GRNDaD: Big Data and Sickle Cell Disease

Sophie Lanzkron, MD, MHS\textsuperscript{1}, Deepa Manwani, MD\textsuperscript{2}, Payal Desai, MD\textsuperscript{3}, Julie Kanter, MD\textsuperscript{4} Jane Little, MD\textsuperscript{5}

\textsuperscript{1}Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; \textsuperscript{2}Department of Pediatrics, Albert Einstein College of Medicine, Bronx, NY; \textsuperscript{3}Department of Medicine, Ohio State University, Columbus, OH \textsuperscript{4}Department of Medicine, University of Alabama at Birmingham, Birmingham, AL \textsuperscript{5}Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, NC,

Running title: Big data SCD

Word count: 1821

Corresponding Author:
Sophie Lanzkron, MD, MHS
Johns Hopkins University School of Medicine
1830 E. Monument St. Suite 7300
Baltimore, MD 21205
slanzkr@jhmi.edu
410-502-8642
Abstract

Big data is being used in the pursuit of precision medicine in the general population. Applying these tools to people with sickle cell disease (SCD), is essential for ensuring that patients receive the most appropriate, customized therapy for their disease. In order to apply these tools there must be large numbers of willing, fully phenotyped participants in collaborative registries. The Globin Regional Data and Discovery (GRNDaD) registry was developed by SCD providers at multiple sites, working collaboratively, to respond to unmet clinical needs and the lack of a large multisite registry. The specific goals of GRNDaD are to 1. prospectively obtain high quality curated data on the evolving sickle cell population; 2. Improve adherence to guideline-based care; 3. Provide a platform for ongoing quality improvement across sites; 4. Allow real-time investigation of therapies and to collaborate broadly to address research questions using GRNDaD as a shared platform. GRNDaD’s current strength lies in the generous participation of people living with SCD, collaborative investigators, and the opportunity to conduct quality improvement activities across a large number of sites. GRNDaD will serve as the data collection tool for the Health Resources and Services Administration Sickle Cell Treatment Demonstration Program and for the newly established National Alliance for Sickle Cell Centers (https://www.sicklecellcenters.org /). GRNDaD is a robust collaborative registry that will be used to identify genetic markers that will help predict outcomes and lead to a better understanding of the natural history of sickle cell disease in the modern era of novel therapies.
There is little consensus on how to define “big data”. Gandomi et al¹ describe the three V’s; volume, variety and velocity as a framework to describe big data. Volume is the magnitude of data, variety is the structural heterogeneity in a dataset and velocity is the rate at which data are generated and analyzed. In applying these concepts to medicine there has been a focus on using genomics, proteomics, metabolomics along with data from the electronic health records, imaging and even social media to enable big data in the pursuit of precision medicine.² Examples of precision medicine include genetic markers that predict reduced efficacy or toxicity of therapies such as variants in CYP2C9 and warfarin or lack of TPMT in the use of mercaptopurine.³ The development of computational infrastructure to store and compute complex data has made the ability to expand the use of precision medicine beyond genomics and many such studies are ongoing in the general population. Applying these tools to rare diseases, like sickle cell disease (SCD), is essential for ensuring that patients receive the most appropriate, customized therapy for their disease. In order to apply these tools there must be large numbers of willing, fully phenotyped participants in collaborative registries.

