Dig Dis Sci (2016) 61:3236–3245
DOI 10.1007/s10620-016-4278-z

ORIGINAL ARTICLE

Development of an Inflammatory Bowel Disease Research Registry Derived from Observational Electronic Health Record Data for Comprehensive Clinical Phenotyping

Alyce J. M. Anderson[1] · Benjamin Click[2] · Claudia Ramos-Rivers[2] · Dmitriy Babichenko[3] · Ioannis E. Koutroubakis[2] · Douglas J. Hartman[4] · Jana G. Hashash[2] · Marc Schwartz[2] · Jason Swoger[2] · Arthur M. Barrie III[2] · Michael A. Dunn[2] · Miguel Regueiro[2] · David G. Binion[2]

Received: 3 June 2016 / Accepted: 10 August 2016 / Published online: 12 September 2016

Abstract

Background Inflammatory bowel disease (IBD) is a heterogeneous collection of chronic inflammatory disorders of the digestive tract. Clinical, genetic, and pathological heterogeneity makes it increasingly difficult to translate efficacy studies into real-world practice. Our objective was to develop a comprehensive natural history registry derived from multi-year observational data to facilitate effectiveness and clinical phenotypic research in IBD.

Methods A longitudinal, consented registry with prospectively collected data was developed at UPMC. All adult IBD patients receiving care at the tertiary care center of UPMC are eligible for enrollment. Detailed data in the electronic health record are accessible for registry research purposes. Data are exported directly from the electronic health record and temporally organized for research.

Results To date, there are over 2565 patients participating in the IBD research registry. All patients have demographic data, clinical disease characteristics, and disease course data including healthcare utilization, laboratory values, health-related questionnaires quantifying disease activity and quality of life, and analytical information on treatment, temporally organized for 6 years (2009–2015). The data have resulted in a detailed definition of clinical phenotypes suitable for association studies with parameters of disease outcomes and treatment response. We have established the infrastructure required to examine the effectiveness of treatment and disease course in the real-world setting of IBD.

Conclusions The IBD research registry offers a unique opportunity to investigate clinical research questions regarding the natural course of the disease, phenotype association studies, effectiveness of treatment, and quality of care research.

Keywords Inflammatory bowel disease · Phenotyping · Natural history · Registry

Introduction

Inflammatory bowel disease (IBD) consists of two main entities: Crohn's disease (CD) and ulcerative colitis (UC). IBD is estimated to affect up to two million Americans, with an increasing annual incidence of 39.4 cases per 100,000 person-years in North America [1]. CD and UC result in morbidity, disability, and heightened mortality, generating approximately $6.3 billion in direct healthcare costs and an additional $3.6 billion in indirect costs due to loss of productivity [2–4]. The clinical course of IBD is variable and often unpredictable. IBD severity ranges from mild symptoms to severely debilitating disease. Therefore, IBD encompasses a large spectrum of severity, disease duration, disease course, and complexity of disease-related

Miguel Regueiro and David G. Binion are co-senior authors.

The work was performed at the University of Pittsburgh and UPMC.

✉ David G. Binion
binion@pitt.edu

[1] School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

[2] Division of Gastroenterology, Hepatology, and Nutrition, Department of Medicine, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15213, USA

[3] School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA

[4] Department of Anatomic Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

extra-intestinal manifestations. This heterogeneity of disease over time and across individuals significantly limits our ability to translate results from randomized controlled clinical trials into clinical practice [5]. For example, the effectiveness of therapeutic drugs varies over time within individuals, and across individuals with different degrees of IBD severity [6–8].

Given our incomplete understanding of disease heterogeneity and the inherent limitations of clinical efficacy studies in IBD, we sought to define clinical subtypes of disease by examining disease course patterns and the effectiveness of medical therapy in a tertiary care clinic. To this end, we developed a research registry of IBD patients at UPMC. The aims of the IBD research registry are to: (1) organize clinical information and define the natural history of IBD; (2) develop a research platform for association studies and the delineation of clinical phenotypes; and (3) examine effectiveness and quality of care measures in the setting of IBD. We propose that a database using prospective observational health record data will support and facilitate natural history, disease phenotyping, and effectiveness research in chronic illness. This manuscript details the design, development, challenges faced, and implementation of the IBD research registry at a large tertiary care center in the USA.

Materials and Methods

Registry Setting

The IBD research registry is an IRB-approved patient registry maintained at UPMC in Pittsburgh, PA. The registry was created in 2001 by a gastroenterologist who is also the principal investigator (MR). The principal investigator originally began enrolling IBD patients in the National Institutes of Digestive, Diabetes and Kidney Disease genetics consortium and used the registry in parallel to prospectively consent all IBD patients visiting the UPMC Digestive Disorder Center. The rationale for the initial development of the registry was to gather clinical data for IBD patients who were participating in the genetic discovery arm of the consortium. The registry was initially managed as a research tool, outside of the daily clinical practice and clinician access. In 2008, an initiative was formed to incorporate observational healthcare data into the registry. To achieve this goal, the registry was moved into an electronic, Health Insurance Portability and Accountability Act (HIPAA)-compliant secure environment that could be readily accessed by gastroenterologists, support staff, and IRB-approved research collaborators. Concurrently, UPMC introduced an outpatient electronic health record system (EpicCare, Epic Systems, Verona, WI, USA) for all affiliated sites in the UPMC system, which includes over 20 hospitals and 500 clinics across Western Pennsylvania. This system-wide electronic medical record allowed all clinical data across numerous healthcare facilities to be captured.

