Administrative data identify sickle cell disease: A critical review of approaches in U.S. health services research

Scott D. Grosse¹, Nancy S. Green², Sarah L. Reeves³,⁴
¹Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, Atlanta, Georgia
²Department of Pediatrics, Division of Pediatric Hematology, Oncology and Stem Cell Transplantation, Columbia University Medical Center, New York, New York
³Department of Pediatrics, Susan B. Meister Child Health Evaluation and Research Center, University of Michigan School of Medicine, Ann Arbor, Michigan
⁴School of Public Health, Department of Epidemiology, University of Michigan, Ann Arbor, Michigan

Abstract
To identify people living with sickle cell disease (SCD) and study their healthcare utilization, researchers can either use clinical records linked to administrative data or use billing diagnosis codes in stand-alone administrative databases. Correct identification of individuals clinically managed for SCD using diagnosis codes in claims databases is limited by the accuracy of billing codes in outpatient encounters. In this critical review, we assess the strengths and limitations of claims-based SCD case-finding algorithms in stand-alone administrative databases that contain both inpatient and outpatient records. Validation studies conducted using clinical records and newborn screening for confirmation of SCD case status have found that algorithms that require three or more nonpharmacy claims or one inpatient claim plus two or more outpatient claims with SCD codes show acceptable accuracy (positive predictive value and sensitivity) in children and adolescents. Future studies might seek to assess the accuracy of case-finding algorithms over the lifespan.

Keywords
billing codes; diagnostic codes; healthcare use; Medicaid; sickle cell disease

Correspondence: Nancy S. Green, Columbia University Medical Center, 630 West 168 St., Black Building, Box 168, New York, NY 10032. NSG11@cumc.columbia.edu 212-305-0494.
CONFLICTS OF INTEREST
The authors have no conflict of interest to disclose.

Publisher’s Disclaimer: DISCLAIMER
Publisher’s Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
INTRODUCTION

Sickle cell disease (SCD) is an inherited blood condition that is most common in populations in sub-Saharan Africa, the Mediterranean basin, the Middle East, and India. Cost and healthcare use analyses for SCD conducted in the United States commonly rely on administrative healthcare databases, notably hospital discharge and claims databases. Researchers have used multiple inpatient and outpatient encounters from both institutional and noninstitutional providers containing International Classification of Diseases (ICD) billing diagnosis codes for SCD to identify individuals living with SCD and track use of healthcare services such as inpatient and preventive services as well as expenditures. Most have utilized Medicaid data, because Medicaid is the leading payer of SCD-related hospitalizations, followed by Medicare and private insurance. The methodological issues discussed in this focused review apply to claims databases for any or all payers.

Although this review is restricted to U.S. administrative claims and encounters databases, both public and private, that contain records on both inpatient and outpatient clinical encounters to identify people living with SCD, databases that include clinical data can yield more accurate case identification. For example, electronic health records (EHRs) combine encounter records with problem lists of patient diagnoses recorded by clinicians, and studies have demonstrated the accuracy of EHR diagnosis codes for the identification of people living with SCD. Similarly, health system databases can identify those with clinical diagnoses of SCD. This critical review is intended to help researchers choose among algorithms in stand-alone administrative databases lacking clinical or laboratory records to identify individuals, especially children and adolescents, living with SCD.

We focus on two types of case-finding algorithms, although other algorithms are also discussed. The first is a generic health services research (HSR) algorithm that requires either ≥1 inpatient claim with a diagnosis code for a condition or ≥2 outpatient claims (including emergency department [ED] encounters not resulting in admission) on separate days or associated with distinct encounters during a reference period. This generic HSR approach is endorsed by the Centers for Medicare and Medicaid Services (CMS) Chronic Conditions Data Warehouse (CCW) for analyses of Medicare or Medicaid claims data for 16 common chronic conditions, and also 42 other chronic health, mental health, and potentially disabling conditions. The rationale for the requirement of multiple outpatient claims with specified diagnosis codes is that outpatient claims are more subject to coding errors and “rule-out” codes for evaluation visits, laboratory tests, or imaging procedures, whereas hospitals have standardization and quality assurance procedures for coding. A second, novel type of algorithm requires ≥3 claims with diagnosis codes in any setting within a reference period; in June 2019 this approach was endorsed by the CCW for one condition, SCD.

