REGULAR ARTICLE

blood advances

Creating an automated contemporaneous cohort in sickle cell anemia to predict survival after disease-modifying therapy

Robert M. Cronin,[1] Kristin Wuichet,[2] Djamila L Ghafuri,[2] Brock Hodges,[2] Maya Chopra,[2] Jing He,[3] Xinnan Niu,[3] Adetola A. Kassim,[4] Karina Wilkerson,[4] Mark Rodeghier,[5] and Michael R. DeBaun[2]

[1]Department of Internal Medicine, The Ohio State University, Columbus, OH; [2]Department of Pediatrics, [3]Department of Biomedical Informatics, and [4]Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, TN; and [5]Rodeghier Consultants, Chicago, IL

Key Points

- To estimate survival, an automated contemporaneous cohort of children and adults with SCA is feasible and efficient.
- Hydroxyurea therapy for at least 1 year is associated with increased survival in adults with SCA compared with no disease-modifying therapy.

The Food and Drug Administration requires contemporaneous controls to compare clinical outcomes for participants receiving experimental gene therapy or gene editing clinical trials. However, developing a contemporaneous cohort of rare diseases requires multiple person-hours. In a single referral center for sickle cell disease, we tested the hypothesis that we could create an automated contemporaneous cohort of children and adults with sickle cell anemia (SCA) to predict mortality. Data were obtained between 1 January 2004 and 30 April 2021. We identified 419 individuals with SCA with consistent medical care defined as followed continuously for >0.5 years with no visit gaps >3.0 years. The median age was 10.2 years (IQR, 1-24 years), with a median follow-up of 7.4 years (IQR, 3.6-13.5 years) and 47 deaths. A total of 98% (274 of 277) of the children remained alive at 18 years of age, and 34.3% (94 of 274) of those children were followed into adulthood. For adults, the median age of survival was 49.3 years. Treatment groups were mutually exclusive and in a hierarchical order: hematopoietic stem cell transplant (n = 22)>regular blood transfusion for at least 2 years (n = 56)>hydroxyurea for at least 1 year (n = 243)>no disease-modifying therapy (n = 98). Compared to those receiving no disease-modifying treatment, those treated with hydroxyurea therapy had a significantly lower hazard of mortality (hazard ratio = 0.38; P = 0.016), but no statistical difference for those receiving regular blood transfusions compared to no disease-modifying therapy (hazard ratio = 0.71; P = 0.440). An automated contemporaneous SCA cohort can be generated to estimate mortality in children and adults with SCA.

Introduction

Myeloablative gene therapy and gene editing expand curative options for children and adults with sickle cell anemia (SCA). With the initiation of the novel cellular therapies, the Food and Drug Administration (FDA) requires a contemporaneous group receiving disease-modifying therapy to compare key clinical outcomes, including mortality. Developing a contemporaneous cohort of rare diseases is costly, inefficient, and requires many person-hours to construct and maintain it. Furthermore, the FDA requires a cellular therapy treatment group to have a follow up for at least 15 years after completion of gene therapy or gene editing trials.[1] More efficient strategies are required to efficiently and effectively develop

Submitted 9 August 2022; accepted 28 October 2022; prepublished online on *Blood Advances* First Edition 9 November 2022; final version published online 26 July 2023. https://doi.org/10.1182/bloodadvances.2022008692.

Data are available on request from the corresponding author, Michael R. DeBaun (m.debaun@vumc.org).

The full-text version of this article contains a data supplement.

contemporaneous comparison groups to comply with FDA requirements for novel treatments of rare diseases.

The widespread use of electronic health record systems creates new opportunities for the secondary use of big data from patient medical records. Options for rare disease clinical history typically require years of following up from a single center or a consortium of multiple centers. Clinical data warehouses of electronic health record data are available at many health care systems across the United States. These data warehouses contain clinical information from multiple years of medical records. Using clinical and laboratory data in a warehouse for creating a cohort of a rare genetic disease is emerging as a reasonable alternative to determine the clinical history of a rare disease.[2-6] Our objective was to close the knowledge gap of constructing an automated contemporaneous cohort in children and adults with SCA to evaluate the impact of disease-modifying treatment on mortality.

In keeping with the overall goal of the Cure Sickle Cell Initiative, a collaborative patient-focused research effort designed to accelerate the advancement of genetic-based cures for sickle cell disease (SCD), we used the clinical data warehouse of electronic health records at the Vanderbilt University Medical Center (VUMC). We tested the hypothesis that an automated contemporaneous cohort of children and adults with SCA could be developed to predict the mortality across different disease-modifying treatment options.

Methods

Study design and participants

We conducted an observational study using VUMC's electronic health record data. The VUMC institutional review board approved this study. It was conducted in accordance with the Declaration of Helsinki.

