NAU
NORTHERN ARIZONA UNIVERSITY
College of the Environment, Forestry, and Natural Sciences

Clinical DSC Package and Repository Management

Avery Bell, Dr. Robert Buscaglia
Department of Mathematics and Statistics

Abstract

Clinical Differential Scanning Calorimetry (DSC) is a biophysical technique that evaluates the denaturation of human blood plasma to create a characteristic curve termed a thermogram. Thermogram curves differ based on the health status of the patient. This project used the programming language R and the library *tidyverse* to compile, organize, clean, and document IRB-protected data for lung cancer, melanoma, and auto immune disorders. The data consisted of thermograms and related patient meta information including age, gender, disease progression, and disease specific attributes. The patient meta information for each dataset was full of typos, inconsistent sample IDs, and extraneous information. All necessary cleaning steps were performed, the variables in each dataset were documented, and an R package was created to hold the DSC datasets. A private R package is managed on GitHub and can be easily accessed by clinical DSC researchers.

This project also resulted in a column expansion program designed to address challenges presented by patient meta information. Often when this data is compiled, a medication column is used to store all the different prescriptions for each patient. The proper format for analysis is to have a column for each medication and a corresponding value in each row to indicate if a patient takes that medication. The process of expanding a column into multiple columns with a single piece of information in each row is called a column expansion. This project resulted in a medication column expansion algorithm that can be used for current and future clinical datasets. Additionally, this project finalized a DSC analysis package, ***tlbparam,*** that provides a toolkit for calculating thermogram-related parameters.

Introduction

Clinical Differential Scanning Calorimetry (DSC) is a biophysical technique the involves the denaturation of human blood plasma. It provides measurements of excess specific heat capacity (C_p^{ex}), that are the result of thermal denaturation of the plasma proteome. Measuring the specific excess heat capacity across a range of temperatures results in a signature termed a thermogram (Figure 1)[1]. Thermograms are of interest because the curve characteristics are different depending on the health status of the patient that provided the sample.