We also summarize the findings of studies that have used diagnoses in medical or laboratory records or newborn screening program data as references to validate SCD case-finding algorithms in children and adolescents using billing codes in administrative databases. The most commonly used measures of the accuracy of case-finding algorithms in general are sensitivity and positive predictive value (PPV); fewer studies report specificity and negative
predictive value.\textsuperscript{16,24} Low sensitivity, the percentage of true cases detected by the algorithm in a validation sample, can adversely affect the representativeness of cases in a study. A high PPV, the proportion of cases identified in an algorithm confirmed to be true positive cases in a validation sample, minimizes misclassification of false positives.\textsuperscript{18} However, PPV may be overstated if prevalence is markedly higher in the validation sample than in the administrative database.\textsuperscript{24} Calculations of specificity and negative predictive value may be subject to overstatement if the validation cohort is enriched with cases, i.e., not representative of the administrative population with diagnosis codes.\textsuperscript{24} Therefore, estimates of specificity and negative predictive in SCD validation analyses are not necessarily comparable. We report in Table 1 all measures using the information available in the original articles, but our focus in the text is on PPV and sensitivity.

2 | ALGORITHMS FOR DETECTION OF SCD

2.1 | Algorithms using ≥1 inpatient claim or ≥2 outpatient claims

The ICD Version 9 Clinical Modification (ICD-9-CM) diagnosis codes for SCD overall are 282.6\textsuperscript{x} and, beginning in 2003, 282.41 and 282.41 for sickle cell-beta thalassemia. The corresponding ICD-10-CM codes, which have been used in U.S. healthcare databases since October 1, 2015, include all D57 codes except D57.3 for sickle cell trait. The ICD-10-CM diagnosis codes include new 6-digit codes for the presence of complications in combination with SCD subtypes.

Numerous peer-reviewed studies that analyzed public or private claims databases have required either 1 inpatient claim with a diagnosis code for SCD or ≥2 outpatient claims with a diagnosis code for SCD on separate days to identify cases of SCD.\textsuperscript{2,3,25–45} One of these studies excluded individuals who also had a claim with a diagnosis code for sickle cell trait,\textsuperscript{40} citing a Wisconsin study that reported that individuals who had diagnosis codes for both SCD (282.60) and sickle cell trait were confirmed to have trait.\textsuperscript{8}

SCD studies have set minimum days between two outpatient claims from 1 to 30 days apart. The number of years of claims data searched to identify SCD cases, i.e., the reference period, has varied across this timespan among the cited studies. The influence of the length of the reference period on the number and characteristics of identified cases using the generic HSR algorithm has not been assessed in the literature. One study that used a 5-year reference period noted that 12% of children who met the SCD case algorithm had no SCD claim in the most recent year; those children had much lower healthcare expenditures.\textsuperscript{2}