Registry Enrollment

The target population for the registry is adult IBD patients, 18 years of age and older. This population is derived from patients presenting to the Digestive Disorders Center at UPMC, a tertiary care clinic with physicians with IBD expertise. Recruitment for the IBD research registry is ongoing and in perpetuity. As a part of the initial clinic visit, all patients are provided a consent form with a description of the IBD registry. All IBD physicians and their staff are co-investigators on the registry. At the time of the clinic visit, the registry is explained to the patient by a co-investigator and the patient has the opportunity to ask questions. The patient may choose to sign the registry consent during the clinic visit, may decline enrollment in the registry, or may elect to take the consent with them to review further and ask additional questions. The consent form describes the participation risks including the most significant potential risk of a breach of confidentiality. Within the consent form, we request access to medical records for research purposes and the ability to approach enrolled patients for future research studies, both of which are critical for ongoing research and recruitment. The consent form does not have a menu of options to avoid unnecessary complexity. Any new research initiatives that would like to link data to the UPMC IBD Registry, or request biological samples, require a separate consent. Patients may withdraw from the registry at any point, and these procedures are outlined in the consent. Patients are also offered the opportunity to participate in the registry during follow-up care. Participation in the registry is optional, and all patients receive the same clinical care whether they are included in the registry or not.

Registry Design and Measures

Variables imported into the IBD research registry are generated as a part of routine outpatient clinical care (Fig. 1). Data extraction from the electronic medical record (EMR) for the purpose of the IBD registry has been occurring since 2009. Clinical data for patients enrolled in the registry are systematically exported from the EMR through the Center for Assistance in Research using eRecord at the University of Pittsburgh, an information technology support group. Data are retrieved from the EMR biannually, categorized according to clinical domain (medications, radiology, laboratories, etc.), and delivered

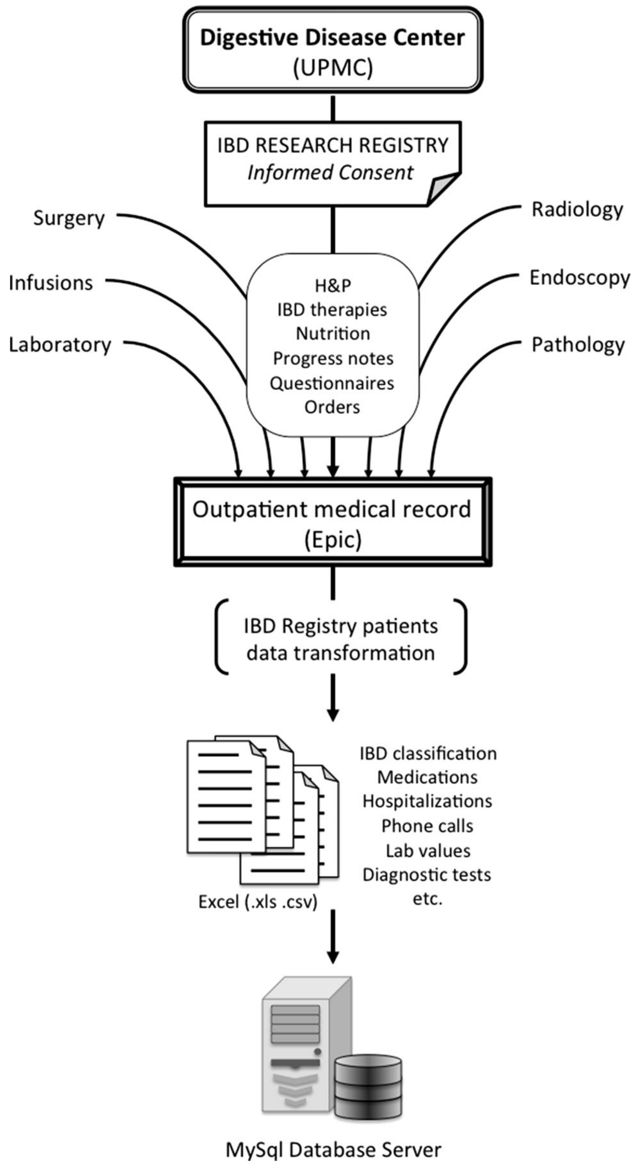

Fig. 1 Flow diagram of inflammatory bowel disease research registry. *IBD* inflammatory bowel disease, *H&P* history and physical

electronically to the registry research team. Automated and manual data transformations are used to separate all values into domain-specific datasets. Clinical events are categorized as binary (categorical) outcomes (0,1) for initial statistical analyses. Laboratory values are imported as raw numbers and are further defined as normal or abnormal (0,1). Annual dichotomous patterns for clinical events, medication prescriptions, and abnormal laboratory values are created. The exact dates of these data points are also preserved for time-to-event analyses. All data are stored behind the HIPAA-compliant, password-protected UPMC firewall, in a secure environment, only accessible by co-investigators listed on the IRB approval who have completed appropriate human subjects research training. Master data lists are password-protected and archived for data integrity.