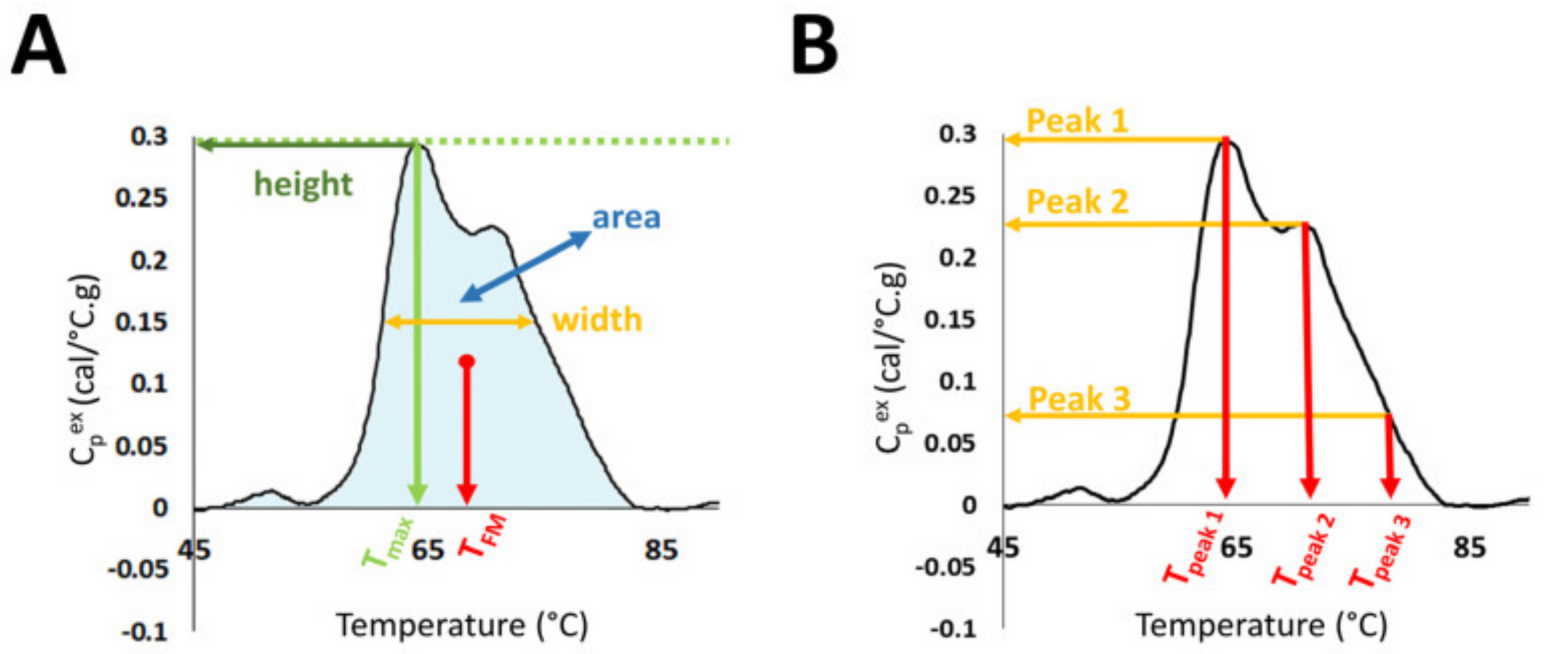


Figure 1. Panel A: Diagram of thermogram characteristics including the maximum height, total area under the curve, full width at half height, maximum peak height, and temperature at peak maximum. Panel B: Position of thermogram peaks 1, 2, 3.

This data management project performed the first step in data analytics, which is to coerce data into the proper format for analysis. It resulted in a fully functional R package that consists of documented DSC datasets ready for immediate analysis, and unique DSC data cleaning tools that address consistent issues, and an analysis package for summarizing DSC results.

Data

The data for this project consists of DSC results for lung cancer, melanoma, and a mixture of autoimmune diseases. A dataset for each disease consisted of two file, one containing DSC results, and the other containing patient meta-information. The patient meta-information contained common patient attributes such as age, gender, and ethnicity, as well as disease specific attributes. For example, the lung cancer meta-information includes a description of a patient's current smoking status, and a description of their smoking history such as the length of time and the number of cigarettes smoked per day. The patient meta-information required the most cleaning, to ensure that key clinical information was not lost or confounded during time of analysis. Figure 2 shows the median thermogram results for several of the diseases and healthy thermogram included in this project. Thermogram curves differ based on disease.