Three published reports have reported information on the accuracy of the generic HSR algorithm to identify cases of SCD in claims data using ICD-9-CM codes, two of which were designed as validation studies (Table 1). Reeves et al. linked newborn screening (NBS) program records of children with laboratory-confirmed SCD diagnoses to 12 months of Michigan Medicaid claims data in 2010 or 2011; in both years, an algorithm using ≥1 SCD inpatient claim or ≥2 outpatient claims on separate days had a PPV of 94.5%.\textsuperscript{20} That is the only published study to date that compared the validity of a generic HSR algorithm to the approach of using ≥3 claims in any setting on separate dates to identify SCD cases.
Two studies used Tennessee Medicaid (TennCare) claims and encounters data to identify likely SCD cases with confirmation by either NBS or clinical records. First, Halasa et al. ascertained probable SCD cases based on the presence of SCD codes in 1 inpatient claim or 2 outpatient claims at least 30 days apart. Among 363 children born during 1996 to 2003 who met the algorithm, 312 (PPV 86%) had SCD confirmed in NBS diagnoses. The authors also reported sensitivity of 91%. Supporting evidence comes from a nonvalidation study by Eckrich et al. that confirmed that 88.3% of 653 children who had Medicaid claims linked to medical records at one of two SCD treatment centers in Tennessee and had SCD diagnosis codes in ≥1 SCD inpatient claim or ≥2 outpatient claims at least 30 days apart had SCD.

### 2.2 Algorithms using 3 or more claims with ICD diagnosis codes for SCD

Investigators at the University of Michigan and Michigan Department of Health and Human Services (2014) developed and validated an algorithm based on ≥3 claims on separate dates in any setting and position in a single year of Medicaid claims to ascertain SCD in children. The algorithm was derived from a data-driven process to assess the predictive power of 37 claims-based algorithms for SCD. The authors selected algorithms that reflected combinations of settings (inpatient, outpatient, home health care, ED, blood transfusion), medication categories (antibiotic prophylaxis, hydroxyurea), evaluation or consultation claims, and an overall count of SCD claims, irrespective of type of service. The gold standard of NBS program records was linked to Michigan Medicaid claims data during 2010 to compare the accuracy of all 37 algorithms. The receiver-operating curve, which balances sensitivity and specificity, was high for four algorithms; an algorithm requiring ≥3 SCD claims in any setting and position in 12 months resulted in a PPV of 95.0%, which was slightly higher than for the generic HSR algorithm, 94.5%. Reeves et al. validated both algorithms with 2011 Medicaid claims and found PPVs of 95.8% and 94.5%, respectively, although the HSR algorithm had slightly higher sensitivity, 90.2% vs 89.7%. Reeves et al. argued that the algorithm requiring ≥3 SCD claims in any setting had the advantages of simplicity and ease of calculation.

The Sickle Cell Data Collection (SCDC) program is a collaboration with the Centers for Disease Control and Prevention (CDC) in two states to date, California and Georgia. The SCDC has used the presence of SCD diagnosis codes in 3 or more claims or encounters in a 5-year reference period to ascertain SCD cases. In California, Paulukonis et al. used ≥3 clinical encounters in administrative databases (Medicaid claims and state hospital and ED discharges) with SCD ICD-9-CM during a 5-year period, 2004 to 2008, to identify probable SCD cases. In Georgia, Snyder et al. reviewed medical, laboratory, and newborn screening program records for individuals ≤21 years of age seen at Children’s Healthcare of Atlanta during 2004 to 2008 with medical records linked to three administrative claims and encounters databases (Medicaid, Children’s Health Insurance Program, and State Health Benefit Plan) for the same 5-year surveillance period. The authors assessed the PPV and sensitivity of 12 administrative case definitions based on from ≥1 up to 6 SCD-related encounters, defined as nonpharmacy medical claims (including laboratory and radiology claims) on separate service dates within a 5-year reference period (no restriction for continuous enrollment). In addition, they assessed a 13th algorithm, discussed in the next section. The PPV increased monotonically with the number of encounters, from 90.0% to
99.0%, and sensitivity decreased from 100.0% to 90.0%. Using ≥3 or more encounters as the criterion, sensitivity was 96.0% and PPV was 97.4%, compared with 98.4% and 94.8% for ≥2 or more encounters. The PPV was unchanged when the surveillance period was reduced from 5 years to a 12-month period within adjoining calendar years, although sensitivity was slightly reduced.

