

Data Sources

Medicaid is a collection of state-run health insurance plans. We used Medicaid data from 5 of the largest Medicaid plans (Florida, New York, Pennsylvania, Ohio and California) from 1999 to 2012. We used only Medicaid fee for service as recording of diagnoses is less complete in Medicaid managed care plans. We also used two sources of commercial health insurance data. Optum's de-identified Clinformatics® Data Mart Database (CDM) is a collection of anonymized patient-level insurance data. This study included CDM data from 2000 to 2020. We also used data from Carelon Research, a wholly-owned, independently operated subsidiary of Elevance Health, Inc (formerly Anthem, Inc). Carelon data in the Healthcare Integrated Research Database (HIRD®) spanned from 2006 to 2020 and included commercial plan members from across the United States with concentration in fourteen states with Anthem Blue Cross and Blue Shield and affiliated blue plans. There is no overlap of Medicaid, HealthCore and CDM data used in this research.

Each data set contains physician encounter data, including diagnoses recorded using International Classification of Disease 9th edition, Clinical Modification (ICD9-CM) (prior to October 1, 2015) or 10th edition (ICD10-CM). Prescription drug claims are coded using National Drug Codes (NDC) and include quantity dispensed and days supplied. Procedures, including infused medications, are classified using American Medical Association's Current Procedural Terminology (CPT) and Centers for Medicare and Medicaid Services (CMS) Common Procedure Coding System (HCPCS) codes.

For this research, we only included patients with inpatient, outpatient, and prescription medication benefits, with the exception of HIRD where only medical coverage was required for inclusion. Patients were also required to have their date of birth and sex recorded in the database. There is no distinction of sex assigned at birth and gender in the databases. We assumed that the variable aligned with the recording of sex within the United States Census data.

In CDM, race and ethnicity has been collected from public records (e.g., driver's license records) for approximately 30% of individuals and is imputed for the other members using an algorithm based on first and last names and US Census data zip codes (zip + 4). This method is estimated to have 97% specificity, 48% sensitivity, and 71% PPV for predicting the race of black individuals.¹ Although Hispanic ethnicity is distinct from race, it is included in a single variable in these data sets and as such is reported together with race. Only Black, Asian, White and Hispanic are included due to incomplete data and lower accuracy of the variables for other races.² HIRD did not contribute to the ethnicity estimates for this analysis.

Case Ascertainment. Among plan enrollees < 20 years of age, we used a combination of medical diagnoses, diagnostic procedures and prescription medications to identify

prevalent IBD cases, which was previously validated in HIRD and Medicaid data.³ The prevalence algorithm required one IBD diagnosis by any provider regardless of specialty combined with either one or more IBD diagnosis by a gastroenterologist or surgeon or a therapy with a medication for IBD in the absence of another indication. We previously established that the PPV of this algorithm was 94% for two or more diagnoses by a gastroenterologist or surgeon and receipt of IBD specific medications, 92% for two or more diagnoses by a gastroenterologist or surgeon without receipt of IBD medications and 78% for one diagnosis by any provider and receipt of IBD medications. We required a minimum of 4 years of continuous benefits prior to December 31, 2017 for HIRD and CDM data and December 31, 2012 for Medicaid beneficiaries to be included in the estimate of prevalence. We categorized patients as having Crohn's disease vs ulcerative colitis if the most recent diagnosis was the same as the most frequent diagnosis on or before the data of measurement. If the number of Crohn's disease diagnoses equaled the number of ulcerative colitis diagnosis or the most recent diagnosis was not the same as the most common, we categorized the patient as IBD not further specified. Such patients were included in counts of total IBD but not counts of CD or UC individually.

Statistical analysis

To compute pooled prevalence estimates, we first computed age- and sex-specific estimates within each data set and subsequently pooled the commercial insurance data using a fixed effects meta-analytics method. Next, we computed age-, sex- and race-specific insurance-weighted estimates by pooling the data and applying weights proportional to the insurance coverage of Americans based on the 2018 census. Finally, we computed age- and sex- standardized estimates of the national prevalence of IBD using direct standardization to the 2020 USA census data.

Stratum specific variance estimates were computed in the same manner. Weighted strata-specific variance estimates were summed to get the overall variance.

Prevalence estimates stratified by race utilized data from CDM and Medicaid only, as HIRD data did not include race data. Estimates over time were based on data from CDM and HIRD only, as Medicaid data were not available beyond 2012.

In addition to providing prevalence estimates for persons < 20 years of age datasets, we also estimated prevalence among those < 18 of age.