In order to provide access to the data for research collaborators, and to enable advanced data analysis, the data are completely de-identified and imported into a relational database or statistical package. The de-identification process creates unique identifiers for each record and generates a patient lookup list used to link the original data structure. This lookup list is stored behind the UPMC firewall, and patient identifiers are separated from the data. The relational database is deployed on a secure HIPAA-compliant server accessible via a virtual private network by co-investigators listed on the IRB approval. The de-identification process also allows for collaboration with outside institutions if multi-site collaborations arise.

Table 1 Measurements collected in the inflammatory bowel disease research registry

Patient demographics
Disease-related information
Age at diagnosis
Disease duration
Disease localization according to Montreal classification
Disease phenotype according to Montreal classification
Patient questionnaires
SIBDQ
HBI
UCAI
Endoscopic data
Pathology (surgical and procedural)
Radiological data
Comorbidities—ICD-9 codes
Laboratory
Standardized IBD laboratory panel
Fecal microbial testing including *Clostridium difficile*
Medications
IBD-related prescriptions
Other prescriptions
Healthcare utilization
IBD clinic visits
Emergency department visits
Hospital admissions
Telephone encounters
Radiology
Surgical procedures
Total charges for healthcare services

IBD inflammatory bowel disease, *SIBDQ* Short Inflammatory Bowel Disease Questionnaire [37, 38], *HBI* Harvey–Bradshaw Index [10], *UCAI* ulcerative colitis activity index [47], *ICD-9* International Classification of Diseases, Ninth Revision

Variables collected include demographic information (Table 1) and initial IBD classification of CD or UC. Related comorbidities are recorded through the use of administrative International Classification of Diseases, Ninth Revision and Tenth Revision (ICD-9, ICD-10) codes and EMR problem lists. Laboratories, vitals and objective data collected and entered into the EMR during the patient visit are organized by subject and visit date. Healthcare utilization measures include telephone encounters, emergency department visits, IBD clinic visits, hospital admissions, endoscopic and radiological procedures, and surgeries. Financial healthcare charges are also organized by year. Medication data include prescriptions for biologics, immunomodulators, steroids, 5-aminosalicylates and iron supplementation for each calendar year. Data on psychiatric and opiate pain medications are also collected as markers comorbid. Additional medications that are not part of routine IBD care can be retrieved using search algorithms.

As a part of the standardized visit in the UPMC Digestive Disorders Center, each patient is asked to complete health-related questionnaires, such as the published version of the Short Inflammatory Bowel Disease Questionnaire (SIBDQ), the Harvey–Bradshaw Index for CD (HBI), and an ulcerative colitis activity index (UCAI) for UC [9–11]. The questionnaires are administered at every visit to inform clinical care, regardless of a patient's registry inclusion status. Patients complete hard copies of the questionnaires in their clinic room. Individual component subscores and total scores for each visit are recorded in the EMR by clinic support staff and are available for export into the registry. These standardized clinical measures allow prospective measurements of patient-reported disease clinical activity and health-related quality of life.

The outpatient EMR has many standard data entry fields, which can be exported in collaboration with the Center for Assistance in Research using eRecord at the University of Pittsburgh. When research projects would benefit from the addition of a variable that is collected in clinical care, we can make additional data extraction requests for retrieval of these data. One example would be family history of IBD and other coexistent diseases. These data are routinely collected at patient visits and was not originally a part of the initial pre-specified registry dataset. Requesting new information allows the research team the ability to create and manipulate new variables and populations of interest as new research questions arise.

In addition to data collected in standard data entry fields, the outpatient record contains health information that is not standardized in the patient chart. This valuable information appears in patient discharge summaries, surgical notes, endoscopic reports, pathology reports, clinic notes, and other free-text entry fields. Text data from free entry fields are exported into the IBD research registry in de-identified ASCII text files, which allows for the use of natural language processing toolkit and the R Project for Statistical Computing for analysis and retrieval of textual information from free-text files [12, 13]. We recently began work on implementing a combination of Apache OpenNLP natural language toolkit and Apache cTakes natural language processing (NLP) system for extraction of information from clinical free text to improve the processing of natural language text [14, 15]. There are several major challenges to extracting meaningful information from free-text data, such as pathology or endoscopy reports. These challenges include lack of structure in the narrative, multiple spellings and synonyms for terms of interest, as well as issues with context negation. For example, when extracting presence or absence of *Clostridium difficile* from pathology reports, 16 different spellings for "*Clostridium difficile*" were found. In order to address these issues, we began to use NOBLE tools developed at the University of Pittsburgh, Department of Biomedical Informatics, to create dictionaries and ontologies for terms of interest [16]. Furthermore, NOBLE tools have built-in NLP support for identifying context negation, helping resolve issues with false-positive term identification. While this approach is still in its development phases, we have been validating its accuracy with manual search and classification of terms of interest.

Quality Assurance

The IBD clinic has a standardized intake visit and also standard laboratory panels which help facilitate more uniform data capture. Many of these uniform datasets and data capture strategies are utilized at the goodwill of clinicians. Our clinicians are not incentivized to use certain notes or pre-populated laboratory order sets. Data quality and integrity are monitored by manual review after data export. Data validation occurs at the extreme values of each measure. For example, a patient weight is verified if the value does not fall within a pre-specified data validation range. Random manual EMR verification ("spot checking") is also performed on each dataset to ensure accuracy. Data linkage and matching are performed using unique patient identifiers. Missing data are imputed using the medical record, and manual data extraction from the EMR is performed on a case-by-case basis.

