
- Agency for Health Care Policy and Research
- Agency for Health care Research and Quality
- “American Federation for Aging Research, Medical Student Summer Research Training in Aging Scholarship”
- American Federation of Aging Research
- American Geriatrics Society
- Astra Zeneca
- Atlantic Philanthropies Inc.
- Bristol-Myers Squibb
Appendix B

Complete List of Articles Reviewed


19 Boustani, M. A., Campbell, N. L., Khan, B. A., Abernathy, G., Zawahir, M., Campbell, T., Tricker, J., Hui, S. L., Buckley, J. D., Perkins, A. J., Farber, M. O. & Callahan, C. M. Enhancing care for hospitalized older adults...


© 2015 Health Libraries Group

*Health Information & Libraries Journal*


60 Dexter, P. R., Blevins, L., McDonald, C. & Overhage, M. Participation in a prototype federated system for sharing medical content. *AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2007*, 938.


70 Finnell, J. T., Overhage, J. M., Dexter, P. R., Perkins, S. M., Lane, K. A. & McDonald, C. J. Community clinical data exchange for emergency medicine patients. *AMIA*
A review of EHR-based outcomes research, Brian E. Dixon et al.


Marrero, D. G., Vandagriff, J. L., Gibson, R., Fineberg, S. E., Fineberg, N. S., Hiar, C. E. & Crowley, L. E.


142 Perkins, A. J. & Clark, D. O. Assessing the association of walking with health services use and costs among...

143 Perkins, A. J., Kroenke, K., Unutzer, J., Katon, W., Williams, J. W. Jr, Hope, C. & Callahan, C. M. Common comorbidity scales were similar in their ability to predict health care costs and mortality. Journal of Clinical Epidemiology 2004, 57, 1040–1048.


A review of EHR-based outcomes research, Brian E. Dixon et al.


Zhu, V. J., Belsito, A., Tu, W. & Overhage, J. M. Data for drugs available through low-cost prescription drug programs are available through pharmacy benefit manager and claims data. BMC Clinical Pharmacology 2012, 12, 12.


Received 9 December 2014; Accepted 16 October 2015

© 2015 Health Libraries Group

Health Information & Libraries Journal