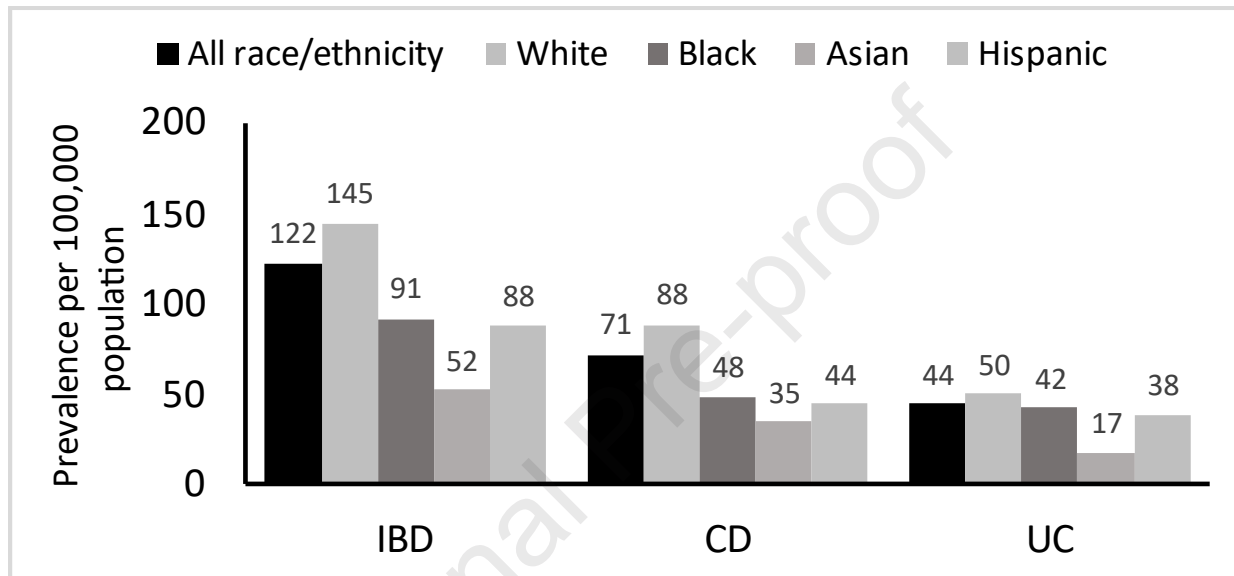
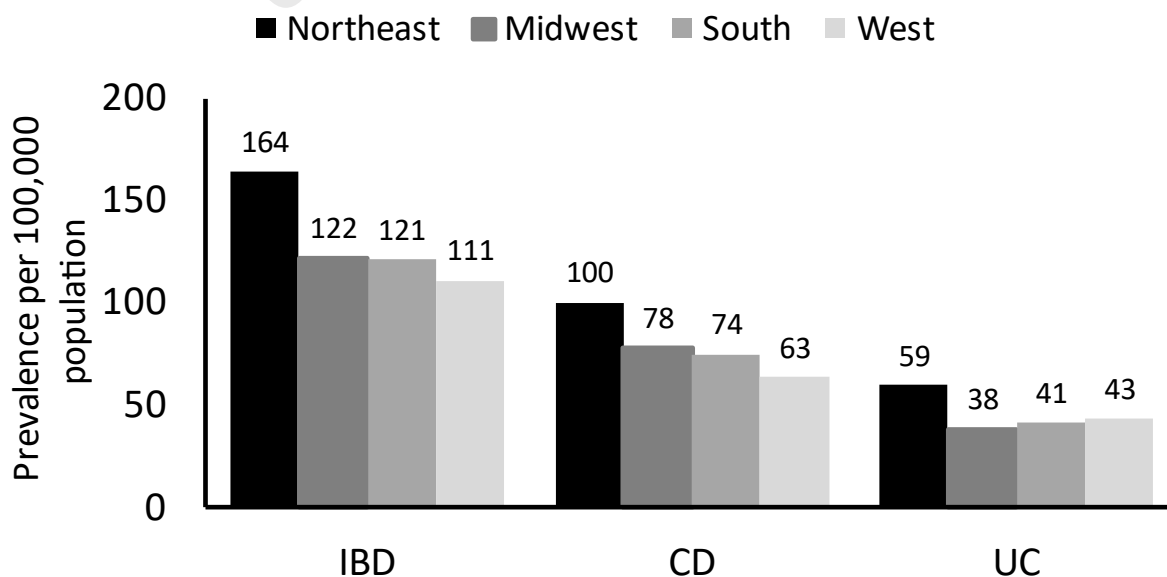
As we did not have access to Medicaid data beyond 2012, we conducted a sensitivity in which we inflated the Medicaid prevalence estimate by the ratio of the prevalence in 2018 to 2011 based on commercial claims and then repeated the pooling procedures.

1. DeFrank JT, Bowling JM, Rimer BK, et al. Triangulating differential nonresponse by race in a telephone survey. *Prev Chronic Dis* 2007;4:A60.
2. Jarrin OF, Nyandege AN, Grafova IB, et al. Validity of Race and Ethnicity Codes in Medicare Administrative Data Compared With Gold-standard Self-reported Race Collected During Routine Home Health Care Visits. *Med Care* 2020;58:e1-e8.
3. Dawwas GK, Weiss A, Constant BD, et al. Development and Validation of Claims-Based Definitions to Identify Incident and Prevalent Inflammatory Bowel Disease in Administrative Healthcare Databases. *Inflamm Bowel Dis* 2023;29:1993-1996.

Supplemental Figure legends

Supplemental Figure 1a. Age- and sex-adjusted prevalence of IBD, Crohn's disease, and ulcerative colitis by race and ethnicity.

Supplemental Figure 1b. Age- and sex-adjusted prevalence of IBD, Crohn's disease, and ulcerative colitis by region.

Supplemental Figure 1a**Supplemental Figure 1b**

Supplemental Table 1 United States Prevalence of and Estimated number of Patients with Pediatric Crohn's Disease, Ulcerative Colitis, and Total IBD (age < 18 years)

CD Prevalence per 100,000	Estimated number in US	UC Prevalence per 100,000	Estimated number in US	Total IBD Prevalence per 100,000	Estimated number in US
57 (53-60)	41,504	36 (33-39)	26,439	98 (94-103)	71,802