Visualizing Public Health Data

Anamaria Crisan, MSc
PhD student with Drs. Jennifer Gardy & Tamara Munzner
UBC School of Population and Public Health
Primary Research Question

To **what extent** and in **what ways** does the visualization of genomic, administrative, and contact network data support decision making for communicable disease prevention and control
Primary Research Question

To what extent and in what ways does the visualization of genomic, administrative, and contact network data support decision making for communicable disease prevention and control

aka. “How is visualization of communicable disease (public health) data useful? Can I quantify how useful it is?”
The Structure of this Talk

How I came to ask this question
Communicating with non-technical experts
Communicating cancer risk to patients
Statistics and data visualization
The Structure of this Talk

How I came to ask this question
Communicating with non-technical experts
Communicating cancer risk to patients
Statistics and data visualization

How I plan to answer this question
Data Visualization Research
Integration with Evaluation from Public Health
Examples of Work
Part 1:

How I came to ask the question
Disclaimer

I’ll be talking about a project I worked on while employed at GenomeDx Biosciences.

Everything I am presenting is publically available, but this doesn’t mean that I endorse their products or the products of their competitors.

Furthermore, I am relaying high level details of my own thought process during and after this project, not the thoughts of others at the organization.
I’m not an artist. I’m a data analyst.
Eventually I had Explain my Work to Experts with Different Backgrounds

I often used data visualization to explain the results of data mining and statistical techniques

But one day I got tasked with a rather challenging problem...
The Question:

The task: We had developed a genomic biomarker panel to assess a man’s risk of metastatic prostate cancer following prostatectomy

How do we communicate “risk”?
I wanted to take more ownership of the question “how do we communicate risk?”
I wanted to take more ownership of the question “how do we communicate risk?”

There wasn’t a simple answer
Just show a Number ...

---

### Framingham Risk Score - RESULTS

Your patient's Framingham Risk Score is **< 1%**

#### 2009 CCS Canadian Cholesterol Guidelines Recommendation

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>Initiate/consider treatment if any of the following:</th>
<th>Primary LDL-C targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW (FRS &lt; 10%)</td>
<td>LDL-C ≥ 5.0 mmol/L</td>
<td>≥ 50% reduction</td>
</tr>
</tbody>
</table>

Adapted from Genest et al. Can J Cardiol. 2009.1

Clinical judgment should be used regarding the timing of pharmacological therapy in low risk patients. Please consult guidelines for complete recommendations.

Clinicians should exercise judgment when implementing lipid-lowering therapy; lifestyle modifications will have an important long-term impact on health and the long-term effects of pharmacotherapy must be weighed against potential side-effects.

---

http://bit.ly/1Knjr19
Evidence from Risk Communication Literature

(difficult to understand) Probability < Frequency < Visualization
60% 6 in 10

(easier to understand)

Numeracy: the ability to reason with numbers

Individuals with low numeracy have a difficulty interpreting numbers and probabilities.

Visualizations can help people with low numeracy make sense of data.

But, there is some evidence that low numeracy affects reasoning with graphs as well.

Example: Data Visualization in Shared decision Making

STUDY DESIGN

Quasi-randomized trial with four conditions
Outcome: correctly calculating the risk (essentially a math test)

RESULTS

Visualization improved comprehension of both doctors and patients
Visualization improved concordance between doctors and patients

Garcia-Retamero et. al (2013) "Visual representation of statistical information improves diagnostic inferences in doctors and their patients"
Yes! Data visualization was more than a “nice to have”!
Show a Number and a Picture

Example Report: OncotypeDx DCIS report

Breast Cancer Report - Node Negative

Prognosis

Patient ID: DOE, JANE
Sex: Female
Date of Birth: 01-Jan-1950
Medical Record/Patient #: 56887777
Date of Surgery: 25-Sep-2008
Specimen Type ID: Breast/SURG-0001

Recurrent Score Result

OncoType DX Breast Cancer Assay uses RT-PCR to determine the expression of a panel of 21 genes in tumor tissue. The Recurrence Score result is calculated from the gene expression results and ranges from 0-100.

