Extensive Patient Data Facilitates the Development of Personalized Cancer Treatment Plans

Phil McNamara[1], Carolyn Brewster[1], Anna Lundberg[1], David Degnan[1], Jaclyn Smith[2], Kevin Matlock[2], Gans Srinivasa[2]

[1]Bioinformatics & Genomics Masters Program, University of Oregon [2]Omics Data Automation, Portland, OR

INTRODUCTION

Cancer is responsible for approximately 14.6% of all human deaths, and survivability is dependent on multiple factors which include age, genetics, disease progression, and health status at diagnosis.[1,2] The interaction of these variables drives the uniqueness of each patient's cancer, complicating the choice of treatment.

Electronic health records (EHR) provide large amounts of medical data, which increase exponentially with modern automation.[2] EHR are designed for specific medical uses, and their format makes it difficult to ask meaningful research questions about the data. The creation of a database which relates all of this information together in a cleaned and standardized format will allow for exploration across multiple patients and types of data.

HYPOTHESIS

Extensive diagnostic, treatment, and demographic information for thousands of cancer patients organized in an intuitive, relational database can be used to research effective personalized cancer treatment plans.

WHY USE A RELATIONAL DATABASE?

Due to the export process, the data arrived in a disconnected and poorly-organized form. This dataset is particularly valuable because of the depth of information available for each patient. A **relational database** allows us to store the data in a modular format which still preserves interrelatedness.

The database structure enforces strict data types that allow users to ask complex questions based on the type of information involved. For example, a researcher could ask for the projected survival time based on patients' biological sex, BMI, number of surgeries, and white blood cell count. This kind of data exploration could also lead to the discovery of previously unknown risk factors.

GOALS

- Compile patient data into an intuitive relational database
  - Standardize data formats
  - Remove illogical typos
- Explore data to find trends
- Integrate with tissue images for more comprehensive and specific prediction

Patient Profile

Subject ID	**SN-00005810**
Age	75
Race	White
Ethnic Group	Non-Hispanic
Gender	F
Tobacco User	N
Status	Alive

Real data from an actual patient in the database

Aug 2012		Sep 2012	2014	2016	2017
Screening mammogram, mass discovered	Ultrasound confirms tumor	Lumpectomy, starts hormone therapy	Follow-up mammogram	Last mammogram, cancer clear	Last follow-up on therapy

FLOW — 1,784 entries
Includes details about heart and respiratory rates, pain ratings, and food intake

SURGERY — 1 entry
Lumpectomy (partial breast removal) with sentinel lymph node dissection

HOSPITALIZATIONS — 15 entries
Contains family history of breast cancer as well as surgical admissions, clinical trials, and long-term medication monitoring

VISITS — 69 entries
Includes mammograms and pre-operative exams

LAB RESULTS — 41 entries

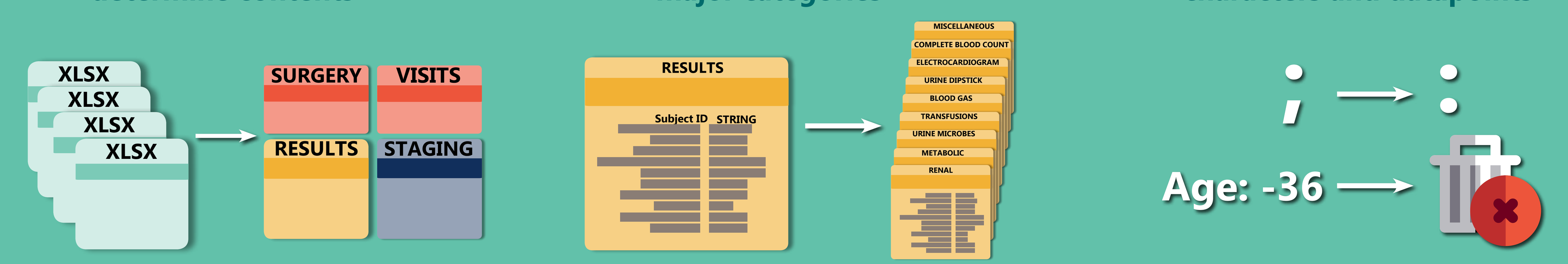
TUMOR STAGING — 8 entries
- ER+ breast cancer
- 20-50mm tumor
- 1-3 lymph nodes involved
- Microscopic evidence of metastasis

