SAS macro code to assemble fingernail photo image (RGB) text files generated by a phone app in pre-term infants at bedside in Neonatal ICU (NICU): Processing large volume of text files to build accurate and clean analytical datasets for image analysis.

Neeta V Shenvi M.S.[#], Kirk A Easley M.A.Stat [#], Ravi M Patel M.D. [$], Cassandra Josepheson, M.D.[!], and Amita Manatunga Ph.D[#].

Department of Biostatistics and Bioinformatics[#], Rollins School of Public Health, School of Medicine[$], Emory University, Atlanta, Georgia 300322.

Cancer and Blood Disorders Institute[!], Johns Hopkins All Children's Hospital,St. Petersburg, FL

ABSTRACT

This project presents the development and application of SAS macro code to process and integrate a large volume of RGB text files derived from fingernail and palm photos of preterm infants in the NICU bedside.

Study nurse takes infant nail pictures in NICU bedside using a smart phone, identify a region of interest (a rectangle) on the picture. The phone app converts pixel intensities into numerical values for color (red, green, blue) and creates a text file. The text file from phone app is then uploaded to data coordinating center servers for further processing.

The phone app generated ~1700 txt files from 89 infants (weekly longitudinal nail pictures for first 13 weeks of life during their stay in the NICU).

We describe SAS 9.4 code to manage and merge these files to create analytic datasets. These datasets are further merged with infant longitudinal characteristics variables that are saved in a separate database.

Our long term goal is to apply various image analysis algorithms on these analytic datasets to devise noninvasive methods to predict hemoglobin and detect anemia.

Clinical and Research Background

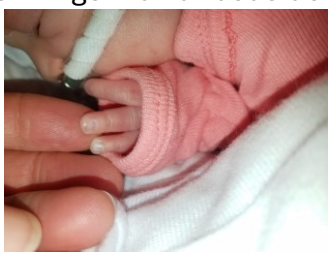
Anemia is a common and serious condition among preterm infants, often necessitating multiple blood transfusions during NICU stays. Traditional hemoglobin measurement requires repeated invasive blood draws, which are burdensome and can exacerbate anemia. This project, conducted at Emory University in collaboration with Johns Hopkins All Children's Hospital, explores a novel approach: leveraging mobile phone–based nailbed photography and colorimetric analysis as a surrogate for hemoglobin estimation. Using the AnemoCheck®Mobile app developed by Dr. Mannino, research staff captured weekly photographs of infants' finger nails, toenails, and palms during their first 13 weeks of life. Each photo was processed by the app into pixel intensity values (RGB), generating pipe-delimited text files that formed the foundation of the analytic workflow.

Data Collection and Scope

Ninety preterm infants with birth weights below 1250 grams were enrolled from three hospitals across Georgia. Over the course of the study, the NICU staff captured approximately 1,700 high-resolution image text files for first 13 weeks of infants' life during their stay in the NICU bedside (Figure 1 and 2). Alongside image data, critical longitudinal variables were recorded, including hemoglobin values, complete blood counts, metabolic panels, transfusion history, and other clinical outcomes. This multimodal dataset offered a rich source for studying potential correlations between nailbed coloration and anemia.