The findings are applicable to women who have stage I or II node negative (N0-1), estrogen receptor positive (ER+) breast cancer and will be treated with 5 years of tamoxifen (tam). It is unknown whether the findings apply to other patients outside these criteria.

Clinical Experience: The following results are from a clinical validation study that included 669 patients from the NSABP B-14 study. The study included female patients with stage I or II, N1+, ER+ breast cancer treated with 6 years of tamoxifen.

Prognosis: 10-Year Risk of Distant Recurrence after 5 Years of Tam, Based on the Recurrence Score Result (from NSABP B-14)

10-Year Risk of Distant Recurrence

Tam Alone
7% (95% CI: 4%-9%)

Graph showing the 10-Year Risk of Distant Recurrence for different risk groups.
Show a Number and a Picture

Example Report: Myriad Prolaris Prostate Cancer Test Report

Prolaris Score: 1.8

- Considerably More Aggressive Than Average AUA Low Risk
  
  **Interpretation:** The Prolaris Score of 1.8 indicates that this cancer is considerably more aggressive than the average cancer in the American Urology Association (AUA)* Low Risk category.

- 10-Year Prostate Cancer-Specific Mortality Risk: 5% (95% CI: 2.1-11.1%)
  
  **Interpretation:** The patient has a 10-year mortality risk of 5% if managed conservatively. Mortality risks could be altered by various therapeutic interventions.

The above chart illustrates the AUA Low Risk category, which is composed of patients with varying degrees of cancer aggressiveness. Cancer aggressiveness can be stratified within this category based upon Prolaris Scores, which are indicated below the graph.*

- US Distribution Percentile: >99%
  (For AUA Low Risk)
  
  **Interpretation:** >99% of patients in the AUA Low Risk category have a lower Prolaris Score.

**CLINICO-PATHOLOGIC FEATURES USED FOR ANALYSIS**

PSA Prior to This Biopsy: 5.2

Patients with similar clinico-pathologic features, as defined by their CAPRA score, have the same a priori 10-year prostate cancer-specific mortality risk. The addition of the Prolaris Score further differentiates this risk, as illustrated in the above graph, which is specific to this patient's CAPRA score. The orange line depicts the relationship between the Prolaris Score and the mortality risk with the 95% confidence interval indicated by dashed lines and the patient's Prolaris Score indicated by the orange dot.
Example Report: Decipher Prostate Cancer Test Report

Primary population: Men, who are susceptible to red-green colour blindness
Example: Deciding upon an Intervention

Baseline Visualization

Helping breast cancer patients decide between multiple treatment options.

Zikmund-Fisher (2013). A demonstration of “less can be more” in risk graphics.

Beyond Building Pretty & Cool Visualizations

Data visualization is not art
Beyond Building Pretty & Cool Visualizations

Defining Data Visualization

(I argue data visualization is much more about design)

Ideas taken from @rachelbinx’s 2016 OpenVis talk
And http://featureguru.com/art-vs-design.html
There’s more a Visualization than Meets the Eye

Final Data Visualization

TB incidence rates overlain on geography (BCCDC reportable disease dashboard)

Iceberg Ideas borrowed from @rachelbinx’s 2016 Open Vis talk
There’s more a Visualization than Meets the Eye

Final Data Visualization

TB incidence rates overlain on geography
(BCCDC reportable disease dashboard)
But there was a lot that went into creating that simple visualization

Data
- **We rarely visualize raw data**
  We often derive we data
- We combined multiple dataset
- Data has issues of quality
There’s more a Visualization than Meets the Eye

But there was a lot that went into creating that simple visualization

Alternative choices
Picked this choice of visualization over others
There’s more a Visualization than Meets the Eye

But there was a lot that went into creating that simple visualization

Visual & Interactive Design

Visual Design:
How data visualized data looks

Interaction Design:
How to interact with the data visualization
There’s more a Visualization than Meets the Eye