DATA

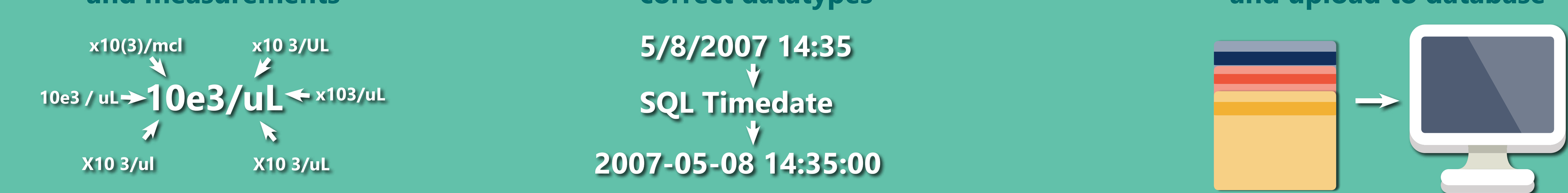
7,552 cancer patients
Spanning 2 decades
Over 1.5 million rows of data
Available data varies based on patient consent

METHODS

1. Format tables and determine contents
2. Separate tests into major categories
3. Remove problematic characters and datapoints



4. Standardize units and measurements
5. Determine and enforce correct datatypes
6. Optimize table parameters and upload to database



DESIGN

LEGEND

TABLE NAME	
Description of contents	
Primary Key	**TYPE**
Data Field Name	TYPE

VITAL SIGN TABLES

HOSPITALIZATIONS — 67,371 recorded procedure and diagnosis data from hospital stays

Subject ID	**STRING**
Hospital Admission	DATE/TIME
Admission Diagnosis	STRING
Admission Procedures	STRING

VISITS — 483,995 doctor visits with vital signs

Subject ID	**STRING**
Visit Date Time	DATE/TIME
Weight	REAL
Height	REAL
BMI	REAL
Blood Pressure	INTEGER
Visit Diagnosis	STRING

FLOW — 483,963 records of longitudinal test data (e.g. breathing, oxygenation, heart rate)

Subject ID	**STRING**
FSD ID	INTEGER
Recorded Date	DATE/TIME
Entry Date	DATE/TIME
Description Name	STRING
Measured Name	STRING
Measured Value	REAL
Unit	STRING

DEMOGRAPHICS — 7,552 sets of demographic information on patients in the database

Subject ID	**STRING**
Age	REAL
Race	STRING
Ethnic Group	STRING
Gender	CHARACTER (M/F)
Tobacco User	CHARACTER (Y/N)
Status	STRING (ALIVE/DECEASED)

SURGERY — 6,480 data points collected during surgery

Subject ID	**STRING**
Surgery Date	DATE/TIME
BMI	REAL
Height	REAL
Weight	REAL
Surgery Diganosis	STRING
Surgery Procedures	STRING
Surgery Description	STRING
Anesthesia Type	STRING
Hemoglobin A1C	REAL
Pre-Op Glucose	INTEGER
Post-Op Glucose	INTEGER
Pre-Op CBG	INTEGER
Diabetes	STRING
ASA Score	STRING
HPRO Type	STRING
Surgery Duration	INTEGER
Wound Class	STRING

TUMOR STAGING — 38,750 evaluations of cancer progression

Subject ID	**STRING**
Tissue Type	STRING
Diagnosis	STRING
Metastasis	STRING
RBC Size Distribution	STRING
Tumor Path	STRING
Lymph Node Path	STRING
Metastasis Path	STRING
Estrogen Level	REAL
Progesterone Level	REAL
Metastasis Site	STRING
Tumor Grade	STRING
Percent Tumor	INTEGER
Percent Necrosis	INTEGER
Gene Marker Test	STRING
Chromosome Location	STRING
Gene Marker Result	STRING
Tissue Image ID	STRING