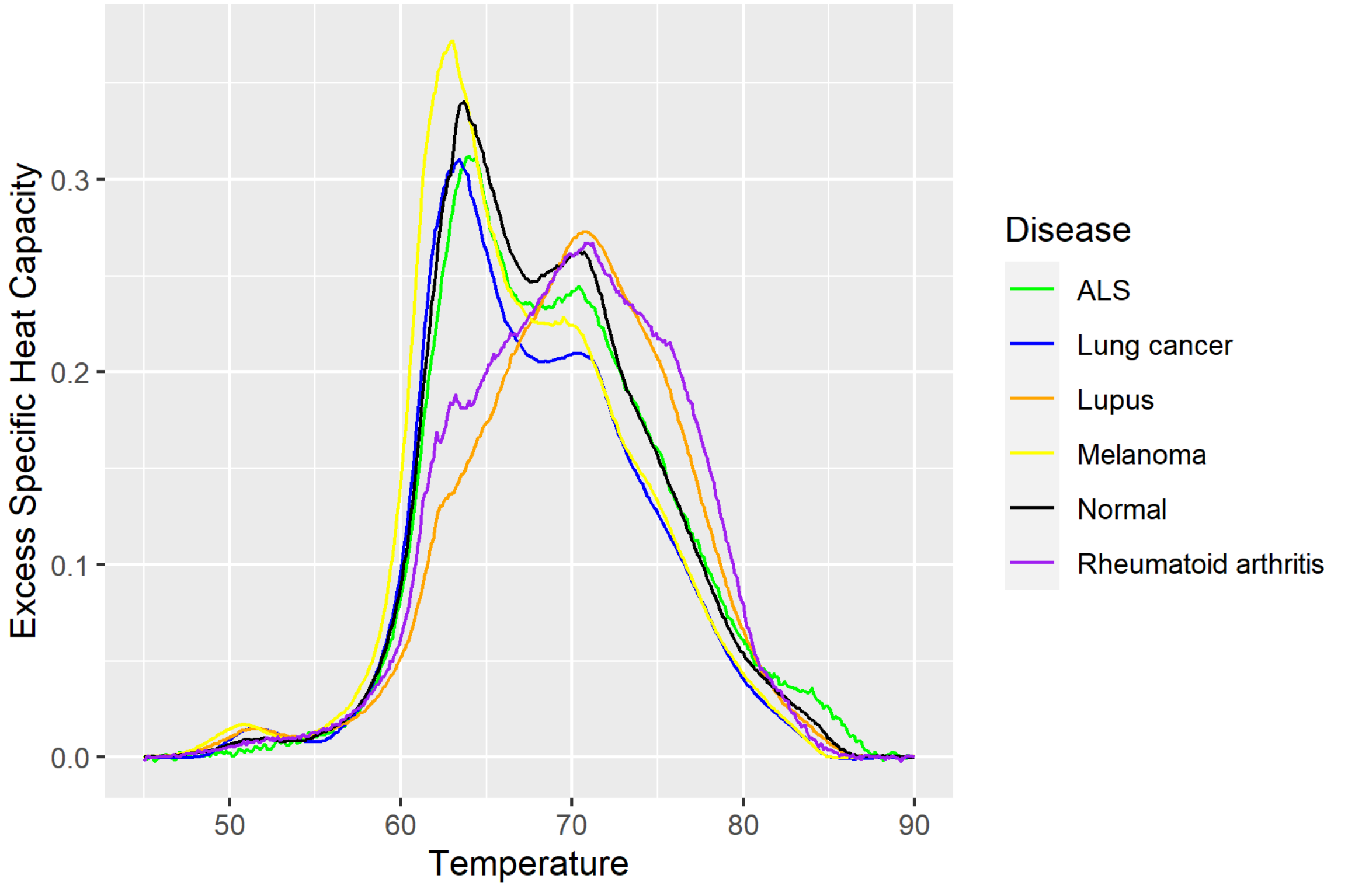


Figure 2. Median thermogram results for five diseases and healthy thermogram results (Normal). The thermogram curve for each disease is unique, and it is hypothesized that the difference in thermogram presentation can be used for disease diagnosis and monitoring.

Package Development

Cleaning and packaging each dataset followed a similar set of steps. First the DSC files were organized and placed in a data-raw directory. Accessing those files, the data was imported into an R environment. R scripts were used to clean each dataset. Each dataset is unique and required different cleaning steps. The data-raw directory also stored the specialized scripts that were developed to wrangle each dataset. This step resulted in a data structure called a data frame that held the cleaned data, these data frames are stored in the data directory. Next each variable was documented by creating an informational document that describes each column of a data frame. The directories and R scripts were compiled into a single R package. This R package is stored on GitHub and can be easily accessed by DSC primary investigators.