In sickle cell disease, registries have been profoundly helpful, but have suffered from institutional attention deficit, lacking a sustained, widespread, and unified approach; registry support in SCD has also been economically disadvantaged compared to other chronic diseases, like cystic fibrosis (very deep foundation support) and Hemophilia (longstanding stable support from the Centers for Disease Control). This matters in the modern world for two reasons. First, because of dramatic improvements in lifespan, associated with disease modification in children, previous multi-centered observational studies in SCD do not accurately describe the modern population of people with the disease. Prior studies were cross-sectional without longitudinal assessments and done prior to the availability of a validated quality of life tool in SCD⁴ or they provided detailed data on
specific aspects of disease management such as transfusion but lacked many of the clinical characteristics to provide a full clinical phenotype.\textsuperscript{5} Neither the modern patient with sickle cell disease, nor their family or providers, have complete information about how SCD will affect them across the their lifespan. Second, the chronic lack of widespread comprehensive care in SCD, including detailed prospective registries, has left this population unprepared for curative-intent therapy. The seeds of cure are being cast into a garden that is unprepared, lacking context from longstanding prospective registries through which to understand what treatments are most beneficial. The largest observational trial of SCD was the Cooperative Study of Sickle Cell Disease (CSSCD), which ceased follow-up in the 1980’s, when the average life expectancy for a patient with sickle cell anemia was less than 20 years of age.\textsuperscript{6} Since then, there have been no large sustained multi-site prospective cohort studies that capture the shifting phenotype of the modern population of people living with SCD. We therefore, lack information about how current therapies are being used in practice, the long-term effect of these therapies, the natural history of sickle cell disease and we lack the tools needed to identify high-risk patients who are most likely to benefit from potentially high-risk disease-modifying treatments.

The complexities of using and analyzing big data change when trying to apply the same tools to rare diseases, like sickle cell disease. There are no published validated panels of biomarkers for clinical manifestations of SCD.\textsuperscript{7} Many of the machine learning statistic-based tools used to analyze big data require a large amount of data from a sufficient number of subjects.\textsuperscript{8} These large sample sets are often not available for people with rare diseases and what is available often lacks robust information on clinical phenotypes associated with the patient data and biospecimens. A recent publication by Liggett et al used the NHLBI Trans-Omics for Precision Medicine (TOPMed) consortium to examine the presence of clonal hematopoiesis in 3090
individuals with SCD compiled from 30 distinct cohorts. The cohorts used were historical and the population was relatively young with the majority of the samples being from subjects under the age of 30 years. In addition, although data on therapies were reported, valuable information necessary to assess risk profiles, such as years of exposure to hydroxyurea and the use of chronic transfusion therapy, was not. In order to take advantage of new technologies being used to evaluate big data, establishment of large research consortia with well-described populations of people living with SCD and expertise in all multiomic platforms is crucial.

The Globin Regional Data and Discovery (GRNDaD) Registry (https://grndad.org) was developed by SCD providers at multiple sites, working collaboratively, to respond to unmet clinical needs and inadequate national infrastructure. GRNDaD collects detailed patient-level data through which to document a robust clinical phenotype. GRNDaD was initially funded by the Doris Duke Foundation in 2015, with the goal of building a shared database that could be used to define the phenotype of the contemporary patient who is living with SCD and to examine the evolving natural history of SCD. Funding allowed investigators to go through a lengthy and iterative process to develop data collection forms for children and adults living with SCD, which included variables and patient reported outcome (PRO) tools that were important for clinical care, quality improvement, and research. The pilot funding allowed GRNDaD to integrate processes for ethical user-friendly regulatory alignment (a single IRB), inclusive and appropriate data use agreements, and to develop a HIPAA-compliant secure web accessible REDCap data repository. Importantly, a framework evolved in which coordinators at each site worked collaboratively to maintain and improve these processes. Currently there are 12 sites contributing data to GRNDaD and over 1500 subjects have been consented, with an average age of 33 years with almost 25% over the age of 39. The specific goals of GRNDaD are to: 1. prospectively
obtain high quality curated data on the evolving sickle cell population (comorbidities, labs, imaging, PROs). 2. Improve adherence to National Heart Lung Blood Institute (NHLBI), American Society of Hematology (ASH), and GOT Transition guideline-based care. 3. Provide a platform for ongoing (investigator- and patient-driven) quality improvement across sites. 4. Allow real-time investigation of therapies, established and novel, and finally to collaborate broadly to address research questions using GRNDaD as a shared platform. All investigators who contribute data have access to all of GRNDaD data by completing a study proposal document which is reviewed by the executive committee who assesses if the study is feasible and not in conflict with another project. Once the review is complete the data is sent to the investigator for analysis. The next steps for GRNDaD entail working with collaborators across the world (Africa, Europe, South America) to collect data harmoniously, using a shared codebook with the same variables. This goal will allow for a more robust sample size and will further lend itself to big data analysis. We are not yet routinely collecting samples for biobanking but our consent does allow for this when resources are available to collect and store samples in a consistent and reliable manner.