In June 2019, CMS endorsed a CCW case-finding algorithm for SCD requiring ≥3 claims on separate dates with ICD-9-CM or ICD-10-CM codes for SCD within 5 years of look-back data,50,51 In place of a single 5-year look-back period ending with the year for which services or expenditures are assessed, the current SCDC approach uses a rolling 5-year period to assess SCD case status.52,53 No restriction is placed on minimum length of continuous enrollment.

2.3 | Combinations of diagnosis codes with SCD-associated treatments, procedures, and complications

Some investigators have considered using procedure or drug codes in addition to SCD diagnosis codes to identify probable SCD cases in claims data. Paulukonis et al. proposed the combination of an ICD-9-CM code for SCD in at least two healthcare encounters, independent of setting, and at least one code for an associated treatment, procedure, or complication of SCD,54 an approach recently adapted by other researchers.55,56 However, Snyder et al. reported that algorithm had a false-negative rate (1 minus sensitivity) of 14.2%, compared with 4.0% for one requiring 3 diagnostic claims.23

3 | DISCUSSION

Administrative data are often used for population-level assessments of utilization of care or expenditures, especially for conditions with low prevalence. For example, U.S. insurance claims databases, both Medicaid and private insurance, have been used to estimate medical costs,2,3,32,33 uptake of antibiotic prophylaxis,19,26,28,31,34,48 documented receipt of immunizations,31,34,47,57 use of hydroxyurea,55,57,58–63 and receipt of transcranial doppler (TCD) screening among individuals with SCD.27,31,38,46,47,64–66 Some of those studies, especially ones published in the past 5 years, merged NBS or clinical databases, which were used to identify cases of SCD, with linked claims data that were used to track utilization of services.37,58,64–66 Some of those studies used clinical or NBS records to identify cases of sickle cell anemia associated with homozygous sickle disease (HbSS) or hemoglobin S-beta thalassemia0 to calculate quality indicators for preventive services with recommendations specific to sickle cell anemia, e.g., TCD screening and hydroxyurea.58,60,64–66 It is challenging to identify cases of SCD subtypes using administrative data.

Use of 2 years of data to ascertain cases of certain chronic conditions has been reported to improve performance relative to a single year of data,67,68 which is consistent with the default CCW algorithm for chronic conditions. Snyder et al. found similar predictive value for SCD requiring multiple claims to occur within either 5 years or a 12-month period within adjoining calendar years, not limited to continuous enrollment, although the authors cautioned that their results might not be generalizable.23 Use of multiple years of claims data may lead to improved identification of individuals with milder disease phenotypes who
have fewer healthcare encounters. Amendah et al. reported that among children with SCD identified using 5 years of either Medicaid or Commercial claims data, 12% had no SCD claims during the fifth year despite continuous enrollment during that year.²

Algorithms using ≥3 claims in any setting to identify children with SCD were found in the Michigan study to have a slightly higher PPV than an algorithm that used ≥1 inpatient claims or ≥2 outpatient (including ED) claims on separate days, although the sensitivity of each algorithm was comparable (Table 1).²⁰ In the validation study using Georgia pediatric data, an algorithm requiring ≥3 claims in any setting with a SCD code had a higher PPV than an algorithm requiring ≥2 claims with SCD codes. However, the advantage of requiring 3 claims in both studies was modest in magnitude; other algorithms may have similar performance.

SCD diagnosis codes in inpatient claims were more predictive of SCD case status in the Michigan study than were SCD codes in outpatient claims.²⁰ A study using data from a children’s hospital found that only a minority of individuals who had a single admission with a SCD code and no outpatient SCD encounters had SCD.²³ However, that finding cannot be generalized to claims data because false-positive SCD diagnoses in inpatient settings are more common in statewide claims data than in records from children’s hospitals.²²

The two published studies that sought to compare and validate multiple claims-based SCD case-finding algorithms each have limits to generalizability. One limitation is that the PPV in Medicaid claims data, as used in the Michigan study, may be higher than in employer-sponsored insurance. The Georgia study also included CHIP data and claims data for state government employees and their dependents but did not separately evaluate PPV by plan type.