But there was a lot that went into creating that simple visualization

Motivations

Increasing public awareness
Allocate Resources
Monitor program progress
Target outreach programs
But there was a lot that went into creating that simple visualization

Tasks
(Atomized components of the motivation)

Communicate rates of TB by Health Service Delivery Area (HSDA) region

Overlay descriptive statistics on geography
There's more to data visualization than simply communicating numerical data
Example: Hypothesis Generation

Allowed John Snow to form the hypothesis of what may be leading to the cholera outbreak

John Snow’s Visualization of the 1854 Cholera Outbreak
Example: Hypothesis Generation

Allowed John Snow to form the hypothesis of what may be leading to the cholera outbreak

John Snow’s Visualization of the 1854 Cholera Outbreak
Example: Checking Assumptions of Statistical Models

Anscombe’s quartet, four datasets that have near identical descriptive statistics but that look very different when visualized.

Anscombe, F. (1973) “Graphs in Statistical Analysis”

Data visualization has long complemented applied statistical practices. Consider Tukey’s classic “Exploratory Data Analysis”, which is rife with suggestions for how to visualization data.
Example: Visualizing Public Health Data

Note that several health regions are comprised of aggregated geographical areas.

Tuberculosis, 2003 to 2014, All BC

Legend:
- Region Rate
- Canada Rate
A Data visualization in 3 Questions:

**Why? (Motivation)**
Why do you need to visualize data?

**What? (Data)**
What kind of data is being visualized?

**How? (Visual and Interaction Design)**
How is data being visualized?
A Data visualization in 3 Questions:

**Design**
- Why?

**Evaluation**
- Does the visualization solve a relevant problem?

**What?**
- Are you using the right data, or deriving the right data?

**How?**
- Are the visual and interactive design choices appropriate?
Steps to Design and Evaluate a Data Visualization

**Why**
- Domain situation
  - Observe target users using existing tools

**What**
- Data/task abstraction
- Visual encoding/interaction idiom
  - Justify design with respect to alternatives

**How**
- Algorithm
  - Measure system time/memory
  - Analyze computational complexity
  - Analyze results qualitatively
  - Measure human time with lab experiment (lab study)
  - Observe target users after deployment (field study)
  - Measure adoption

**DESIGN**

**EVALUATION**

Munzner (2014) "Visualization Analysis and Design"
### Steps to Design and Evaluate a Data Visualization

#### Methodology

<table>
<thead>
<tr>
<th>Qualitative Methods, Domain Knowledge</th>
<th>Design &amp; Cognitive Science</th>
<th>Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative &amp; Quantitative Methods</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Why

**Domain situation**
- Observe target users using existing tools

#### What

**Data/task abstraction**
- Visual encoding/interaction idiom
  - Justify design with respect to alternatives

#### How

**Algorithm**
- Measure system time/memory
- Analyze computational complexity
- Analyze results qualitatively
- Measure human time with lab experiment *(lab study)*
- Observe target users after deployment *(field study)*
- Measure adoption

---

*Image of diagram showing the steps.*
Part 2:

How I plan to answer the question
How Data Visualization is like Statistical Modelling

Model selection is a design problem

\[ Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \]

- Input data (to fit the model)
- Parameters
How Data Visualization is like Statistical Modelling
“Parameters” of Visual and Interaction Design!

Five dimensions are plotted in 2D
(4 continuous dimensions & 1 categorical dimension)
Basic Building Blocks of Data Visualization

“Parameters”

Munzner (2014) “Visualization Analysis and Design”
How Data Visualization is like Statistical Modelling

“Parameters” of Visual and Interaction Design!

- Colour = Continent
- Transparency = Density

Reveal detail on hover

Health expenditure vs. life expectancy, 2011

Canada
- Health Expenditure: $5629.726739
- Life Expectancy: 80.9292439
- Region: North America
- Population: 34483975

Legend:
- East Asia & Pacific
- Europe & Central Asia
- Latin America & Caribbean
- Middle East & North Africa
- North America
- South Asia
- Sub-Saharan Africa
How Data Visualization is like Statistical Modelling

“Parameters” of Visual and Interaction Design!