Figure 1: Infant Finger nails

Infant left finger nail at bedside in NICU



Infant right finger nail at bedside in NICU

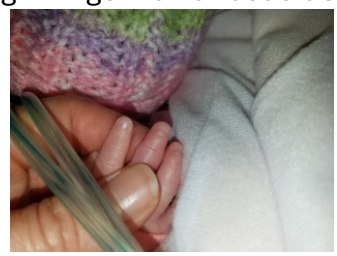
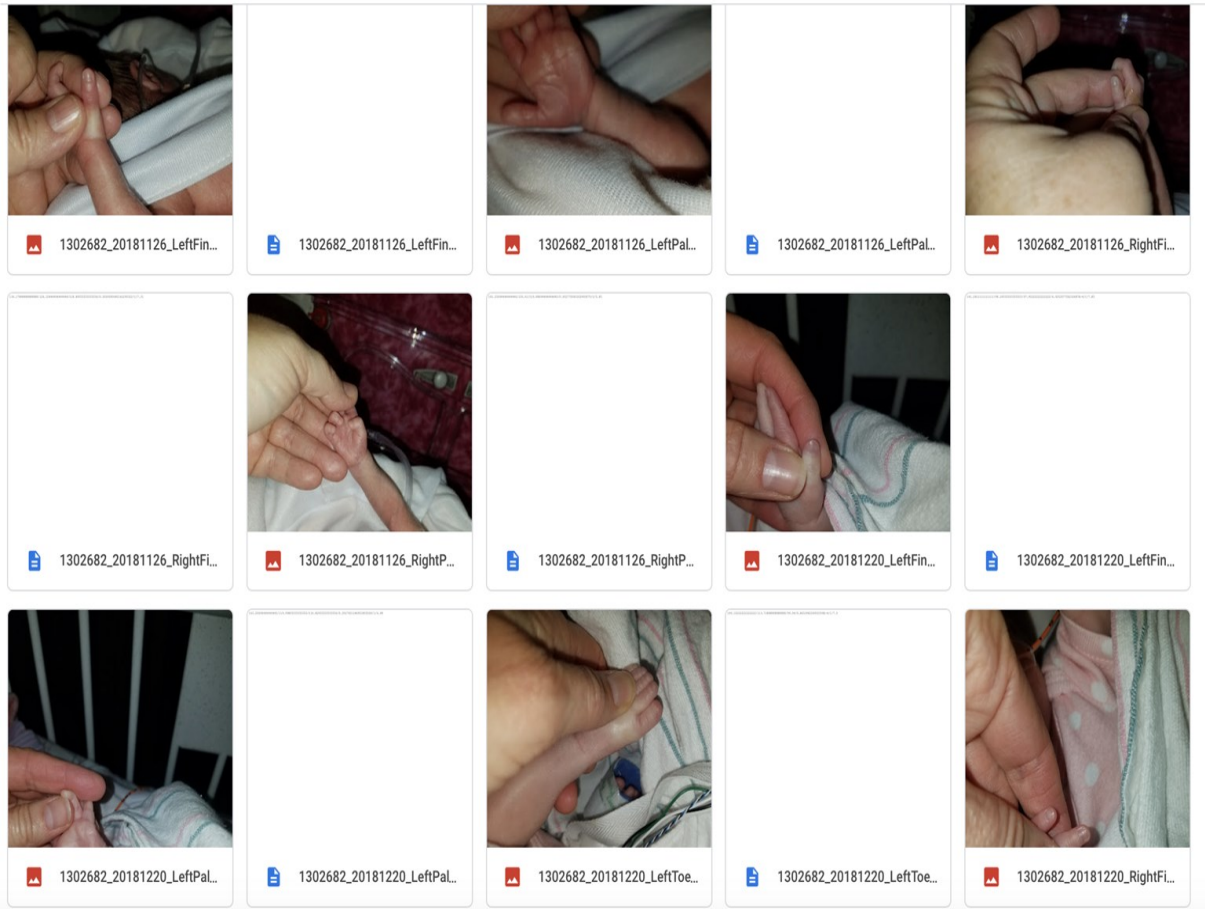


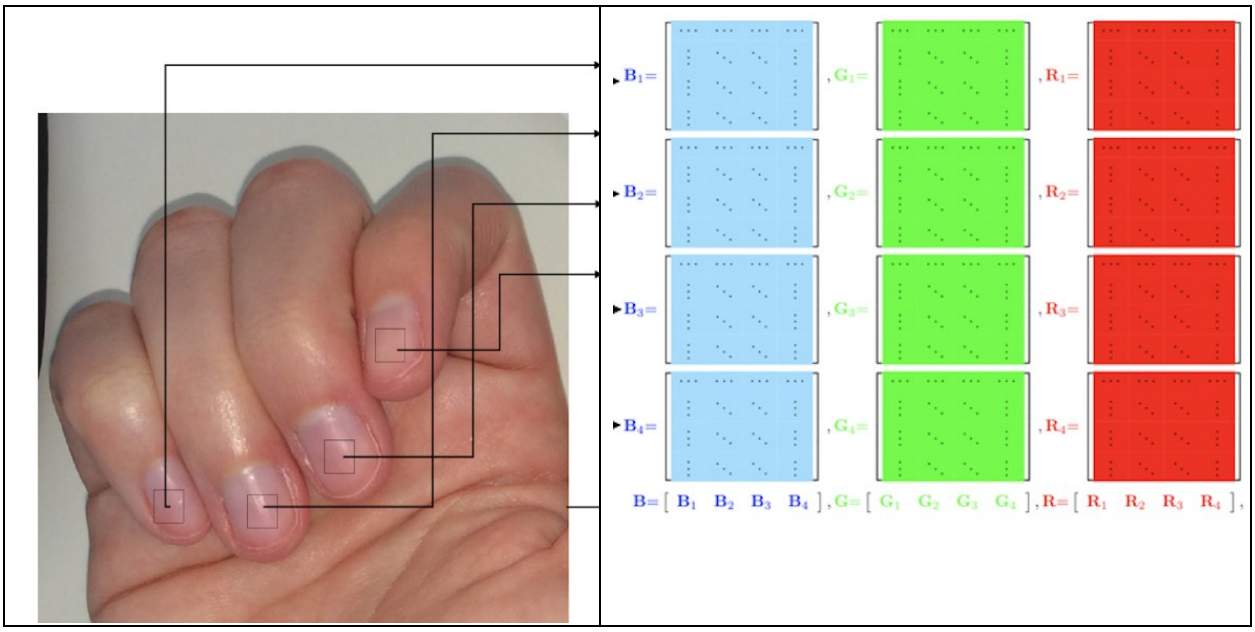
Figure 2: Infant Finger nails and palm at bedside in NICU.