Methods for the full automated cohort development of children and adults with SCA. We developed an automated cohort using algorithms to identify individuals with SCA, defined as having either hemoglobin SS (HbSS) or HbS β thalassemia^0 phenotype, in the electronic health record to create our cohort. These algorithms were adapted from other published algorithms.[5,7,8] The algorithm is described in the supplemental Appendix (supplemental Table 3; supplemental Figures 1-3; supplemental Appendix Methods). Individuals with SCA were identified from the VUMC Research Derivative, a fully identified clinical data warehouse drawn from the VUMC electronic health record. Individuals were included if there was an opportunity to prescribe a disease-modifying therapy such as hydroxyurea therapy or blood transfusion. Individuals in the cohort had SCA and consistent medical care within our system, defined as being seen for more than 0.5 years, and not having an interval greater than 3.0 years between consecutive visits (Figure 1). The full details of the automated cohort creation are listed in the supplemental Appendix (supplemental Table 10; supplemental Figures 1-3; supplemental Appendix Methods). Descriptions of the data extraction process for key data elements can be found in supplemental Table 6. Data were obtained from VUMC between 1 January 2004 and 30 April 2021, which included a cohort of more than 3 million individuals evaluated at the medical center during that period.

Methods for the hybrid cohort development (automated component and manual electronic health record review) of children and adults with SCA. We also used another method to develop a cohort of individuals followed at VUMC to compare clinical features and overall mortality rates with the fully automated cohort. We refer to this cohort as the hybrid cohort because the construction included a low-level automated component that can be done without advanced computer programmer

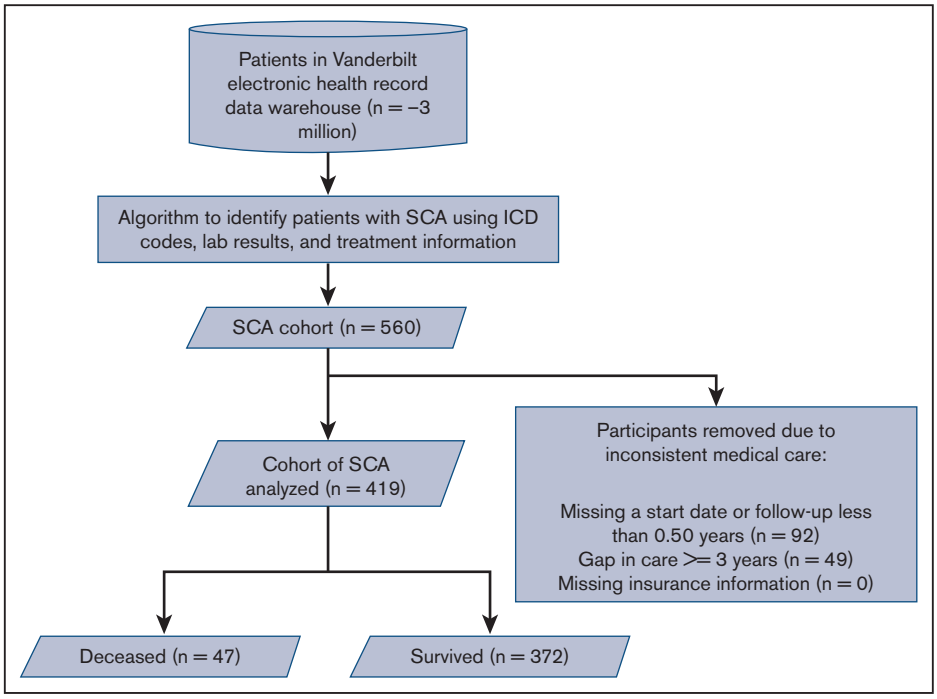

Figure 1. CONSORT flow diagram of the automated SCA cohort from our electronic health record with exclusion criteria. SCA, sickle cell anemia (SS and Sbeta0); ICD, International Classification of Disease.

skills and a manual review of the electronic health record compared with the fully automated cohort. The essential covariates of the automated component of the hybrid cohort, deemed likely to be generalizable to most electronic health record systems, included age, sex, follow-up time, the gap between visits, and prescriptions for hydroxyurea. The key covariates of the manual chart review included the presence and type of SCA, disease-modifying treatment type, such as chronic transfusion and transplant, and mortality.

Death as the primary outcome measure for both automated and hybrid cohorts of children and adults with SCA. Our primary outcome measure to test our hypothesis was mortality. To determine the patient's vital status (alive or dead) for the automated cohort, we used structured fields for vital status and dates of death information automatically extracted from our data warehouse when available. These data fields were derived from multiple sources, including electronic health records; the social security death index, which contains only the fact of death and was available on 31 December 2016 in the data warehouse; or an external death data source, Datavant, an organization that augments the social security administration death master files with information from newspapers, funeral homes, and memorials to construct an individual-level database of >80% of US deaths annually.[9] When a deceased person did not have a death date available from automated methods, we used the last visit date as a proxy for the date of death.