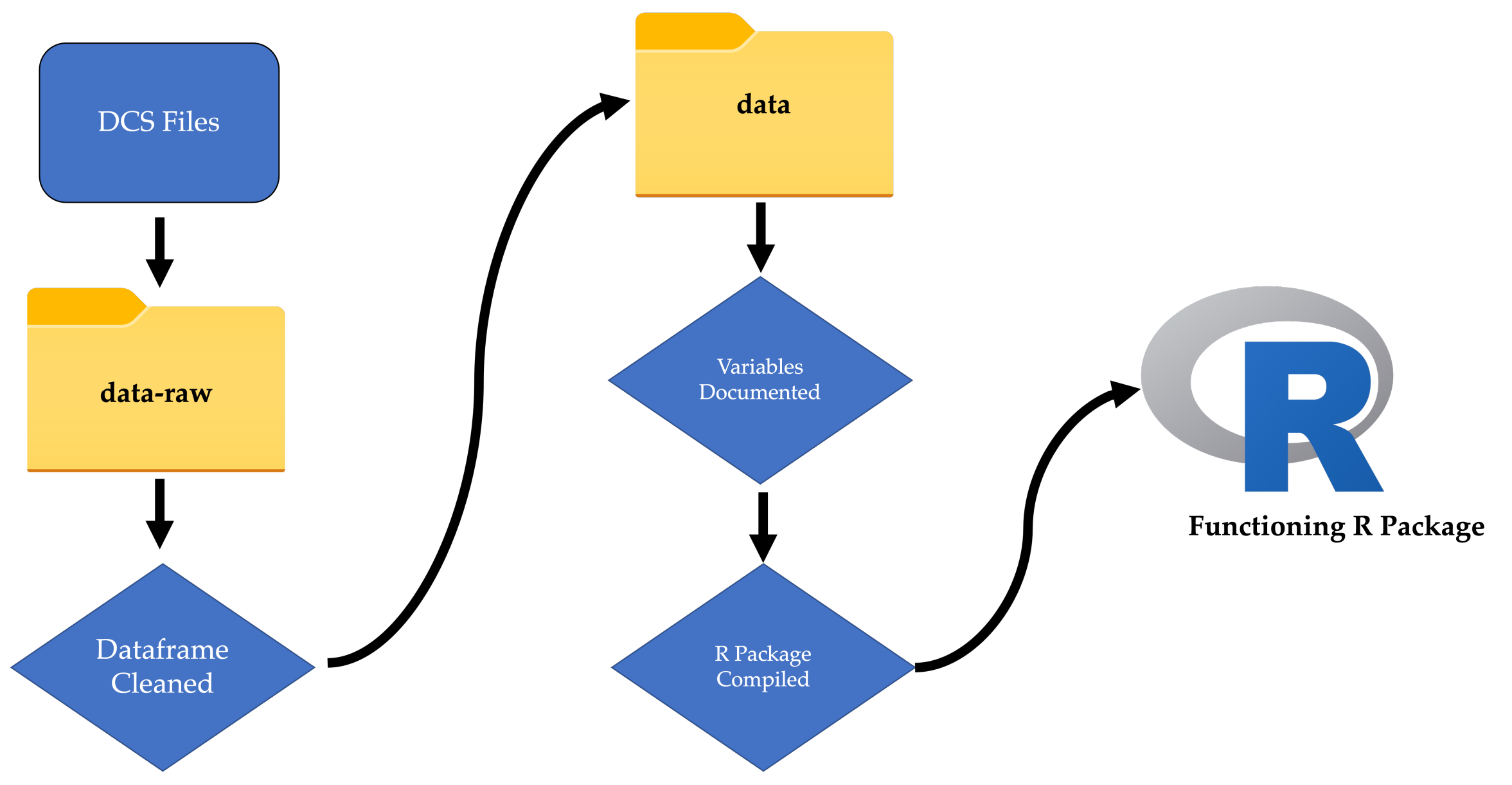


Figure 3. Package development workflow that was used to clean and document each DSC dataset. Original data files are stored in a data-raw directory, cleaned and an image of the cleaned data is stored in a data directory. The variables in each dataset are documented, and then the package is compiled.

Data Wrangling

Common cleaning tasks for each dataset:

- Identifying mismatched sample ids

sample_id	age	gender
a123	37	Female

sample_id	45.0	45.1	…
A123	0.001	0.0014	…

↓

sample_id	age	gender	45.0	45.1
a123	37	Female	NA	NA
A123	NA	NA	0.001	0.00143

Figure 4. An example of the sample_id inconsistencies that resulted in thermogram and meta information being joined incorrectly. Joins are case sensitive; differences in letter case prevent joining thermograms and corresponding meta information, resulting in NA values being insert into the datasets. Sample id inconsistencies were identified and corrected.

- Modifying column contents to correctly factor a variable
  - For example, the lung cancer dataset contains a Smoking Status column:

Factors
Current Smoker
Current smoker
Ex-smoker
Unknown
Non-smoker**
Never smoked
" "

→

Factors
Current smoker
Ex-smoker
Unknown
Never smoked

Figure 5. Row content correction requires properly factored variables, such as the Smoking Status column in the lung cancer dataset. All instances of "Current Smoker" was changed to "Current smoker", all instances of "Non-smoker**" was changed to "Never smoked", and empty rows were changed to "Unknown". The rows in the Smoking Status column can now be factored, and the values can be treated as categories.

- Correcting variable names
  - Replacing spaces with underscores
  - Consistent variable naming
  - Removing extra characters

Sample ID	Patient ID	Smoking Status:	Years Smoked:

↓

sample_id	patient_id	smoking_status	years_smoked

Figure 6. Example of the variable name corrections that took place for each dataset.

- Correcting datatypes
  - Converting numbers stored as strings to a numeric type
  - Converting strings to categories (factors)
- Column Expansions
  - A single column with multiple pieces of data expanded into multiple columns with a single piece of data.