The same parameters can be combined in different ways to yield different visualizations.
How Data Visualization is like Statistical Modelling

A finale note on parameters

For brevity, I haven’t exhaustively described all the different components, which I’ve called parameters, that can be a part of data visualization

For more in depth details consider: Visualization Design and Analysis (2014) by Tamara Munzner
How Data Visualization is like Statistical Modelling

OPTIMIZATION!

Finding the best model

Searching the parameter space for a model that yields the lowest error
How Data Visualization is like Statistical Modelling

The “Design Space” metaphor

[Diagram showing a grid of data points labeled as good, okay, or poor]
How Data Visualization is like Statistical Modelling

OPTIMIZATION!
The “Design Space” metaphor

Sedlmair 2012
Progressively Identify the Right Visualization

The “Design Space” metaphor

Use “why, what, and how” framework to guide the selection of the optimal design choice

Sedlmair 2012
The Importance of Thinking Broadly

Use “why, what, and how” framework to guide the selection of the optimal design choice

Munzner (2014) “Visualization Analysis and Design”
Designs for Visualizing Health Data (http://www.vizhealth.org/)
How Data Visualization is like Statistical Modelling

A final note

Data visualization and statistical modelling are not identical, even though at a high-level they share similar research processes.

I’ve presented one aspect of visualization research, but there are others I haven’t touched upon.

I’ve emphasize problem driven work – finding the right visualization for a specific motivation or task – but there also exists technique and systems type research.
How to Implement Data Visualizations

Matthew Brehmer's totally subjective ranking of vis design tools
BUT.....

How do we design good visualizations for public health?
Motivations Underlying my Doctoral Work

Decision Support
For communicable disease prevention and control

Design Space
Characterizing and evaluation the design space of public health microbial genomics
Motivations Underlying my Doctoral Work

Methodology
Designing and evaluating data visualizations through a public health lens

Decision Support
For communicable disease prevention and control

Design Space
Characterizing and evaluation the design space of public health microbial genomics
DECISION SUPPORT

Visualizing Tuberculosis data at the British Columbia Centre for Disease Control
WHY
Combining Data will Prepare us for the Pandemics of the Future
But, that’s a lot of data....
Can Visualizing TB data help Decision Support?

We wanted to create an interactive and visual tool that allowed our public health stakeholders to analyze the different data types.

We want to understand how this tool can be used by different public health stakeholders:

- Medical Health Officers
- TB Clinicians
- TB Nurses
- Researchers
- Epis / Biostats
Treatment
Genomic
Contact Network
Patient Data
Outcomes
Geography / Location
Treatment

Genomic
TB whole genome genotyping

Outcomes

Patient Data

Contact Network

Geography / Location
HOW
An iterative approach to development allows us to get feedback before committing to ineffective design choices.
The Big Picture
But this takes a lot of time & effort

Latent TB
Patient Demographic & Treatment Data

Active TB
Genomic (Whole TB Genome Sequencing Data)

Active TB
Genomic (MIRU-VNTR) & Contact Data

Active TB
Patient Demographic & Treatment Data

Effort

Time
Introducing EpiCOGs

EpiCogs is a data viewer and currently a sandbox environment for developing data visualizations.
Factors Influencing the Current Design

Task: Filter patients and identify where they are
Filter patients from the side panel, and interactively update the line list & map based upon those interactions

Task: Follow-up on selected patients
Select patients view – a subset of the data. For outreach nurses, and request to include driving directions.

Task: Incorporate existing statistical methods
Analysis modes, allows epidemiologists and biostatisticians to integrate their R methods into EPI COGS

Task: Provide overview of key metrics
Predefined analysis modules that in the future will be migrated to “reports” section.
Factors Influencing the Current Design

Needs of individuals
Gathered through meetings, dialogue with individuals, and various iterations of EpiCOGs

Technology Changes
Support for data visualization tools in R improved greatly allowing for the creation of better data visualizations

Data Driven Interface and Analysis
Created a data driven interface that is responsive to the user’s data.