For the hybrid cohort, the vital status and date of death included only those recorded in the electronic health record. When the fact of death was available without a date of death, the last day of follow up in the clinic was substituted as the date of death. Further descriptions of the automated and manual data collection methods are described in Appendix supplemental Table 10.

Estimate of person-hours allocated to the development of automated and hybrid cohorts of children and adults with SCA. We estimated the number of person-hours in the program based on the personnel allocated to the grant that worked on the project from 2019 to 2021. The research personnel included 2 research coordinators and 2 data analysts with research computer programming skills to collect information from the data warehouse. The team also included 1 senior project manager and 2 physicians.

Clinical and laboratory data

From our data warehouse, we obtained demographic variables, insurance status, laboratory results, dates of the first and last hematology encounters within our health care system, medications, procedures such as transfusions and hematopoietic stem cell transplants (HSCTs), and the date of death. The demographic variables included the date of birth and sex. Laboratory results from Hb fractionation analyses (eg, isoelectric focusing or high-performance liquid chromatography) were used in automated algorithms to determine the sickle cell type based on standard cutoffs determined by previous literature.[10,11] Medications, regular red blood cell transfusions, and HSCTs were obtained using computerized algorithms. A participant was considered to have hydroxyurea therapy if they had 1 year of prescriptions for hydroxyurea in our electronic health record, which was defined as being followed for 1 year after initiating hydroxyurea prescriptions. A participant was considered to receive regular blood transfusion therapy if the participant had at least 9 transfusions in 12 months for 2 years.[12,13] If 2 or more transfusions occurred within 13 days, they were counted as only 1 transfusion. The coding of the type of disease-modifying therapy was mutually exclusive and hierarchical with treatment priorities: HSCT > blood transfusion > hydroxyurea therapy; for example, a participant who had been on hydroxyurea and chronic transfusion is coded as transfusion. All adults receiving HSCT had myeloablative-matched related donor or reduced intensity–related haploidentical HSCT.[14] Total follow-up time was calculated from the first hematology encounter or first disease-modifying treatment (ie, hydroxyurea prescription, chronic transfusion, or transplant) until death or the last hematology visit.

Statistical analysis

Data were summarized as counts and percentages for categorical variables, and depending on the data distribution, as either means with standard deviations or medians and interquartile range for continuous variables. Cumulative incidence curves were generated using the Kaplan–Meier method. To identify associations with mortality, we constructed a baseline multivariable Cox regression model with time-varying effects for treatment type. Mortality was adjusted for age at the start date, sex, Medicaid status, and the type of treatment in our models. A patient was classified as either having Medicaid insurance for none of the visits (0%), all visits (100%), or some visits (1%-99%). The proportional hazards assumption was assessed, and Martingale and deviance residuals were used to assess model assumptions and the effect of outlier cases. $P < .05$ was considered significant for all analyses. Statistical Package for the Social Sciences version 27 (IBM Corp, Armonk, NY) was used to perform all analyses. All primary analyses were completed with the automated cohort, in which all data were obtained automatically using computer programming scripts. Additional sensitivity analyses were done with a hybrid cohort, which included automated and manually checked data and are presented in the supplemental Appendix.

Results

More than 3 million unique participants were represented in the VUMC data warehouse during the sampling interval of 1 January 2004 and 30 April 2021. Our automated methods identified 988 participants with SCD (HbSS, HbSC, and other SCD compound heterozygotes). Our algorithms identified an SCD phenotype for 883 of these 988 participants, and 560 were designated to have SCA (Figure 1), demonstrating a 63% proportion of all adults with SCD. We removed 92 participants without visit dates during the assessment period or follow up of fewer than 0.50 years and 49 participants with a gap in the care of 3 or more years. Our final cohort of 419 participants with SCA had a median age of 10.2 years at the first visit, 49.9% were male, and had a median follow up of 7.4 years. A total of 25.8% (108 of 419), 37.5% (157 of 419), and 36.8% (154 of 419) had Medicaid insurance for none of the visits, all of the visits, and some of the visits, respectively (Table 1). The incidence rate of death was 0.15 per 100 patient-years for children younger than 18 years (95% confidence interval [CI], 0.03-0.44) and 2.89 per 100 patient-years for adults older than 18 years (95% CI, 2.01-3.88). Among adults, those 18 to 36 years had an incidence rate of death of 2.36 per 100 patient-years and those older than 36 years (95% CI, 1.57-3.42) had an

Table 1. Baseline characteristics and primary outcome by type of therapy (n = 419).