An important limitation is that the SCD validation studies only included pediatric subjects. The Michigan study was restricted to children up to 18 years of age, and the Georgia study was restricted to individuals up to 21 years of age. It is unknown whether billing codes are equally predictive of SCD case status in adults. One study reported that ICD-9-CM codes for SCD were more predictive of case status in pediatric hospital EHRs than in adult hospital EHRs,⁸ but similar comparisons have not been made with claims data.

Validation data sets can be either population-based or provider-based. Findings may be more generalizable from the Michigan study, which encompassed Medicaid enrollees managed by all types of providers in the state, than the Georgia study, whose validation data were restricted to individuals seen at a children’s hospital with an SCD clinic and affiliated facilities.

A final consideration is the number of years of data used to validate claims. As already mentioned, health services researchers have frequently found that use of more than 1 year of claims data, specifically at least 2 years, improves the accuracy of case-finding algorithms. The Georgia study analyzed 5 years of claims data and compared the accuracy of algorithms using 2 calendar years (12 months from first claim) versus 5 calendar years. In contrast, the Michigan study used a single calendar year of data to identify SCD cases, 2010, which it
replicated with a separate analysis of 2011 claims data. It did not assess the accuracy of algorithms pooling the 2 years of available claims data.

In June 2019, CMS endorsed a CCW case-finding algorithm for SCD requiring ≥3 claims on separate dates with ICD-9-CM or ICD-10-CM codes for SCD within 5 years of look-back data and used it to estimate the prevalence of SCD in Medicaid and Medicare claims data. The CCW condition algorithm is the same as the SCDC algorithm used by Snyder et al., including use of the most recent 5 years of claims data as the reference period and exclusion of pharmacy claims, even though CMS pharmacy claims do not contain diagnosis codes. In place of a single 5-year look-back period ending with the year for which services or expenditures are assessed, the current SCDC approach uses a rolling 5-year period to assess SCD case status. No restriction is placed on minimum length of continuous enrollment.

Some researchers use the presence of ≥1 claim in any setting with an SCD diagnosis code to identify putative SCD cases. However, owing to the frequency of false-positive diagnoses in outpatient claims, that approach can lead to misclassification and result in underestimation of the use of services or costs among individuals with SCD.

4 | CONCLUSION

Researchers can use stand-alone administrative databases for research on healthcare utilization among persons living with SCD. Evidence from validation studies conducted using data for children and adolescents indicates that algorithms that require multiple SCD codes, particularly in records of outpatient claims or encounters, can yield acceptable accuracy of SCD case ascertainment, although the accuracy may vary across the lifespan. Researchers can decide which algorithm best suits their study purposes, e.g., the assessment of uptake of services recommended for those with specific SCD subtypes, taking into account the number of years of available data.

ACKNOWLEDGMENTS

This work received support from NHLBI grant funding 5K01HL132057 (Reeves). Helpful comments on previous versions were provided by Thomas Adamkiewicz, MD, and Amanda B. Payne, MPH.

Funding information

National Heart, Lung, and Blood Institute, Grant/Award Number: 5K01HL132057

Abbreviations:

CCW Chronic Conditions Data Warehouse
CDC Centers for Disease Control and Prevention
CMS Centers for Medicare and Medicaid Services
ED emergency department
EHR electronic health record

Pediatr Blood Cancer. Author manuscript; available in PMC 2021 December 01.
REFERENCES


54. Paulukonis ST, Harris WT, Coates TD, et al. Population based surveillance in sickle cell disease: methods, findings and implications from the California registry and surveillance system


### TABLE 1
Assessments of the validity of ICD-9-CM–based algorithms for identifying pediatric SCD in U.S. healthcare databases