Policies and Procedures
Existing policies and procedures at the BCCDC inform the utility of such a tool and how it can integrate into existing workflows
Initial Work & Next Directions

Much initial work was to understand the tool’s feasibility

Could it meet the needs of stakeholders?
How could it integrate (security & workflow)?
How could it be supported long term? (Choice of R)
Could we build a useful tool in R?

Next phases will explore genotypes, genomics, and contact networks

Right now, users can filter based on assigned genotype clusters (which will show patients on map), but we’re working towards better visual and interactive design for these data
This is an Open Source Project

TRY THE DEMO:
https://amcrisan.shinyapps.io/EpiCOGSDEMO/

GET THE CODE
(& contribute to the project!):
https://github.com/amcrisan/EpiCOGS/
Call for Guinea Pigs!

To make relevant tools I need feedback!
If you want to be involved and get project updates let me know!

E-mail: anamaria.crisan@bccdc.ca

Twitter: @amcrisan

Web: cs.ubc.ca/~acrisan
Design Space
Exploring the Public Health Microbial Genomics Design Space
WHY
Can we Define the Design Space for Microbial Genomics?
WHAT
Can we Define the Design Space for Microbial Genomics?

Research literature and public documents already contain visualizations that are commonly used Public Health and specifically for microbial genomics.

Annotate those visualizations to develop a code set for “why, what, how”
Example: Outbreak Narratives
<table>
<thead>
<tr>
<th>Domain Specific Terminology</th>
<th>How</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 Genomic</td>
<td>Phylogenetic Tree</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Hierarchical Clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Matrix</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.1 Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1.1 Clades</td>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Map Colour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.1.2 Derived</td>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Arrange Order</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Map Colour</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.1.3 Genotype</td>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Facet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.2 Node Distance</td>
<td>Align root</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Annotation</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Scatter Plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.3 SNV/Ps</td>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Map Colour</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.4 Spatial Distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1.5 Temporal Distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Part 3:

Take home messages
Beyond Building Pretty & Cool Visualizations

Data visualization is not art
It is a research process.
Take Home Messages

Data Visualization is not an art or graphic design project
Relevance (utility) and usability trump aesthetics
Take Home Messages

Data Visualization is not an art or graphic design project
Relevance (utility) and usability trump aesthetics

Deciding upon the most appropriate data visualization can be a research problem
Design & Evaluation
Think about “why, what, and how” framework
Parallels to finding the right statistical model
Take Home Messages

Data Visualization is not an art or graphic design project
Relevance (utility) and usability trump aesthetics

Deciding upon the most appropriate data visualization can be a research problem
Design & Evaluation
Think about “why, what, and how” framework
Parallels to finding the right statistical model

Think broadly, progressively find the right data visualization

The Design Space Concept
Iterative development
Genomics is Becoming more Important

Fiona Brinkman @fionabrinkman · May 26
This would work not be possible without these fine people

The Gardy Lab
Dr. Jennifer Gardy
Jennifer Guthrie

PHSA Reference Laboratory
Dr. Patrick Tang
Hope Lapointe
Clare Kong

BCCDC CDPACS
Ciaran Aiken
Laura MacDougall
Mike Coss
Sunny Mak
Mike Otterstatter
Robert Balshaw

UBC Computer Science
Dr. Tamara Munzner
The InfoVis group

BCCDC Clinical TB Clinical Team
Clinicians
Dr. Maureen Mayhew
Dr. James Johnston
Dr. Jason Wong (CPS)
Dr. Victoria Cook

Nurses
Nash Dhalla
Michelle Mesaros

Epidemiologists
Dr. David Roth

The large team of individual’s from BC’s HAs and HSDAs without whom there would be no data.