Variable	No disease-modifying treatment (n = 98)	Hydroxyurea (n = 243)	Transfusion (n = 56)	HSCT (n = 22)	P*
Age at start of follow up (y), median (IQR)	11.5 (1.1-27.6)	9.0 (0.5-21.4)	8.8 (5.0-21.5)	20.5 (13.2-27.3)	.032
Sex (male), n (%)	43 (43.9)	119 (49.0)	32 (57.1)	15 (68.2)	.132
Follow-up time (y), median (IQR)	2.5 (1.1-5.7)	9.1 (4.9-14.0)	11.7 (7.2-16.5)	4.2 (2.8-11.4)	<.001
Largest gap in care (y), median (IQR)	0.6 (0.3-1.0)	0.8 (0.5-1.2)	0.6 (0.3-1.0)	0.6 (0.5-1.0)	.002
Use of Medicaid, n (%)					<.001
0%	41 (41.8)	50 (20.6)	8 (14.3)	9 (40.9)	
1%-99%	22 (22.4)	99 (40.7)	27 (48.2)	9 (40.9)	
100%	35 (35.7)	94 (38.7)	21 (37.5)	4 (18.2)	
Death, n (%)	12 (12.2)	23 (9.5)	12 (21.4)	0 (0.0)	.027†
Rate of death/100 patient-years (95% CI)	3.01 (1.56-5.26)	1.00 (0.64-1.51)	1.82 (0.94-3.18)	0.0 (not applicable)	<.001‡

IQR, interquartile range.
*χ^2 test for counts and Kruskal–Wallis test for medians, unless otherwise noted.
†Fisher exact test.
‡Poisson regression.

incidence rate of death of 4.72 per 100 patient-years (95% CI, 2.70-7.67). A majority received a hydroxyurea prescription (58.0%) for at least a year, significantly fewer individuals received at least 1 year of regular blood transfusions for at least 2 years (13.3%), and only a small percentage of individuals received an HSCT (5.3%). Baseline laboratory values for the automated cohort are presented in supplemental Table 1.

Receiving hydroxyurea was associated with significantly lower mortality

Hydroxyurea therapy, transfusions, and transplant had significantly lower mortality than no disease-modifying treatment ($P \leq .001$, $P = .011$, and $P = .007$, respectively). None of the 3 treatments had significantly different survival compared with the other 2. A Kaplan–Meier curve was constructed to evaluate the different treatments (Figure 2). A Cox regression model was then constructed using sex, age at the first visit, Medicaid insurance status throughout the sampling frame (0%, 1%-99%, and 100%, respectively), and time-varying treatment type, with the same categories for treatment and with no disease-modifying treatment used as reference categories in the model (Table 2). Age was associated with increased mortality (hazard ratio [HR], 1.09; $P < .001$). Compared with those on no disease-modifying treatment, participants prescribed with hydroxyurea therapy for at least a year had a decreased hazard of mortality (HR, 0.38; $P = .016$). Medicaid insurance status was associated with an increased hazard of mortality (Medicaid 1%-99%: HR, 3.05; $P = .038$ and Medicaid 100%: HR, 7.44; $P \leq .001$) compared with the reference category of never having Medicaid insurance. No statistical difference was observed between those treated with regular transfusions and those who did not receive disease-modifying therapy (HR, 0.71; $P = .440$). Adults receiving HSCT had the highest survival (100%), but they were the lowest number of individuals (n = 22), and the regression coefficients could not be reliably estimated because there were no deaths.

The automated cohort had more extensive information than the hybrid cohort

We identified 432 individuals with SCA based on a manual electronic health record review. Of the 432 patients, 383 (81.8%) were also in the automated cohort and 49 were only in the hybrid cohort. There were 36 patients who were only in the automated cohort and not the hybrid cohort. For the primary outcome measure, mortality, the automated cohort had 47 deaths, whereas the hybrid cohort only had 29 deaths. The median survival age in the automated and hybrid cohorts, using only those who entered as adults, was 49.3 years (95% CI, 42.9-54.9) and 46.1 years (95% CI, 43.0-58.2), respectively. The demographic and laboratory features of the automated cohort were similar to those of the hybrid cohort (supplemental Tables 2-4). However, the follow-up time, the proportion of deaths, and the incidence rate of death differed on when comparing the automated with the hybrid cohort (supplemental Table 2).

The automated cohort required more time to create than the hybrid cohort

The automated approach required ~6500 hours (Table 3). Two programmers worked efficiently at ~35% effort for more than 2 years (4000 hours) and 1 supervisor at 40% effort for more than 2 years (2500 hours). Two experienced data extractors worked almost full-time for more than 2 years to pull the manual record in more than 1000 individuals with SCD for an estimated ~5350 hours. The data extractors were supervised by a project manager and 2 experienced clinicians for ~200 hours for more than 2 years.