Registry Team

The registry is coordinated by an analytical research scientist who processes data extraction requests and organizes master files. We have found that having a dedicated staff member has revitalized the UPMC IBD Registry. The staff analytical research scientist allocates duplicate data appropriately to active researchers, preserving data integrity of the master files. Trained clinical personnel consent individuals to join the registry. Hard copies of disease activity indices and quality of life metrics are completed by patients in their clinic rooms and entered into the official medical record by clinical support staff. All other research-related operations, including the data entry of research questionnaires, are assigned to the primary investigator on each IRB-approved substudy. We also receive information technology assistance from the Center for Assistance in Research using eRecord at the University of Pittsburgh, and data science collaborators from University of Pittsburgh, School of Information Sciences.

Ethical Considerations

The research registry is an IRB-approved protocol (Protocol #0309054), open for continuous enrollment, and undergoes renewal as dictated by IRB regulations. All subsequent data linkage protocols and research questions involving the database require separate IRB approval to ensure the protection of human subjects. Subjects are able to withdraw from the IRB research registry at any time with a written request to the principal investigator.

Results

The initial registry cohort, in collaboration with other NIDDK genetics consortium institutions, has been an instrumental part of published genome-wide association studies and other genetic discoveries in IBD [17, 18]. With ongoing enrollment, the registry continued to grow after the revitalization initiative in 2008 and currently includes over 2565 patients participating in the IBD research registry (Table 2). Using annual visit trends, we estimate that approximately 70 % of the IBD patients in our clinic are actively participating in the registry. Each newly consented registry participant provides us with new data going forward, but also all retrospective data contained in their outpatient EMR from 2009 to the time of the data pull. This allows for backfilling of the data for each new registry member while avoiding manual chart review. IBD registry participants represent over 700 unique zip codes and represent a wide geographic area (Table 2). The median age is 43.8 years, and the vast majority of participants are Caucasian, while just under half the participants report full-time employment (Table 2).

Disease-related information is a critical component of the registry with nearly 90 % of IBD registry participants having defined disease phenotype based on Montreal classification (Table 3) [19]. The average disease duration is 17.4 years, and 22 % of patients have had a history of IBD-related surgery prior to 2009 (Table 3). Nearly half of the participants have CD (Table 3).

Table 2 Inflammatory bowel research registry demographics

IBD registry participants (*n*)	2565
Age[a], years (median, IQR)	43.8 (32.9–57.6)
Race, *n* (%)	
Black	60 (2.3)
White	2390 (93.2)
Other	8 (0.3)
Not specified	107 (4.2)
Living status (*n*, % alive)	2500 (97.5)
Number of zip codes	748
Employment status, *n* (%)	
Full time	1184 (46.2)
Part time	59 (2.3)
Self-employed	60 (2.3)
Student	190 (7.4)
Retired	206 (8.0)
Not employed	460 (17.9)
Not specified	403 (15.7)

IQR interquartile range

[a] Age calculated as of October 1, 2015

Table 3 Inflammatory bowel disease research registry disease information

	Total (%)
Disease classification, *n* (%)	
Crohn's disease	1313 (51.2)
Ulcerative colitis	910 (35.5)
Indeterminate colitis	7 (0.3)
IBD—unclassified	190 (7.41)
Disease duration, median (IQR)	15.0 (10–22)
Patients with Montreal classification, *n* (%)	2238 (87.3)
Patients with history of IBD surgery, *n* (%)	563 (22.0)
IBD questionnaires (*n*; median, IQR)	
SIBDQ	9905 52 (40–61)
HBI	10,446 4.84 (0–55)
UCAI	10,446 2.0 (0–6.0)

IQR interquartile range, *IBD* inflammatory bowel disease, *SIBDQ* Short Inflammatory Bowel Disease Questionnaire [9], *HBI* Harvey–Bradshaw index [10], *UCAI* ulcerative colitis activity index [11]

The effort to achieve our primary aim to organize prospectively collected, longitudinal clinical information has resulted in over 500 gigabytes of temporally organized data. We have organized over 1.3 million laboratory values and 124,658 prescriptions since 2009 (Table 4). Routinely collected utilization measures, including office visits, telephone calls, surgeries, hospitalizations, emergency room visits, and radiological or endoscopic procedures, have been organized by year (Table 4). We have organized over $310 million of total financial healthcare charges incurred by patients in the IBD registry and are exploring financial charge data as a new phenotype of disease severity (Table 4).

Our second aim was to develop a research platform for the definition of clinical phenotypes. The data have resulted in multiple clinical phenotypes that have been published and are associated with increased levels of healthcare utilization or predictive of poor disease outcomes (Table 5). These phenotypes include patients with high-volume telephone calls, persistent or recurrent anemia, CRP elevations, and peripheral eosinophilia [20–24]. We developed a set of tools written in Python programming language to search unstructured text data and identify patients with features of interests. These features included presence of granulomas on pathology reports, as well as presence or absence of *Clostridium difficile* in endoscopy reports, both of which have resulted in meaningful subgroups for analysis [25, 26]. Projects are underway to link the clinical phenotypes to genotype signatures and utilize genetic data to understand the relationship of drug metabolism polymorphisms and patient data in our population.