1. Schneider G, Kaliappan A, Nguyen TQ, **Buscaglia R**, Brock GN, Hall MB, DeSpirito C, Wilkey DW, Merchant ML, Klein JB, Wiese TA, Rivas-Perez HL, Kloecker GH, Garbett NC. The Utility of Differential Scanning Calorimetry Curves of Blood Plasma for Diagnosis, Subtype Differentiation and Predicted Survival in Lung Cancer. Cancers (Basel). 2021 Oct 23;13(21):5326. doi: 10.3390/cancers13215326. PMID: 34771491; PMCID: PMC8582427.

Column Expansion

An issue present in all datasets was a single column that stored multiple pieces of information. For example, the autoimmune dataset contained a medication column that held each patients' prescribed medication. The proper format for analysis is an individual column for each medication, and either the prescribed dosage, or an "N" to indicate that the patient does not take that medication. There are no built-in *tidyverse* commands that can easily expand a single column, so a column expansion function was developed to perform this task.

Medication
Prednisone (10 mg), NyQuil (30 ml), Prednisone (20 mg)

↓

Medication	Prednisone_mg
NyQuil (30 ml), Prednisone (20 mg)	10

↓

Medication	Prednisone_mg	NyQuil_ml
Prednisone (20 mg)	10	30

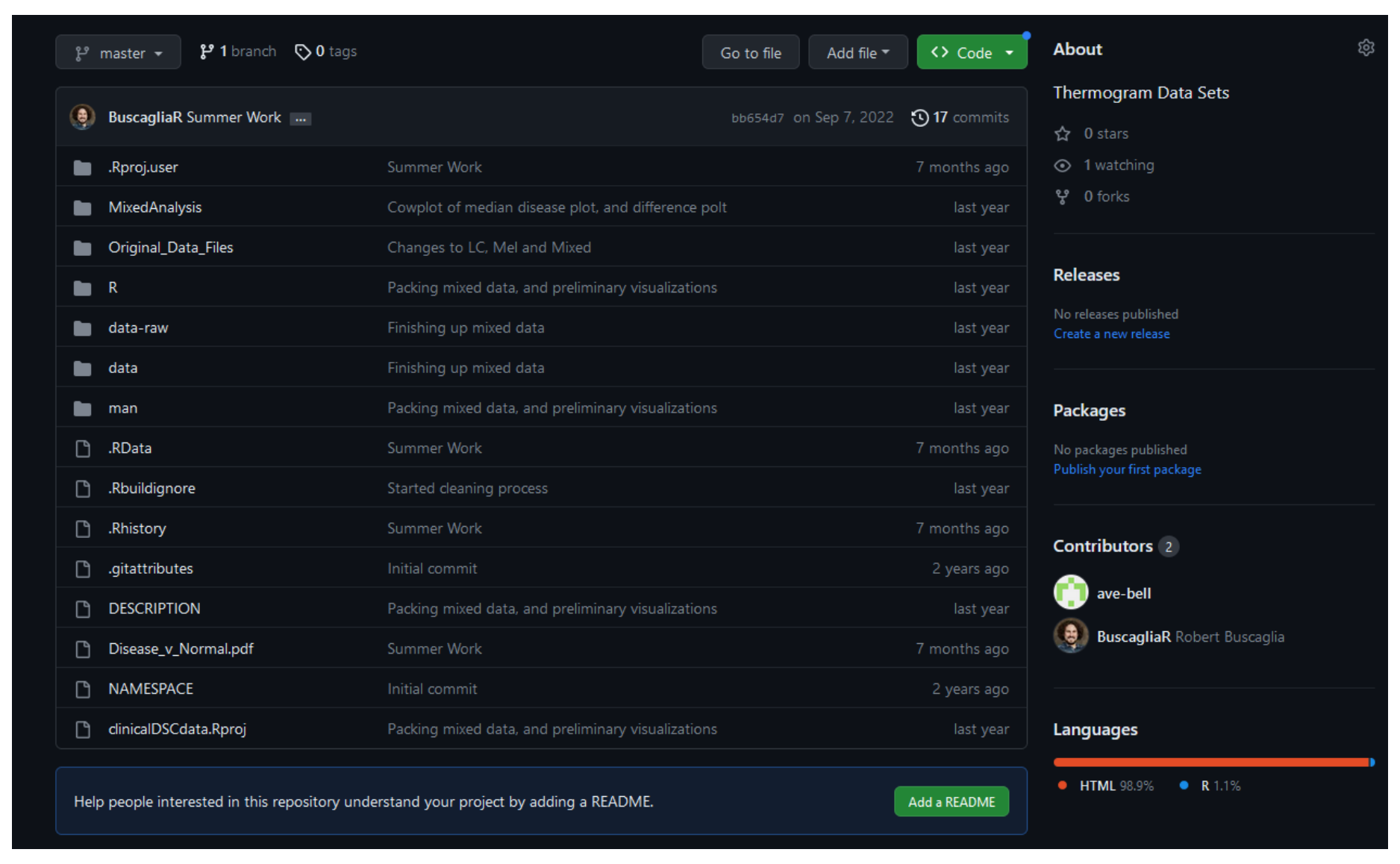
↓

Medication	Conflict	Prednisone_mg	NyQuil_ml
Prednisone (10 mg), NyQuil (30 ml), Prednisone (20 mg)	Prednisone_mg (10)	20	30

Figure 7. Medication column seen in raw data is cumbersome and requires an expansion before analysis can take place. Any conflict (i.e. multiple dosages) is recorded and provided as notes back to practitioners for follow up discussion.

Repository Management

The R package of cleaned datasets is stored on private GitHub repository. GitHub is a code management software that can be easily accessed and managed. It holds all versions of the package, can be easily updated, and clinical DSC investigators can easily download and use the package.



BuscagliaR Summer Work … bb654d7 on Sep 7, 2022 17 commits

.Rproj.user	Summer Work	7 months ago
MixedAnalysis	Cowplot of median disease plot, and difference polt	last year
Original_Data_Files	Changes to LC, Mel and Mixed	last year
R	Packing mixed data, and preliminary visualizations	last year
data-raw	Finishing up mixed data	last year
data	Finishing up mixed data	last year
man	Packing mixed data, and preliminary visualizations	last year
.RData	Summer Work	7 months ago
.Rbuildignore	Started cleaning process	last year
.Rhistory	Summer Work	7 months ago
.gitattributes	Initial commit	2 years ago