Discussion

Myeloablative gene therapy and gene editing clinical trials have expanded curative therapy options for individuals with SCA; however, the FDA mandates contemporaneous comparison groups to demonstrate the relative benefits and harms of these novel therapies. Traditionally, creating contemporaneous cohorts has been done through manually curated SCA registries; however, these registries are expensive to initiate and even more challenging to maintain. As a proof of principle, we demonstrated the feasibility that the medical record data warehouse of a large single medical center could be used to create a contemporaneous cohort of children and adults with SCA. Furthermore, this cohort can be used to assess the mortality rate as a comparator group to

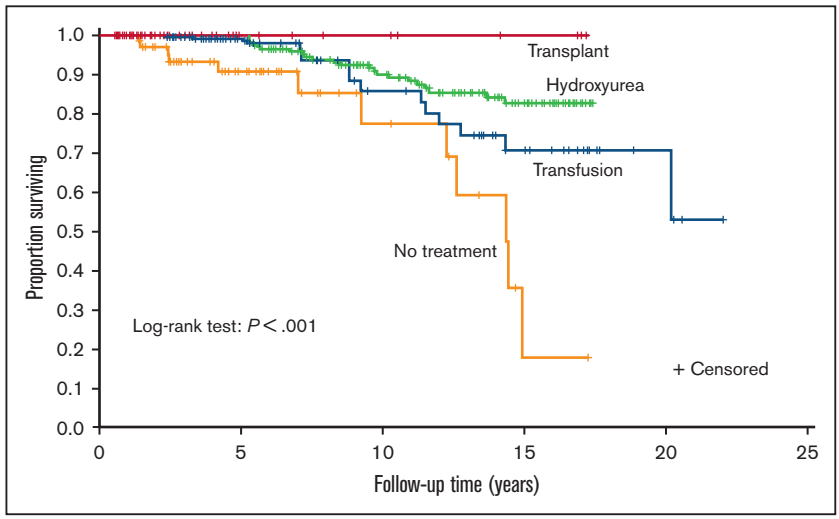

Figure 2. Kaplan–Meier curve of disease-modifying therapy for the cohort of children and adults with SCA. A Kaplan–Meier curve was constructed to evaluate the different treatments. Treatment groups were mutually exclusive and in a hierarchical order: HSCT (n = 22) > regular blood transfusion for at least 2 years (n = 56) > hydroxyurea for at least 1 year (n = 243) > no disease-modifying therapy (n = 98).

experimental cellular therapies. In our cohort, those with HSCTs had the highest survival, although only ~5% of the total SCA cohort received this curative therapy. Individuals receiving at least 1 year of hydroxyurea therapy, ~60% of the cohort, had a higher survival than individuals receiving no disease-modifying treatment, ~23% of the cohort. Importantly, our study cohort of individuals with SCA consisted of ~63% of the expected proportion of individuals with SCA in large cohort studies.[15,16]

Our automated approach estimates similar overall and median survival as previous hybrid or manual data extraction approaches from other medical centers and our own. Our results for the median age of survival in children and adults with SCA are similar to those of prior studies.[17,18] In one of those studies, the median age of survival (49.3 years) of our automated cohort was similar to that of a recent manually constructed cohort (49.2 years).[17] Also, the overall survival to 18 years of our cohort was 98%, similar to the expected mortality rates in other large cohort studies conducted in high-income settings.[19-21]

Our study showed a similar survival advantage as other studies for adults with SCA prescribed with hydroxyurea compared with those not prescribed with hydroxyurea.[22-25] Previously, using a manual chart review approach, our VUMC team showed that hydroxyurea therapy was not associated with improved survival in adults with SCA;[17] however, those analyses differed from our current approach. Specific differences between the 2 cohorts, which may account for these discordant results may be related to the approach for classifying individuals as being treated with hydroxyurea and the mutually exclusive hierarchal modeling strategy. Our results are similar to observational studies with a median follow up of at least 10 years showing improved survival in children[19] and adults[22,24,26] treated with hydroxyurea compared with individuals not treated with hydroxyurea.

Table 2. Cox regression models for mortality with demographic and therapy covariates in the cohort of children and adults with SCA (n = 419)

Variable	HR	95% CI	*P*
Age at the first visit, y	1.086	1.07-1.11	.000
Sex (male)	1.610	1.61-3.00	.133
Therapy*			
Hydroxyurea	0.379	0.17-0.83	.016
Transfusion	0.705	0.29-1.71	.440
Transplant	0.000	0.00 - undefined	.975
Medicaid†			
1%-99% of visits	3.052	1.06-8.76	.038
100% of visits	7.440	2.44-22.72	.000

*Reference category is no therapy.
†Reference category is 0% of Medicaid visits.

An important clinical finding in our regression analysis is the high HR for death in individuals receiving Medicaid insurance compared with adults who always had another insurance type (eg, Medicare, other dual plans, self-pay, or private insurance). Others have shown that the presence of Medicaid and lower socioeconomic status are associated with mortality in adults with SCA.[27,28] Together, these studies emphasize the importance of screening adults with Medicaid coverage to assess other social determinants of health that may put these individuals at risk for earlier mortality.