Table 4 Inflammatory bowel disease research registry total number of measurements

	Total number organized
Laboratory values	1,308,993
Clinic visits	36,747
Telephone encounters	645,888
Emergency room visits	7378
Hospital admissions	3508
Endoscopies	8472
Surgeries	1304
Radiology (*n*)	
CT	7716
MRI	2585
X-ray	10,569
Comorbidities[a]	2152
Prescriptions	
Biologics[b]	5976
Immunomodulators	6897
Systemic steroids	6867
5-ASA	4652
Other	100,236
Total charges organized ($)	$310.3 Million

5-ASA 5-aminosalicylic acid medications

[a] Based on the International Classification of Diseases, Ninth Revision (ICD-9) codes contained in patient-specific problem lists

[b] Biologics included anti-tumor necrosis factor agents (infliximab, adalimumab, certolizumab pegol)

Finally, we have overcome numerous challenges during the development and implementation of a longitudinal natural history database (Table 6). An ongoing challenge is the quantification of patient follow-up. With natural history data, it is difficult to distinguish whether a patient did not have an endoscopy because they were lost to follow-up, or if they are feeling well and did not require endoscopic evaluation. To ensure patients in hypothesis-driven studies resulting from registry data are only included if they are active in our practice, we organize outpatient EMR encounters to quantify a patient's telephone activity, email exchanges, clinic visits, or emergency department use in a calendar year. These data are the backbone of all inclusion and exclusion criteria for multi-year studies and are not routinely accessible in registries unlinked from the EMR or cohorts based on administrative datasets.

Another facet of data organization that commonly accompanies longitudinal data is the identification of study observation intervals. To address this, we prospectively lock our data on a calendar year basis. This strategy allows for cross-sectional association studies to be repeated on each annually locked dataset and provides internal validation for the evaluation of trends over time. Previous

Table 5 Example clinical phenotypes explored using the inflammatory bowel disease research registry

Phenotypes	Risk of adverse health outcome
Silent Crohn's disease: CRP elevation without clinical symptoms [21]	Increased risk and rate of hospitalization
Persistent/recurrent anemia [22]	Associated with increased healthcare utilization
High telephone encounters [20]	Increased risk of hospitalization and/or emergency room use
Peripheral eosinophilia [23, 24]	Increased patient charges and healthcare utilization
Obesity in IBD [48]	Use of lower dosing of IBD-related medications
Long-term lipid profiles in IBD [49]	Dyslipidemia is associated with more severe disease
High healthcare utilization in IBD [27]	Associated with unemployment, psychiatric disease, narcotic use, and medical comorbidities

CRP C-reactive protein, *IBD* inflammatory bowel disease

Table 6 Challenges and solutions in creating and maintaining a registry

Challenges	Solutions
Data extraction from the electronic medical record	Active and ongoing partnerships with outpatient medical record support teams at our local institution. Our local partners facilitate data transfer requests
Standardization of data capture	Patient encounters are standardized, regardless of inclusion in the registry. There are standard laboratory orders and questionnaires
Quantification of patient follow-up	Participants are considered “active” if they had at least one phone call or office visit in the calendar year
Complex longitudinal data	Initially, patient data are organized by calendar year. Time-stamped data are available for more complex longitudinal data analyses and time-to-event analyses
Historical data on newly consented registry participants	Each data extraction from the electronic medical record provides historical data on each patient from 2009 to date of extraction. This overcomes the problem of manual filling of historical data as in other non-electronic medical record-derived registries
Recruitment and retention	All clinic physicians are actively recruiting IBD patients to join the registry. By utilizing the electronic medical record as a data source, we greatly reduce participant burden and increase retention. Physician and staff data entry burden is also minimized

studies from our group have employed the data to generate prediction models with data from one calendar year and perform a validation of the prediction model in subsequent years [27]. The annual trend data are the primary way in which data are curated; however, all raw data from the EMR are maintained in a time-stamped manner that allows for a granular approach if any particular study requires individual data elements.

Discussion

This paper outlines the design, development, challenges, and implementation of an IBD research registry at a tertiary care center. We describe successful implementation of an IBD research registry generated from the outpatient medical record and linked surveys related to patient-reported disease activity and quality of life. Given that the majority of IBD patients are managed in the outpatient setting, the outpatient registry allows for the examination of real-world IBD subgroups, treatment patterns, disease trajectories, and clinical effectiveness in a large IBD cohort.

We have made it a research priority to define clinical subtypes of disease that relate to poor health outcomes and have demonstrated that routinely collected patient care data over time can be organized to provide the framework for such studies. The use of routinely collected observational patient data from the EMR allows for rapid implementation of research findings at other institutions [28]. Many studies use point measurements of disease activity indices, quality of life scores, or biomarkers, but with the registry’s data, we are able to evaluate patterns of these markers and trends over time which probably reflects disease severity with better accuracy. Additionally, we are capturing healthcare data from real-world patients and clinical practice, which has been advocated by the Institute of Medicine to facilitate rapid comparative effectiveness research [29]. These large datasets include patients that would be excluded from participation in randomized, controlled clinical trials due to comorbid illness or complex disease history [5]. Thus, research findings generated from the IBD research registry are a closer reflection on real-world IBD compared to highly controlled trials.