<table>
<thead>
<tr>
<th>Author</th>
<th>Case definition</th>
<th>Study period and reference period for algorithm</th>
<th>Age group</th>
<th>Source of ICD-9-CM data</th>
<th>Validation data source</th>
<th>Validation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halasa et al. (2007)</td>
<td>1 inpatient claim or ≥2 outpatient claims ≥30 days apart with a diagnosis code for SCD Note: ICD-9-CM codes of 282.41 and 282.42 were not included</td>
<td>10 years, 1995–2004, for both</td>
<td>Children, born in state, 1996–2003</td>
<td>Tennessee Medicaid claims, 1995–2004</td>
<td>Tennessee NBS data, 1996–2003</td>
<td>PPV 86.0% (312 confirmed cases out of 363 who met SCD case-finding algorithm) Sensitivity 91.2% (312 true positives and 32 false negatives among 344 children with SCD confirmed by NBS records)</td>
</tr>
<tr>
<td>Reeves et al. (2014)</td>
<td>1 inpatient claim or ≥2 outpatient claims 1 day apart with SCD codes during 12-month period</td>
<td>1 year, 2010 or 2011, for both</td>
<td>Children, ages 0–18 years, born in state</td>
<td>Michigan Medicaid claims, 2010 and 2011</td>
<td>Michigan NBS data, 1987–2010</td>
<td>2010 data: PPV 94.5% Sensitivity 90.2% Specificity 90.4% 2011 data PPV 94.5% Sensitivity 90.2% Specificity 90.4%</td>
</tr>
<tr>
<td>Reeves et al. (2014)</td>
<td>Presence of ≥2 claims (any setting) on separate dates with SCD codes during 12-month period</td>
<td>1 year, 2010 or 2011, for both</td>
<td>Children, ages 0–18 years</td>
<td>Michigan Medicaid claims, 2010 and 2011</td>
<td>Michigan NBS data, 1987–2010</td>
<td>2010 data: PPV 95.0% Sensitivity 90.7% Specificity 91.3% 2011 data: PPV 95.8% Sensitivity 89.7% Specificity 87.9%</td>
</tr>
<tr>
<td>Snyder et al. (2019)</td>
<td>Presence of ≥2 claims any setting on separate dates with SCD codes during the 5-year period Additional ICD-9-CM code included: 282.6</td>
<td>5 years, 2004–2008, for both</td>
<td>Children and young adults, ages 0–21 years</td>
<td>Georgia Medicaid, Children’s Health Insurance Program, State Health Benefit Plan, Georgia Hospital discharge data, 2004–2008</td>
<td>Children’s Healthcare of Atlanta medical and laboratory records, Georgia NBS records</td>
<td>PPV: 97.4%, Sensitivity: 96.0%, Specificity: 76.5%, NPV: 68.2%</td>
</tr>
<tr>
<td>Snyder et al. (2019)</td>
<td>≥2 claims on separate dates with SCD ICD-9-CM codes and ≥1 claim with code for an SCD-associated treatment, procedure, or complication Additional ICD-9-CM code included: 282.6</td>
<td>5 years, 2004–2008, for both</td>
<td>Children and young adults, ages 0–21 years</td>
<td>Georgia Medicaid, Children’s Health Insurance Program, State Health Benefit Plan, Georgia Hospital discharge data, 2004–2008</td>
<td>Children’s Healthcare of Atlanta medical and laboratory records, Georgia NBS records</td>
<td>PPV: 97.4%, Sensitivity: 85.8%, Specificity: 79.0%, NPV: 38.2%</td>
</tr>
</tbody>
</table>

Abbreviations: ED, emergency department; ICD-9-CM, International Classification of Diseases Version 9 Clinical Modification; NBS, newborn screening; NPV, negative predictive value; PPV, positive predictive value; SCD, sickle cell disease

\(^{\text{a}}\) All case definitions include the following ICD-9 codes unless otherwise indicated: 282.60–282.69 and 282.41,282.42.