Our approach to creating an automated contemporaneous cohort had strengths and weaknesses and lessons learned (Table 3). First, we were able to pull the clinical information over 17 years from a large health care system which includes multiple hospitals. Other individuals in small or large health care systems may be able to replicate or extend our tools to mimic our approach for assessment of mortality rate of their institution in their health care system. Second, our work builds on the extensive prior data of automated cohorts in SCD.[5,7,29-32] However, to our knowledge, the integration of disease-modifying therapies and mortality has not been included in other automated cohort studies. Our tools also utilized more information from the electronic health records than existing approaches from administrative data sets which primarily use billing codes. Third, the initial personal-hours investment does not require significant future annual personnel expenditure to estimate the vital status of the cohort. The most significant challenge for establishing an automated cohort is the substantial initial personnel time, informatics expertise, and required SCD clinical expertise. However,

Table 3. Lessons learned after developing an electronic SCD cohort from a data warehouse

Area of consideration	Lesson learned
Personnel of a multidisciplinary team required	1. Context expertise: a hematologist knowledgeable about the clinical care of SCD who can provide content expertise about SCD and medical complications. 2. Informatics expertise: an informatics expert knowledgeable about clinical information systems (electronic health records, laboratory systems, and clinical data warehouses), data standards, terminologies, and research informatics. 3. Programmer expertise: a computer programmer knowledgeable about extracting the data from the data warehouse and electronic health record. 4. Research coordinators: research personnel knowledgeable about SCD-related medical complications with the ability to identify distinct complications in the electronic health records. Preferably at least 2 individuals extract data from the electronic health records, so at least 10% of the data extracted can be double-checked.
Effort for the team needed	Initial creation and validation of cohort: 1. 200 h of context expertise 2. 2500 h of informatics expertise 3. 4000 h of programmer time 4. 5350 h of validation time (this can vary depending on cohort size and selected SCD phenotypes) An ongoing effort is needed to assess and update the cohort to double check data structures and integrity. This effort could be substantially more if moving to a new electronic health record or data warehouse system. The minimal amount of time required annually: 1. 40 h of programmer time 2. 40 h of validation time
Practical lessons learned	1. A recurring feedback loop is needed between members of the team and ongoing assessment of data quality and validation work. The feedback loop included the following: a. Weekly meetings between all members of the team to resolve and troubleshoot issues. b. Weekly programmer-specific meetings to develop algorithms and complete data extraction. 2. Anticipate data fragmentation requiring vigilance to ensure complete data capture.
Exportability	1. Every electronic health record data warehouse has its unique attributes, including data structures, terminologies, access, and security that the team needs to familiarize themselves with and use. 2. Changes in electronic health record data entry and clinical information systems may require modifications to code and extract data or additional validation to ensure data is correct and accurate. 3. Harmonizing data systems across electronic health records is essential. There are certain areas in which harmonization can occur (eg, laboratory values) and those in which harmonization can be difficult (eg, reports and clinical notes).

We have identified the most salient strategies for creating a contemporaneous cohort of children and adults with SCD from a data warehouse.

after the contemporaneous cohort has been established, the results can be updated annually at a fraction of the original cost. Because of the comprehensive and efficient approach of the automated cohort, we have already elected to use our SCD automated cohort as a contemporaneous comparison group for individuals with SCD who received curative therapy (5U01HL156620).

The automated and hybrid approach significantly differs in capturing mortality in the cohort. The automated cohort used our data warehouse, which obtains mortality data from our electronic health record and external sources, and found 18 more deaths than the hybrid approach during the same period. The lack of mortality data in the electronic health records was a major limitation in estimating the mortality rate. Otherwise, our hybrid cohort had similar patient characteristics and laboratory values as the automated cohort (supplemental Tables 2-4).

Several limitations occurred in our study to determine the mortality rate in the automated and hybrid cohorts. First, we used prescriptions of hydroxyurea as a marker for hydroxyurea therapy without confirming adherence to therapy. However, prescription history is a reasonable approach for disease-modifying treatment assignment and is similar to previous large cohorts showing hydroxyurea therapy to be associated with a survival benefit.[19,22,24,26] Second, we could not account for hydroxyurea prescriptions outside of VUMC. However, our clinical practice has not significantly changed during the sampling frame, and relatively few adults and fewer children have hydroxyurea prescribed outside of a VUMC provider. Despite these 2 limitations, the preponderance of evidence indicates that in long-term follow-up cohort studies, hydroxyurea therapy decreases mortality as compared with no disease-modifying treatment in adults with SCD. Another limitation in our analysis plan is the fragmentation of care, in which individuals have irregular follow-up visits. Because of the inability to adequately treat and observe the effects of treatment on our primary outcome of mortality, we excluded people with gaps of greater than 3 years. Finally, we could not evaluate the impact of the recent FDA-approved drugs, L-glutamine, crizanlizumab, and voxelotor on mortality. However, over the next decade, our automated cohort will be poised to evaluate the effects of these medications on mortality.