<u>Nail photo pixel data processing using app</u>:

This study utilizes a phone app created by Dr. Mannino: AnemoCheck®Mobile Mannino et al. (2018)**.** The NICU research staff take infants' nail pictures from six different anatomic regions (left and right fingernail, toe, palm) at NICU bedside. The NICU research staff then use the phone app to identify a region of interest (a rectangle) on the photo (Figure 3). The app converts pixel intensities into numerical values for color (**RGB**) and creates a pipe separated text file. This text file is then uploaded to study servers for further processing. The phone app generated ~1700 text files from 90 infants from first 13 weeks of life during their stay in the NICU).

Figure 3: The phone app is used to detect pixel intensity of the specified region.



Processing Workflow with SAS

The project required the creation of an automated, efficient system to handle the large number of files and transform them into usable datasets for statistical and image analyses. The SAS 9.4 macro code was designed to:

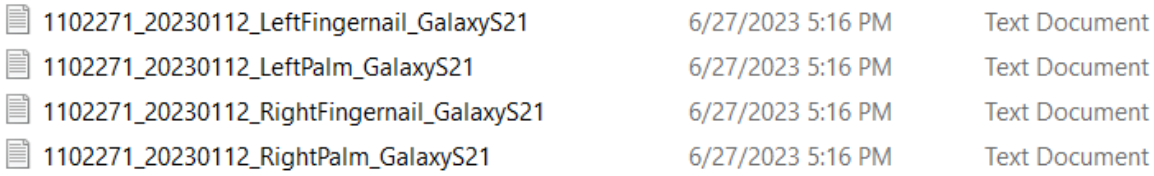
1. **File Import and Management** – Using the filename statement with the PIPE option, the macro first generated a dataset containing all text file names from the study directory. Each filename encoded three key identifiers: infant ID, date of image capture, and anatomical location. SAS string functions parsed these identifiers into structured variables, enabling systematic linkage across datasets.

2. **Sequential Data Extraction** – A looping macro structure read each RGB text file, imported the pixel values, and appended them into a master dataset. Each file contained six variables: red, green, blue pixel intensities, camera exposure, and brightness. These were systematically extracted and aligned with the identifiers dataset.

3. **Dataset Merging and Quality Control** – PROC SQL merged the RGB dataset with the identifiers dataset, creating a unified SAS image dataset. Automated checks flagged missing data and discrepancies, ensuring integrity of the analytic dataset.

4. **Integration with Clinical Data** – The final step merged the image dataset with the infant longitudinal database that housed hemoglobin and other clinical measures. This integration allowed for correlation analyses between pixel-level features and clinical outcomes.

We describe SAS 9.4 macro code that uses external text files to create the desired analytical datasets. Each picture file is given a unique filename. The file name has a composite string made up of 3 identifiers: <<Infant Id>>_<<date picture taken>>_<<body region location>>. Below screenshot shows filenames.