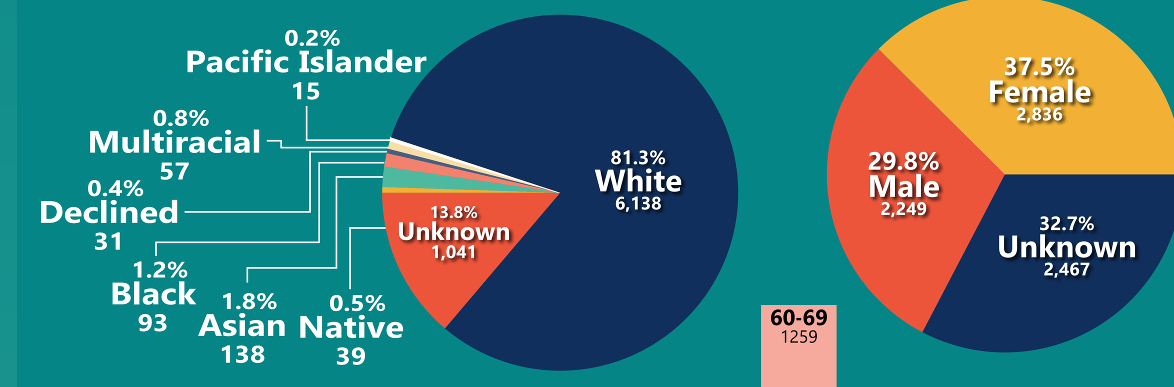
LAB TEST RESULT TABLES

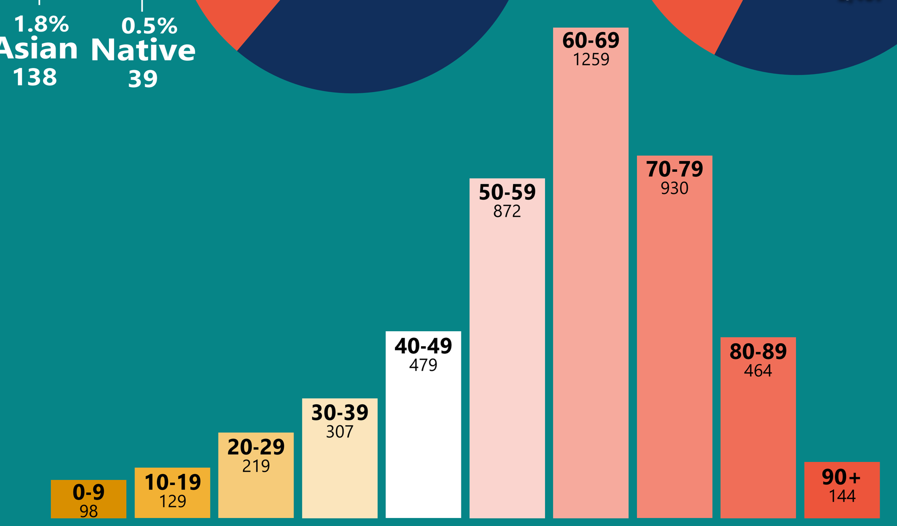
- **MISCELLANEOUS** — 139,173 uncategorized test results
- **COMPLETE BLOOD COUNT** — 127,362 blood counts by cell type
- **ELECTROCARDIOGRAM** — 13,827 ECG results
- **URINE DIPSTICK** — 16,038 urine dipstick tests
- **BLOOD GAS** — 11,625 blood gas concentration data points
- **TRANSFUSIONS** — 9,147 IV transfusions
- **URINE MICROBES** — 18,172 urine cultures
- **METABOLIC** — 113,326 records from metabolic panels
- **RENAL** — 25,482 kidney function lab tests

Subject ID	**STRING**
Lab Status	STRING
Result Status	STRING
Specimen Taken Time	DATE/TIME
Result Time	DATE/TIME
Description	STRING
Test Name	STRING
Test ID	REAL
Reference Range	STRING
Result Flag	STRING

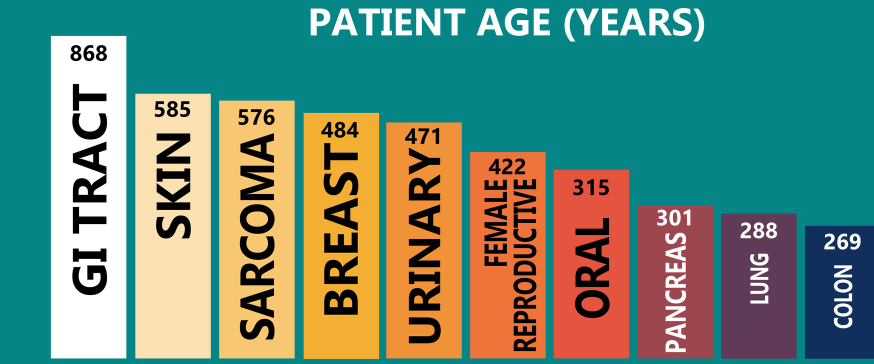
The PostgreSQL (version 10.5) database is hosted on a Red Hat Linux virtual machine at the University of Oregon.[3] Tables are connected by subject ID.