The unique strengths of our study include:

1. Developed for the first time, a contemporaneous cohort of children and adults with SCD who received sustained treatment with disease-modifying therapy from a data warehouse of electronic health records.
2. Unlike previous studies, our approach provided mortality estimates that considered the impact of disease-modifying therapy to estimate mortality. The importance of this strategy cannot be underscored.
3. For the first time, we provided the methodology required to create an SCD electronic cohort from a data warehouse of individuals with SCD.
4. We showed, for the first time, the internal resources required, including the personnel and the necessary time to develop a contemporaneous cohort of children and adults with SCD.

Together, these advantages provide a new opportunity to develop contemporaneous cohorts to evaluate new therapies.

In a large health care system of >3 million individuals, we showed an automated contemporaneous cohort of children and adults with SCA could be generated with a data warehouse to estimate mortality. Members of the SCD and HSCT VUMC teams will use the results of these studies to advise families on the relative benefits and risks of present FDA-approved disease-modifying therapies and emerging strategies to cure SCD in children and adults. For the SCD community, in this article, we also provided computer algorithms to replicate, expand, or create their health care systems contemporaneous SCD cohort to compare to novel curative therapies being evaluated.

Acknowledgments

This work was partly funded by National Institutes of Health (NIH) Agreement OT3HL147810 as part of the Cure Sickle Cell Initiative. The Cure Sickle Cell Initiative is a collaborative, patient-focused research effort designed to accelerate the advancement of genetic-based cures for SCD. The initiative is funded by the National Heart, Lung, and Blood Institute, NIH. This work was, in part, supported by the National Center for Advancing Translational Sciences, Clinical and Translational Science Awards (grant UL1 TR002243). In addition, this work was supported by National Heart, Lung, and Blood Institute, NIH grant K23HL141447 (R.M.C.).

The views and conclusions contained in this article are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. Its contents are solely the authors' responsibility and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the NIH.

Authorship

Contribution: M.R.D., M.R., K. Wuichet, and R.M.C. designed the study; K. Wuichet, D.L.G., B.H., M.C., J.H., X.N., M.R.D., and R.M.C. collected the data; M.R. performed the analyses; M.R.D., K. Wuichet, M.R., and R.M.C. interpreted the results; R.M.C. and K. Wuichet wrote the manuscript; and all authors reviewed and edited the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

ORCID profiles: R.M.C., 0000-0003-1916-6521; K. Wuichet, 0000-0002-5653-2597; M.R., 0000-0001-7258-0073; M.R.D., 0000-0002-0574-1604.

Correspondence: Michael R. DeBaun, Department of Pediatrics, Vanderbilt University Medical Center, 2525 West End Ave, Suite 750, Nashville, TN 37203; email: m.debaun@vumc.org.

References

1. US Food & Drug Administration. Long term follow-up after administration of human gene therapy products. Guidance for industry; 2020. Accessed 3 August 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/long-term-follow-after-administration-human-gene-therapy-products
2. Bremond-Gignac D, Lewandowski E, Copin H. Contribution of electronic medical records to the management of rare diseases. *Biomed Res Int*. 2015;2015:954283.
3. Colbaugh R, Glass K, Rudolf C; Mike Tremblay Volv Global Lausanne Switzerland. Learning to identify rare disease patients from electronic health records. *AMIA Annu Symp Proc*. 2018;2018:340-347.
4. Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. *Kidney Int*. 2020;97(4):676-686.
5. Michalik DE, Taylor BW, Panepinto JA. Identification and validation of a sickle cell disease cohort within electronic health records. *Acad Pediatr*. 2017;17(3):283-287.
6. Garcelon N, Neuraz A, Salomon R, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet J Rare Dis*. 2018;13(1):85.
7. Singh A, Mora J, Panepinto JA. Identification of patients with hemoglobin SS/Sβ0 thalassemia disease and pain crises within electronic health records. *Blood Adv*. 2018;2(11):1172-1179.
8. Snyder AB, Zhou M, Theodore R, Quarmyne MO, Eckman J, Lane PA. Improving an administrative case definition for longitudinal surveillance of sickle cell disease. *Public Health Rep*. 2019;134(3):274-281.
9. Datavant. Mortality data in healthcare analytics; 2022. Accessed 25 February 2023. https://datavant.com/resources/whitepapers/mortality-data-in-healthcare-analytics/
10. Hashmi NK, Moiz B, Nusrat M, Hashmi MR. Chromatographic analysis of Hb S for the diagnosis of various sickle cell disorders in Pakistan. *Ann Hematol*. 2008;87(8):639-645.
11. National Institutes of Health; National Heart, Lung, and Blood Institute. The management of sickle cell disease; 2002. Accessed 25 February 2023. https://www.nhlbi.nih.gov/resources/management-sickle-cell-disease
12. DeBaun MR, Gordon M, McKinstry RC, et al. Controlled trial of transfusions for silent cerebral infarcts in sickle cell anemia. *N Engl J Med*. 2014;371(8):699-710.
13. Kelly S, Rodeghier M, DeBaun MR. Automated exchange compared to manual and simple blood transfusion attenuates rise in ferritin level after 1 year of regular blood transfusion therapy in chronically transfused children with sickle cell disease. *Transfusion*. 2020;60(11):2508-2516.
14. Walters MC, Patience M, Leisenring W, et al. Bone marrow transplantation for sickle cell disease. *N Engl J Med*. 1996;335(6):369-376.
15. Hulihan MM, Feuchtbaum L, Jordan L, et al. State-based surveillance for selected hemoglobinopathies. *Genet Med*. 2015;17(2):125-130.