The SAS macro first imports filenames of all the text files as string variable and creates a dataset. The filename string variable is then split into several string variables using SAS string functions. The next step in the macro is to read RGB data from each file sequentially and creates numeric RGB color variables. The filename dataset and the RGB datasets are then merged using PROC SQL to create SAS image dataset. The macro code also outputs quality control checks such as data discrepancy and data missingness. The final analytical dataset is created using SAS PROC SQL code to merge image dataset with infant longitudinal hemoglobin and other characteristics that are saved in the external database.

SAS macro code:

PROCESS FLOW: For the process flow details please refer to the macro code attached as an appendix.

STEP 1: Save RGB TXT file names as a SAS dataset list.

Using filename statement with the PIPE option we create a fileref which points to the location where all txt files are saved and then with an infile statement a SAS dataset is created with all txt file name.

```
filename nLIST1 pipe 'dir "C:\SESUG2025\1*.txt" /b' ;
```

SAS PIPE option is used to read the name of all the files that start with 1 and end as .txt files at the location specified by the study path.

```
data nLIST ;
  infile nLIST lrecl=200 truncover;
```

```
  input file_name $100.;
run;
```

Below is partial list of filename SAS dataset.

file_name
1202202_20240916_RightFingernail_GalaxyS21.txt
1102611_20230731_RightToenail_GalaxyS21.txt
1201891_20231113_LeftPalm_GalaxyS21.txt
1305681_20210414_LeftFingernail.txt
1201471_20230102_LeftFingernail_GalaxyS21.txt
1303211_20190408_LeftFingernail.txt
1306961_20220112_LeftFingernail.txt
1306121_20210804_LeftFingernail.txt
1302682_20181220_LeftToenail.txt
1305142_20200915_RightFingernail.txt
1303671_20190716_RightFingernail.txt
1304461_20191210_LeftPalm.txt
1305041_20201019_LeftPalm.txt
1305351_20210113_RightFingernail.txt
1302681_20181205_LeftPalm.txt
1201521_20230123_RightPalm_GalaxyS21.txt

STEP 2: Use data step to extract 3 identifiers from file name variable.

The file name has a composite string made up of 3 identifiers: <<Infant Id>>_<<date picture taken>>_<<body region location>>. We create these 3 identifiers using SAS string functions.

```
data nlist;
set nlist;
id=input(substr(file_name,1,7),7.0);  **** Infant id;
pict_dt=input(substr(file_name,9,17),8.0);  **** picture date;
dset=scan(file_name,1,'.');  ***** filename without .txt;
pic_location=substr(dset,18);  ***** body region of picture;

run;
```

Below is partial file name dataset with identifiers.

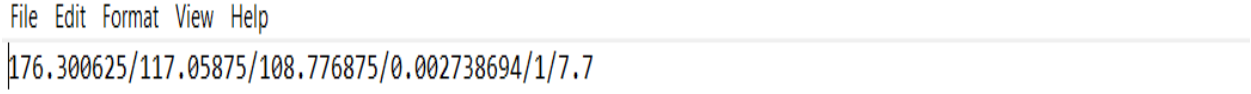
file_name	id	pict_dt	pic_location
1202202_20240916_RightFingernail_GalaxyS21.txt	1202202	20240916	RightFingernail_GalaxyS21
1102611_20230731_RightToenail_GalaxyS21.txt	1102611	20230731	RightToenail_GalaxyS21
1201891_20231113_LeftPalm_GalaxyS21.txt	1201891	20231113	LeftPalm_GalaxyS21
1305681_20210414_LeftFingernail.txt	1305681	20210414	LeftFingernail
1201471_20230102_LeftFingernail_GalaxyS21.txt	1201471	20230102	LeftFingernail_GalaxyS21
1303211_20190408_LeftFingernail.txt	1303211	20190408	LeftFingernail
1306961_20220112_LeftFingernail.txt	1306961	20220112	LeftFingernail
1306121_20210804_LeftFingernail.txt	1306121	20210804	LeftFingernail

STEP 3: Add a sequence number to the filename dataset.

```
***** Add sequence number;
data nlist;
set nlist;
count=_n_;
run;
```

STEP 4: Macro to sequentially read content of each txt file.

STEP 4a: Each txt file has 6 RGB variables that are separated by a pipe character. The content of the file is shown below.



The first 3 variables correspond to numeric values for Red ,Green, and Blue intensities of color respectively. The variable 4 is camera exposure time, variable 5 is camera xXXX and variable 6 is camera brightness value.