Registries have been used in the setting of other chronic disorders and rare diseases [30–34]. Despite the utility of research registries in the setting of chronic disorders, there is a lack of publications outlining registry development and implementation of longitudinal medical records data, especially in the setting of IBD. Others have described methods of patient identification using the medical record; however, this approach generates administrative data without the ability to contact individuals for current and future study recruitment [35, 36]. The EMR-derived registry approach allows identification of patients with unique clinical signatures that may benefit from enrollment in research trials. For example, we are in the process of recruiting patients to a microbiome research trial based on the extremes of documented gastrointestinal infection, which is phenotype data generated from the research registry. Additionally, while we have not yet actively pursued these studies, registries can facilitate linkage with other state and national databases to enrich the data. Furthermore, using a registry approach, we are able to validate variables that appear inaccurate and fill in any missing data using the EMR. Generating the majority of the data from the EMR also avoids data entry burnout that can restrict the potential of registries distinct from the EMR. EMR data pulls also facilitate effortless, unbiased back filling of clinical care data that is not routinely available with prospective registries distinct from the EMR.

We have linked the data in the IBD research registry to a variety of validated healthcare questionnaires in order to quantify comorbidities in our clinic population. Over time, we have collected data using an autonomic dysfunction screening questionnaire (COMPASS-31), persistent stress questionnaire, an intake depression screen, and a fiber and fat dietary intake questionnaire [37–39]. Introducing clinical questionnaires into the routine clinical workflow and linking the results of these questionnaires to patient data in the registry have resulted in studies aimed to validate these questionnaires that have not been used previously in the setting of IBD.

Finally, we now have the infrastructure required to examine effectiveness and quality of care measures in the setting of IBD, with the development of this registry. New research is focusing on the evaluation of the effectiveness of biological therapies within our patient cohort. We are also dedicating research efforts on quality of care metrics including the management of surveillance colonoscopies in patients with colonic IBD, infection rates, medication exposure, and the frequency and outcomes related to micronutrient repletion [40, 41]. Furthermore, the infrastructure currently afforded by this registry allows application of machine learning algorithms to discover patterns in the data. We in the process of testing statistical models that could be used to predict poor health outcomes are developing exploratory data visualization systems to allow clinicians and researchers to observe patient trends over time and rapidly identify clinical events of interest.

In comparison with other population-based cohorts, the EMR-based registry approach has some advantages. We have learned a great deal from Olmsted County to advance our understanding of prevalence and incidence of IBD over time [42]. However, the linked census and healthcare data are based primarily on diagnosis codes, and achieving data granularity requires retrospective chart examination. It is also often cited that a large percentage of Olmsted County inhabitants are also working in health care and are highly educated which may make natural history findings less generalizable to the larger US population [42]. The Ocean State Crohn's and Colitis Registry (OSCCAR) is another registry of a multicenter population that recruited incident cases of IBD in Rhode Island up to 6 months from initial diagnosis [43]. OSCCAR follows patients prospectively at predetermined intervals and has collected extremely valuable data on health outcomes, quality of life, and disease activity while having the added benefit of biological sample collection. While time intervals are consistent across patients, these intervals require dedicated study personnel to prospectively monitor patients and schedule follow-up. Patients in the UPMC IBD registry do not need to engage in research outside of their routine clinical care, which greatly reduces the burden of research on both the participant and research staff. Although the UPMC IBD registry does not capture data at predetermined time points, our aim is to capture real-world healthcare utilization data on a patient level, as they require care for worsening disease.

Research registries from Canada have also contributed to our understanding of IBD. The Alberta IBD Consortium recently published an influential study on IBD phenotypes and medical outcomes using their registry [44]. This study employed intensive manual chart review by two independent data abstractors with clinical expertise. In the context of our registry, the EMR data abstraction methods at the Center for Assistance in Research using eRecord at the University of Pittsburgh at UPMC are automated and uniformly applied to all registry participants and may reduce errors associated with manual data extraction and interpretation. The Manitoba IBD research group has also been influential in advancing our understanding of IBD [45]. The Manitoba group maintains an open-enrollment IBD registry and cohort studies that follow recently diagnosed IBD patients [46]. In addition to registry data, Manitoba's IBD-related epidemiologic studies are strengthened by large administrative datasets that capture universal care. The lack of universal care systems in the USA requires creative solutions to track health outcome and healthcare utilization data on the majority of our patients. We have detailed one solution to this problem through the use of a commonly employed outpatient EMR to serve as the basis of real-world data in a longitudinal IBD research registry.

Despite successful implementation of the IBD research registry, this methodology has limitations. The registry is housed at a tertiary care center and may selectively capture highly severe disease. Even with this potential bias, IBD patients in our registry are similar to IBD patients seen at other centers, in that they experience unpredictable flares with the clinical goal of controlling symptoms and restoring quality of life. Additionally, outpatient data are collected from UPMC satellite clinics and allow us to capture routine care that occurs in the community setting outside the walls of our tertiary care center. We are also limited in the breadth and accuracy of observational data as they are entered into the EMR. To address this, we validate data at the extremes to confirm any potential outliers in an effort to improve data accuracy. This limitation applies to all forms of research utilizing the EMR as a source of real-world data. Finally, with observational and interventional research studies, there is participation bias of subjects who join the registry. We are unable to capture the healthcare states and reasons why persons decline participation in the registry.