16. Michlitsch J, Azimi M, Hoppe C, et al. Newborn screening for hemoglobinopathies in California. *Pediatr Blood Cancer*. 2009;52(4):486-490.
17. DeBaun MR, Ghafuri DL, Rodeghier M, et al. Decreased median survival of adults with sickle cell disease after adjusting for left truncation bias: a pooled analysis. *Blood*. 2019;133(6):615-617.
18. Hamideh D, Alvarez O. Sickle cell disease related mortality in the United States (1999–2009). *Pediatr Blood Cancer*. 2013;60(9):1482-1486.
19. Lê PQ, Gulbis B, Dedeken L, et al. Survival among children and adults with sickle cell disease in Belgium: benefit from hydroxyurea treatment. *Pediatr Blood Cancer*. 2015;62(11):1956-1961.
20. Nouraie M, Darbari DS, Rana S, et al. Tricuspid regurgitation velocity and other biomarkers of mortality in children, adolescents and young adults with sickle cell disease in the United States: the PUSH study. *Am J Hematol*. 2020;95(7):766-774.
21. Telfer P, Coen P, Chakravorty S, et al. Clinical outcomes in children with sickle cell disease living in England: a neonatal cohort in East London. *Haematologica*. 2007;92(7):905-912.
22. Steinberg MH, McCarthy WF, Castro O, et al. The risks and benefits of long-term use of hydroxyurea in sickle cell anemia: a 17.5 year follow-up. *Am J Hematol*. 2010;85(6):403-408.
23. Steinberg MH, Barton F, Castro O, et al. Effect of hydroxyurea on mortality and morbidity in adult sickle cell anemia: risks and benefits up to 9 years of treatment. *JAMA*. 2003;289(13):1645-1651.
24. Voskaridou E, Christoulas D, Bilalis A, et al. The effect of prolonged administration of hydroxyurea on morbidity and mortality in adult patients with sickle cell syndromes: results of a 17-year, single-center trial (LaSHS). *Blood*. 2010;115(12):2354-2363.
25. Araujo OMRd, Ivo ML, Ferreira MA, Pontes ERJC, Bispo IMGP, Oliveira ECLd. Survival and mortality among users and non-users of hydroxyurea with sickle cell disease. *Rev Lat Am Enfermagem*. 2015;23(1):67-73.
26. Fitzhugh CD, Hsieh MM, Allen D, et al. Hydroxyurea-increased fetal hemoglobin is associated with less organ damage and longer survival in adults with sickle cell anemia. *PLoS One*. 2015;10(11):e0141706.
27. Desai RJ, Mahesri M, Globe D, et al. Clinical outcomes and healthcare utilization in patients with sickle cell disease: a nationwide cohort study of Medicaid beneficiaries. *Ann Hematol*. 2020;99(11):2497-2505.
28. Aljuburi G, Laverty AA, Green SA, Phekoo KJ, Bell D, Majeed A. Socio-economic deprivation and risk of emergency readmission and inpatient mortality in people with sickle cell disease in England: observational study. *J Public Health (Oxf)*. 2013;35(4):510-517.
29. Reeves S, Garcia E, Kleyn M, et al. Identifying sickle cell disease cases using administrative claims. *Acad Pediatr*. 2014;14(5):S61-S67.
30. Reeves SL, Madden B, Wu M, et al. Performance of ICD-10-CM diagnosis codes for identifying children with sickle cell anemia. *Health Serv Res*. 2020;55(2):310-317.
31. Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM. The accuracy of hospital ICD-9-CM codes for determining sickle cell disease genotype. *J Rare Dis Res Treat*. 2017;2(4):39.
32. Grosse SD, Green NS, Reeves SL. Administrative data identify sickle cell disease: a critical review of approaches in US health services research. *Pediatr Blood Cancer*. 2020;67(12):e28703.