DESCRIPTION	Packing mixed data, and preliminary visualizations	last year
Disease_v_Normal.pdf	Summer Work	7 months ago
NAMESPACE	Initial commit	2 years ago
clinicalDSCdata.Rproj	Packing mixed data, and preliminary visualizations	last year

Help people interested in this repository understand your project by adding a README. Add a README

About
Thermogram Data Sets
0 stars
1 watching
0 forks

Releases
No releases published
Create a new release

Packages
No packages published
Publish your first package

Contributors 2
ave-bell
BuscagliaR Robert Buscaglia

Languages
HTML 98.9% R 1.1%

Figure 8. Image of private GitHub Repository for DSC datasets.

tlbparam: Thermogram Analysis Toolkit

An analysis toolkit for thermal liquid biopsy (TLB) clinical results developed as part of this project to improve ease of analysis. Thermal liquid biopsy is a growing literature term for clinical DSC. ***tlbparam*** provides a simple to read output from a complex set of biochemical information and calculates summary metrics related to important thermogram characteristics.

Parameter	Description
tarea	Total area under the thermogram signature
max	Maximum observed excess heat capacity
tpeak1	Temperature of peak corresponding to Peak 1 temperature region (60 - 67 C)
peak1	Height of peak corresponding to Peak 1 temperature region (60 - 67 C)
min	Minimum observed excess heat capacity.

Table 9. Example of parameters calculated by ***tlbparam***. Over 15 metrics are computed, not all are included in this table.

Final Products and Future Directions

The results of this project are a functioning R package with DSC datasets ready for common DSC data challenges, a custom data cleaning tool developed for common DSC data challenges, and an analysis package. These packages are stored on GitHub and can be easily accessed by DSC researchers for analysis. The column expansion tool is also stored on GitHub and can be used to perform column expansions on existing and future clinical datasets.

Future research will include building a SQL database to store DSC datasets as they grow larger and more diverse. Additionally, another HURA project will take place, focused on statistical and machine learning analysis of the lung cancer dataset, with the goal of classifying lung cancer thermograms and lung cancer subtypes.

Acknowledgements

This project was funded by the Hooper Undergraduate Research Award. Data was generously made available by Dr. Nichola C. Garbett, Associate Professor of Medicine, University of Louisville.