We have detailed the methods to develop, implement, and utilize a research registry in the setting of IBD and ways in which we have overcome challenges associated with real-world, longitudinal data. Future and current studies utilizing the research registry will be focused on better defining IBD phenotypes in an effort to uncover clinical pathways that can be targeted for treatment. These studies are designed to bring the practice of gastroenterology and IBD clinical management closer to the ultimate goal of personalized medicine.

Acknowledgments We would like to acknowledge the faculty and staff in the Digestive Disorders Center at UPMC who continue to make this ongoing research endeavor feasible. We would also like to thank the generosity of the participants of the IBD research registry.

Funding Alyce J.M. Anderson is supported by the University of Pittsburgh Clinical and Translational Science Institute Grant 5TL1TR000145-09 (PI: Steven E. Reis, MD), Benjamin Click, MD, is supported by the National Institute of Health training Grant 5T32DK063922-12 (PI: David Whitcomb, MD, PhD), and Ioannis Koutroubakis reports support by a sabbatical salary of Medical Faculty University of Crete, Greece. David G. Binion and Michael A. Dunn report support from Grant W81XWH-11-2-0133 from the US Army Medical Research and Materiel Command.

Compliance with ethical standards

Conflict of interest Authors do not report any conflict of interest.

References

1. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*. 2012;142: 46.e42–54.e42. **quiz e30**.
2. Kappelman MD, Rifas-Shiman SL, Porter CQ, et al. Direct health care costs of Crohn's disease and ulcerative colitis in US children and adults. *Gastroenterology*. 2008;135:1907–1913.
3. Longobardi T, Jacobs P, Bernstein CN. Work losses related to inflammatory bowel disease in the United States: results from the National Health Interview Survey. *Am J Gastroenterol*. 2003;98:1064–1072.
4. Bernstein CN, Nugent Z, Targownik LE, Singh H, Lix LM. Predictors and risks for death in a population-based study of persons with IBD in Manitoba. *Gut*. 2015;64:1403–1411.
5. Ha C, Ullman TA, Siegel CA, Kornbluth A. Patients enrolled in randomized controlled trials do not represent the inflammatory bowel disease patient population. *Clin Gastroenterol Hepatol*. 2012;1:1002–1007.
6. Gonzaga JE, Ananthakrishnan AN, Issa M, et al. Durability of infliximab in Crohn's disease: a single-center experience. *Inflamm Bowel Dis*. 2009;15:1837–1843.
7. Chaparro M, Panes J, García V, et al. Long-term durability of infliximab treatment in Crohn's disease and efficacy of dose "escalation" in patients losing response. *J Clin Gastroenterol*. 2011;1:113–118.
8. Chaparro M, Panés J, García V, et al. Long-term durability of response to adalimumab in Crohn's disease. *Inflamm Bowel Dis*. 2012;1:685–690.
9. Irvine EJ, Zhou Q, Thompson AK. The short inflammatory Bowel Disease Questionnaire: a quality of life instrument for community physicians managing inflammatory bowel disease. CCRPT Investigators. Canadian Crohn's Relapse Prevention Trial. *Am J Gastroenterol*. 1996;91:1571–1578.
10. Harvey RF, Bradshaw JM. A simple index of Crohn's-disease activity. *Lancet*. 1980;8:514.
11. Kozarek RA, Patterson DJ, Gelfand MD, Botoman VA, Ball TJ, Wilske KR. Methotrexate induces clinical and histologic remission in patients with refractory inflammatory bowel disease [Internet]. *Ann Intern Med*. 1989;1:353–356.
12. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, R Foundation for Statistical Computing, [cited 2015 Oct 12]. http://www.R-project.org; 2013.
13. Natural Language Toolkit [cited 2016 May 22]. http://www.nltk.org/.
14. The Apache Software Foundation. openNLP [cited 2016 May 22]. https://opennlp.apache.org/.
15. The Apache Software Foundation. Apache cTAKES^{TM}—clinical text analysis knowledge extraction system [cited 2016 Aug 1]. http://ctakes.apache.org/.
16. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS. NOBLE—flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinform*. 2016;14:32.
17. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;1:119–124.
18. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*. 2006;1:1461–1463.
19. Silverberg MS, Satsangi J, Ahmad T, et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can J Gastroenterol*. 2005;19:5A–36A.
20. Ramos-Rivers C, Regueiro M, Vargas EJ, et al. Association between telephone activity and features of patients with inflammatory bowel disease. *Clin Gastroenterol Hepatol*. 2014;12:986–94.
21. Click B, Vargas EJ, Anderson AM, et al. Silent Crohn's disease: asymptomatic patients with elevated C-reactive protein are at risk for subsequent hospitalization. *Inflamm Bowel Dis*. 2015. doi:10.1097/MIB.0000000000000516.
22. Koutroubakis IE, Ramos-Rivers C, Regueiro M, et al. Persistent or recurrent anemia is associated with severe and disabling inflammatory bowel disease. *Clin Gastroenterol Hepatol*. 2015;13:1760–1766.
23. Click BH, Machicado JD, Rivers CR, et al. Su1334 Eosinophilia in patients with inflammatory bowel disease is independently associated with increased healthcare expenditures: a prospective 5-year experience. *Gastroenterology*. 2015;148:S-477–S-478.
24. Machicado JD, Kabbani T, Rivers CR, et al. Sa1187 peripheral blood eosinophilia in patients with inflammatory bowel disease is associated with worse outcomes: A 5-year prospective study. *Gastroenterology*. 2015;S-148:S-251.
25. Johnson C, Hartman DJ, Rivers CR, et al. Sa1876 do epithelioid granulomas function as a biomarker of severity in Crohn's disease? Analysis of a prospective six-year natural history registry. *Gastroenterology*. 2016;150:S387–S388.
26. Anderson AJ, Rivers CR, Click BH, et al. Su1810 clostridium difficile Infection in inflammatory bowel disease predicts future healthcare utilization. *Gastroenterology*. 2016;150:S559.
27. Click B, Ramos Rivers C, Koutroubakis IE, et al. Demographic and clinical predictors of high healthcare use in patients with