THE DATABASE PATIENTS





PATIENT AGE (YEARS)



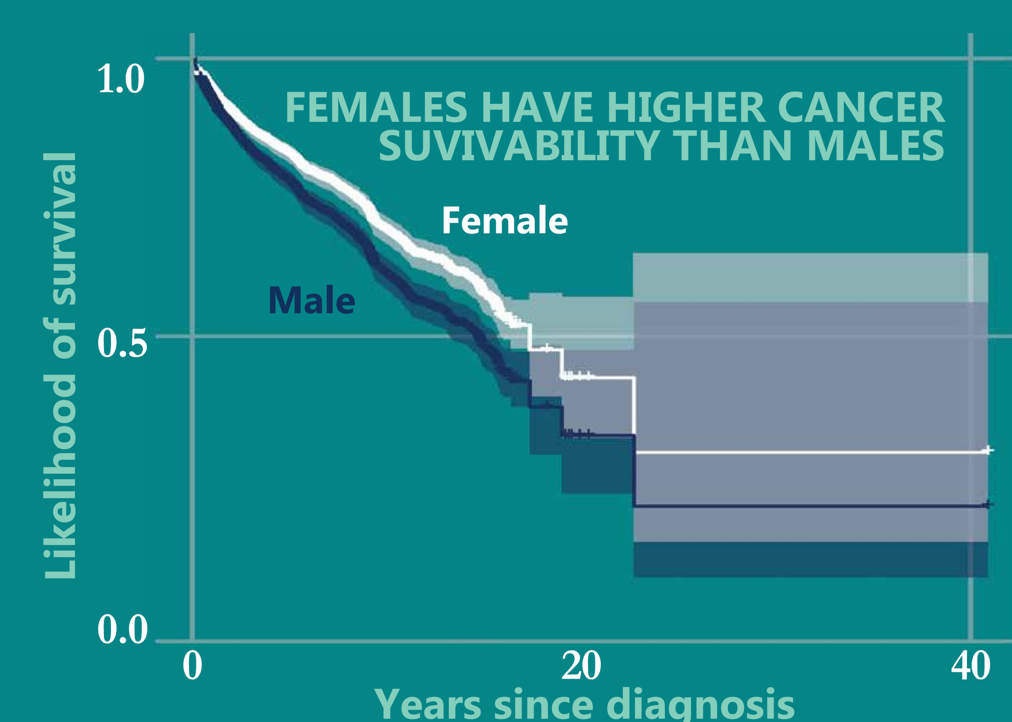
TOP TEN UNIQUE CANCERS REPRESENTED

PROPORTIONAL HAZARDS

- The Cox Proportional-Hazards model, a widely used regression analysis of survival data, estimates survivability based on predictors (called covariates).[4]
- This model assumes that hazard ratios do not change over time.
- This approach uses the date of a deceased patient's final datapoint as a proxy for time of death.

Survival Time   Baseline Hazard   Effect size   Covariate

$$h(t) = h_0(t) \times e^{\underbrace{b_1x_1 + b_2x_2 + \ldots + b_nx_n}_{\text{Hazard ratios}}}$$



FUTURE DIRECTIONS

- Expand database with additional patient data sets
- Integrate with machine learning tissue-imaging diagnostics
- Identify predictors of treatment response to facilitate "all in a day" development of personalized treatment plans

ACKNOWLEDGEMENTS

This project was made possible by the patients and families who shared their medical data for the benefit of others. We are grateful to Omics Data Automation and OHSU for their data, guidance, and support. We would also like to thank the BGMP and our mentors Stacey Wagner, Leslie Coonrod, Clay Small, Jason Sydes, and Pete Batzel for their input and expertise.

UNIVERSITY OF OREGON
BGMP@UO Bioinformatics and Genomics Master's Program
OHSU

REFERENCES

1. Pavlopoulou, A., Spandidos, D., & Michalopoulos, I. (2014). Human cancer databases (review). *Oncology reports, 33(1).*
2. Miller, K., et al., (2016). Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians, 66(4).*
3. PostgreSQL Consortium. History. https://www.postgresql.org/docs/10/biblio.html. Accessed October 2018.
4. Singh, R., & Mukhopadhyay K. (2011). Survival analysis in clincial trials: basics and must know areas. *Perspectives in Clinical Research, 2(4).*

Code, poster, & supplementary materials available at git.io/fpQBf