GRNDaD is onboarding new sites and continues to grow. However, even at our current size, we have made some important observations. Several studies have examined adherence to guideline-based recommendations on screening. We have examined our real-world screening and management practices of hepatic iron overload\(^\text{10}\) and renal disease\(^\text{11}\) in people with SCD. We have found that screening for albuminuria occurred in only 37% of annual observations; similarly, more than a third of patients with a ferritin >1500 mg were not on chelation, and only a quarter had had imaging studies to assess iron accumulation. We also found that albuminuria was associated with a clinical phenotype. Hemoglobin levels were, on average, 0.79 g/dL lower in
those with albuminuria between 30-300 mg/g when compared to those with no albuminuria (P=0.005).\textsuperscript{11}

As we continue to build the capacity of GRNDaD to contribute to natural history studies and multiomic analyses, we are also using GRNDaD to improve care, by both focusing on ways to improve screening and providing tools to improve adherence to guideline recommendations.

For instance, SCD specialty care is not readily available to many people, especially adults, who are living with SCD\textsuperscript{12}. Thus, systems-level reminders are needed to ensure individuals receive guideline-based care. In GRNDaD, we have developed a patient-specific health maintenance sheet that is populated when prompted e.g., prior to a subject’s routine clinic visit. This document summarizes data extracted from GRNDaD on whether subjects have had: guideline recommended laboratory screening (and the results of this testing), annual ophthalmology exams, annual transcranial Doppler testing and recommended vaccinations. The health maintenance sheet case report form also highlights when patients with SCD are not on disease modifying therapy. This tool was developed to guide all providers, regardless of the size of their clinical population, and will add value as GRNDaD expands to other sites. The expectation is that these prompts will facilitate adherence to guideline-based medicine, and increase the likelihood of improved outcomes for patients, e.g. adherence to recommendations for transcranial Doppler testing and use of hydroxyurea. While the 12 active sites currently enrolling subjects are all affiliated with large academic institutions, we are open to expanding to small less-resourced programs. GRNDaD investigators have IRB approval to do virtual consents and so the ability to enroll subjects that live far from academic centers will allow for a more diverse group of enrolled subjects.
There are several other multisite registries including the NHLBI sponsored Sickle Cell Disease Implementation Consortium which has enrolled 2400 subjects who continue to have longitudinal follow-up but no new subjects are actively being enrolled. This registry collects data that is similar to GRNDaD, including a number of PRO’s. Work is ongoing to see how data from both the SCDIC and GRNDaD can be combined on an NIH platform. Another ongoing registry is the American Society of Hematology’s Research Collaborative. This registry is collecting, mostly limited datasets with data extracted directly from the electronic health record using ICD-10 and Systematized Nomenclature of Medicine (SNOWMED) codes from sites that are participating as hubs in the American Society of Hematology Sickle Cell Clinical Trial Network. The utility of this registry, given the limitations of retrieving accurate genotype data using ICD-10 coding remains unclear.

GRNDaD’s current strengths lie in the generous participation of people living with SCD, collaborative investigators, and the unique opportunity to conduct quality improvement activities across a large number of sites and in associating clinical phenotypes with outcomes. GRNDaD will serve as the data collection tool for the Health Resources and Services Administration Sickle Cell Treatment Demonstration Program and for the newly established National Alliance for Sickle Cell Centers (https://www.sicklecellcenters.org/). In the near future, GRNDaD will be used to identify both bio and genetic markers that will help predict outcomes and will be a robust collaborative registry that will be used to understand the natural history of sickle cell disease as the CSSSCD did decades ago.
Author contribution
S.L. wrote the paper and assisted with designed registry and D.M., J.L., J.K., and P.D. assisted with design of registry and manuscript review.