inflammatory bowel disease. *Inflamm Bowel Dis*. 2016;22:1442–1449.

28. Atreja A, Achkar JP, Jain AK, Harris CM, Lashner BA. Using technology to promote gastrointestinal outcomes research: a case for electronic health records. *Am J Gastroenterol*. 2008; 103:2171–2178.
29. Sox HC. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009;4:203.
30. Ali UA, Issa Y, van Goor H, et al. Dutch Chronic Pancreatitis Registry (CARE): design and rationale of a nationwide prospective evaluation and follow-up. *Pancreatology*. 2015; 15:46–52.
31. Howes N, Lerch MM, Greenhalf W, et al. Clinical and genetic characteristics of hereditary pancreatitis in Europe. *Clin Gastroenterol Hepatol*. 2004;2:252–261.
32. Hye RJ, Inui TS, Anthony FF, et al. A multiregional registry experience using an electronic medical record to optimize data capture for longitudinal outcomes in endovascular abdominal aortic aneurysm repair. *J Vasc Surg*. 2015;61:1160–1166.
33. Schmitt-Egenolf M. PsoReg–the Swedish registry for systemic psoriasis treatment. The registry's design and objectives. *Dermatology (Basel)*. 2007;214:112–117.
34. Papp KA, Strober B, Augustin M, et al. PSOLAR: design, utility, and preliminary results of a prospective, international, disease-based registry of patients with psoriasis who are receiving, or are candidates for, conventional systemic treatments or biologic agents. *J Drugs Dermatol*. 2012;11:1210–1217.
35. Liao KP, Ananthakrishnan AN, Kumar V, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One*. 2015;24:e0136651.
36. Ananthakrishnan AN, Cagan A, Cai T, et al. Identification of nonresponse to treatment using narrative data in an electronic health record inflammatory bowel disease cohort. *Inflamm Bowel Dis*. 2016;22:151–158.
37. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav*. 1983;24:385–396.
38. Sletten DM, Suarez GA, Low PA, Mandrekar J, Singer W. COMPASS 31: a refined and abbreviated Composite Autonomic Symptom Score. *Mayo Clin Proc*. 2012;87:1196–1201.
39. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX toolkit: get the most from your measures. *Am J Epidemiol*. 2011;1:253–260.
40. Kabbani TA, Koutroubakis IE, Schoen RE, et al. Association of vitamin D level with clinical status in inflammatory bowel disease: a 5-year longitudinal study. *Am J Gastroenterol*. 2016; 111:712–719.
41. Hashash JG, Chintamaneni P, Ramos Rivers CM, et al. Patterns of antibiotic exposure and clinical disease activity in inflammatory bowel disease: a 4-year prospective study. *Inflamm Bowel Dis*. 2015. doi:10.1097/MIB.0000000000000534.
42. Loftus CG, Loftus EV, Harmsen WS, et al. Update on the incidence and prevalence of Crohn's disease and ulcerative colitis in Olmsted County, Minnesota, 1940–2000. *Inflamm Bowel Dis*. 2007;13:254–261.
43. Sands BE, LeLeiko N, Shah SA, Bright R, Grabert S. OSCCAR: ocean state Crohn's and colitis area registry. *Med Health R I*. 2009;92:88.
44. Moran GW, Dubeau MF, Kaplan GG, et al. Phenotypic features of Crohn's disease associated with failure of medical treatment. *Clin Gastroenterol Hepatol*. 2014;12:434–442.
45. Bernstein CN, Blanchard JF, Rawsthorne P, Wajda A. Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. *Am J Epidemiol*. 1999;15:916–924.
46. Leslie WD, Miller N, Rogala L, Bernstein CN. Vitamin D status and bone density in recently diagnosed inflammatory bowel disease: the Manitoba IBD Cohort Study. *Am J Gastroenterol*. 2008;103:1451–1459.
47. Seo M, Okada M, Yao T, Ueki M, Arima S, Okumura M. An index of disease activity in patients with ulcerative colitis. *Am J Gastroenterol*. 1992;87:971–976.
48. Seminerio JL, Koutroubakis IE, Ramos-Rivers C, et al. Impact of obesity on the management and clinical course of patients with inflammatory bowel disease. *Inflamm Bowel Dis*. 2015. doi:10.1097/MIB.0000000000000560.
49. Koutroumpakis E, Ramos-Rivers C, Regueiro M, et al. Association between long-term lipid profiles and disease severity in a large cohort of patients with inflammatory bowel disease. *Dig Dis Sci*. 2015. doi:10.1007/s10620-015-3932-1.