STEP 4b:The macro code sequentially reads each txt file from the filename dataset, extracts the content and makes the RGB dataset.

```
options mprint symbolgen;
%macro nlistreadin (_i=,_max=,_path=);

  %do i=&_i. %to &_max.;
*****************************;
  proc sql;
      select file_name into: _x from nlist where count=&i.;

  quit;

  PROC IMPORT OUT= WORK.t_&i.
        DATAFILE= "&_path.\&_x."
        DBMS=DLM REPLACE;
    DELIMITER='2F'x;
    GETNAMES=NO;
    DATAROW=1;
RUN;

*********************************;

data t_&i.;
  set t_&i.;

  seq=&i.;
  run;
***********************************;
proc append base=_master_nlist data= t_&i. force;run;
***********************************;
  proc sql;
drop table t_&i.;
  quit;

  %end;
%mend nlistreadin;
%nlistreadin(_i=1,_max=100,_path=C:\SESUU20205);
```

Below is the partial RGB SAS dataset (_master_nlist.sas7bdat).

VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	seq
33.47	42.4225	49.168125	0.019982648	1	2.7	1
132.935	83.949375	83.404375	0.001538428	1	8.54	2
137.520625	89.523125	80.44140625	0.001641718	1	8.44	3
110.7377778	68.80444444	63.37666667	0.001347709	1	6.89	4
195.4125	156.278125	153.80625	0.002946407	1	7.6	5
131.8566667	116.5155556	108.9933333	0.000706215	1	7.82	6
215.5644444	201.1155556	195.5644444	0.000710227	1	7.82	7
169.26	118.6844444	114.1511111	0.000464684	1	8.43	8

The filename dataset and the RGB dataset are then merged using PROC SQL to create SAS image dataset

The _master_nlist.sas7bdat needs to be merged with filename dataset that has infant identifiers.

STEP 5: Merge _master_nlist.sas7bdat with infant identifiers using PROC SQL.

```
PROC SQL;
CREATE TABLE _master_image AS
SELECT A.* ,B.*
FROM _master_nlist AS a
LEFT JOIN nlist AS b ON a.seq=b.count
;
QUIT;
```

The _master_image.sas7bdat is created that has 100 rows and 13 columns.

The partial image dataset is shown below. This dataset has key subject identifiers (subject id and picture date) along with corresponding red, green and blue values.

id	pict_dt	pic_location	R	g	b	exposure	brightness
1202202	20240916	RightFingernail_GalaxyS21	33.47	42.423	49.168	0.019983	2.7
1102611	20230731	RightToenail_GalaxyS21	132.935	83.949	83.404	0.001538	8.54
1201891	20231113	LeftPalm_GalaxyS21	137.521	89.523	80.441	0.001642	8.44
1305681	20210414	LeftFingernail	110.738	68.804	63.377	0.001348	6.89

The final analytical dataset is created using SAS PROC SQL code to merge image dataset with infant longitudinal hemoglobin and other characteristics that are saved in the external database.

Analytical Potential and Long-Term Goals

The resulting dataset provides a robust foundation for testing advanced statistical and machine learning algorithms aimed at predicting hemoglobin concentration noninvasively. By harnessing pixel-level image data and correlating it with laboratory hemoglobin values, the study aspires to validate nailbed imaging as a clinical decision-support tool. The long-term vision is to reduce invasive procedures, lower transfusion dependence, and ultimately improve outcomes for preterm infants.

Broader Implications

This work also highlights the importance of data harmonization in biomedical research. Integrating unstructured data (images) with structured clinical records requires meticulous data engineering. The SAS macro framework developed here can be adapted to other studies involving high-throughput image-to-text pipelines, serving as a model for future neonatal and broader healthcare research.

Conclusion

The project demonstrates how a thoughtful combination of mobile imaging technology, structured data processing, and statistical programming can open new frontiers in neonatal care. The SAS macro workflow not only streamlines the transformation of raw RGB text files into analytical datasets but also bridges the gap between digital health tools and clinical application. As this research advances, it has the potential to redefine anemia detection in preterm infants, offering a noninvasive, scalable, and data-driven solution to an urgent medical challenge.

References:

Mannino, R. G. et al (2018). Smartphone app for non-invasive detection of anemia using only patient-sourced photos. Nature Communications, 9(1).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Neeta Shenvi
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
nshenvi@emory.edu