Conflict-of-interest statement
S.L. research funding Shire, Novartis, Imara, GBT, Consulting-Bluebird Bio, Novo Nordisk, Stocks-Pfizer, Teva DM-Consulting Bluebird bio, GBT, Novartis, Pfizer, Forma PD-Grants/Research Support-GBT, CSL, Shire/Baxalta, University of Pittsburgh, University of Tennessee, NASAC Speaker-Novartis Advisory Board-Forma Honorarium-ASH J.K.-Consultant Novartis, Forma, Graphite Bio,GLG Advisory board-Novartis, Forma, Agios, Beam, Sanofi Honorarium-Novartis, Forma, Agios, Beam, Sanofi, Guidepoint Global J.L. research funding-Bluebird Bio.
References

1. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*. 2015/04/01/ 2015;35(2):137-144. doi:https://doi.org/10.1016/j.ijinfomgt.2014.10.007

2. Vogt H, Green S, Ekstrøm CT, Brodersen J. How precision medicine and screening with big data could increase overdiagnosis. *Bmj*. Sep 13 2019;366:l5270. doi:10.1136/bmj.l5270

3. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. *Health Aff (Millwood)*. 2018;37(5):694-701. doi:10.1377/hlthaff.2017.1624

4. Machado RF, Barst RJ, Yovetich NA, et al. Hospitalization for pain in patients with sickle cell disease treated with sildenafil for elevated TRV and low exercise capacity. *Blood*. Jul 28 2011;118(4):855-64. doi:10.1182/blood-2010-09-306167

5. Karafin MS, Bruhn R, Westlake M, et al. Demographic and epidemiologic characterization of transfusion recipients from four US regions: evidence from the REDS-III recipient database. *Transfusion*. Dec 2017;57(12):2903-2913. doi:10.1111/trf.14370

6. Platt OS, Brambilla DJ, Rosse WF, et al. Mortality in Sickle Cell Disease -- Life Expectancy and Risk Factors for Early Death. *The New England Journal of Medicine*. 6/9/1994 1994;330(23):1639-1644. Not in File.

7. Goodman SR, Pace BS, Hansen KC, et al. Minireview: Multiomic candidate biomarkers for clinical manifestations of sickle cell severity: Early steps to precision medicine. *Exp Biol Med (Maywood)*. Apr 2016;241(7):772-81. doi:10.1177/1535370216640150

8. Brasil S, Neves CJ, Rijoff T, et al. Artificial Intelligence in Epigenetic Studies: Shedding Light on Rare Diseases. *Front Mol Biosci*. 2021;8:648012. doi:10.3389/fmolb.2021.648012

9. Liggett LA, Cato LD, Weinstock JS, et al. Clonal hematopoiesis in sickle cell disease. *medRxiv*. 2021:2021.06.12.21258772. doi:10.10110/2021.06.12.21258772

10. Sears M, Lanzkron S, Hoppe C, et al. Iron Overload Is Under-Recognized and Under-Treated in SCD: A Report from the Grndad Registry. *Blood*. 2018;132(Supplement 1):158-158. doi:10.1182/blood-2018-158

11. Williams E, Brown E, Manwani D, et al. Chronic Kidney Disease Is Under-Screened in SCD and Mild Albuminuria Is Associated with a Drop in Hemoglobin: A Report from the Grndad Sickle Cell Registry. *Blood*. 2019;134(Supplement_1):2284-2284. doi:10.1182/blood-2019-124933

12. Grosse SD, Schechter MS, Kulkarni R, Lloyd-Puryear MA, Strickland B, Trevathan E. Models of comprehensive multidisciplinary care for individuals in the United States with genetic disorders. *Pediatrics*. 1/2009 2009;123(1):407-412. Not in File. doi:123/1/407 [pii];10.1542/peds.2007-2875 [doi]

13. Eisenbrown K, Nimmer M, Brousseau DC. The accuracy of using ICD-9-CM codes to determine genotype and fever status of patients with sickle cell disease. *Pediatr Blood Cancer*. May 2015;62(5):924-5. doi:10.1002/pbc.25432