Identifying Probands at Risk of Developing Hereditary Cancer Based on Family Health History and Clinical Practice Guidelines

Group number: 09
Group members: Jordon RItchie
CPCS 4030/6030, Fall 2020
Project Report

Introduction [MAX 150 words]

Family health history (FHx) is the most important indicator of a probands risk for developing hereditary cancer.[1,2] FHx includes personal and family cancer histories and relies on information related to age of cancer onset, number and types of cancers present, genetic testing and positive mutations, along with other disease specific details. Clinical practice guidelines are sets of criteria that use FHx to evaluate whether or not certain thresholds are met that would result in an increased risk of developing hereditary cancer. We relied on criteria from guidelines published by the American College of Medical Genetics and Genomics (ACMG) and the National Comprehensive Cancer Network (NCCN).[3–5] In the event of a positive indication, the guidelines typically recommend a cancer genetic consultation and/or genetic testing to investigate more closely the proband's risk of developing hereditary cancer.

Dataset [MAX 150 words]

The dataset used here was derived from a dataset collected in November of 2019 during an ad campaign that ran for the duration of the month.[6] We collected data using a user friendly chatbot and used web APIs and ontologies to produce boolean fields for each participant to indicate whether or not criteria from specific guidelines were met. Data is stored in a SQL database. Two data sources were derived from the database. The first has boolean fields for meeting criteria include has_acmg and has_nccn. These indicate that criteria specific to the organization were met regardless of the cancer indicated. The remainder of included boolean fields are specific to indicated cancers and include criteria from both ACMG and NCCN. Additionally, we recorded zip codes in the database from which state, county, and latitude and longitude coordinates could be derived for our first dataset. The second source is much smaller but simply lists the total number of probands that did not meet criteria but do have cancer in their family and what percent breakdown by cancer type.

Design solution [NO WORDS LIMIT]

Overview
I produced two visualizations, one of which has an interactive component. The first visualization is a map of U.S. counties showing the number of participants per county. It allows filtering on several key sub groups of the data. Specifically those subgroups are All, NCCN, ACMG, Breast and Ovarian, Colorectal and Endometrial, Li Fraumeni syndrome, and Gene Mutation. These represent key groupings of guideline criteria met by this set of probands. The second visualization is a stacked bar chart that examines the probands who did not meet criteria more closely. Specifically, the stacked bar chart looks at the number of cancers that run in the families of those who failed to meet clinical practice guideline criteria and indicates the percentage of those cancers by type.

Design Process Visualization 1
Truthfully, a map of the U.S. with a color scale that bins counties by how many participants reported living in the associated zipcode may not have been the best choice for the first data set. If the dataset were substantially larger it would have been more impactful. The most valuable part of the map is a spatial representation of the respondents. Salt Lake City, Utah and South Carolina had the counties with the densest response count (which makes sense as the app that collects the data was predominantly advertised in these states). By using zip code to mark counties and color as a channel, this spatial organization of respondents is readily observable.

The boolean fields in the first data set are not immediately marked on the map, however, they are aggregated behind the scenes in the code and when the map is filtered by guideline type, those fields modify the set of data displayed by the map. This was easily the most challenging part of the visualization. The color scale unfortunately was not a one size fits all and the range of the bins for the color scale had to be dynamically set to accommodate the different filtered data sets. This resulted in very different ranges varying from 8.5—42.5 respondents for all participants to 1—3 respondents for gene mutation participants. With a greater number of participants than the whole data set contained, the need to dynamically set these ranges might go away by reducing the artificial peaks observed in Utah and South Carolina. In any case, the visualization is still useful for understanding the distribution of respondents for different subgroups of the full data set.

Design Process Visualization 2

The stacked bar chart was a great fit for the second data set. Type and number of cancers are important elements for using guideline criteria for evaluating patient FHx. The greater the number of instances of cancer in a family and the greater the percentage of those cancers that fall under brain, breast, colorectal, endometrial, gastric, leukemia, melanoma, pancreatic, prostate, kidney, thyroid, or sarcoma increase, the more likely they are to meet guideline criteria. The marks for this visualization are lines and a color channel is employed to indicate stacks, or percentages of that line that account for certain types of cancers. Number of cancers is listed across the x-axis and percent is marked on the y-axis. Across the top I also included the total number of probands with exactly that number of cancers e.g. 681 probands had exactly 2 cancers in their family despite not meeting criteria. There were no pancreatic cancer cases that fit the requirements to be included in the graph.

Importantly, this graph effectively shows the majority of cancers reported in families that did not meet criteria were listed as 'other' cancers. This means it was not reported to us what kind of cancer ran in their family and guideline criteria could not be properly applied. It also reveals that breast cancer, colorectal cancer, and melanoma, all cancers for which guideline criteria were implemented, comprised the next three largest stacks in the bar chart. These are high risk cancers and generally have significant hereditary contributions to their onset. The most likely reason these reported cancer cases appear in families who do not meet hereditary cancer guideline criteria is advanced age of onset and cancers being split across both sides of the family resulting in high cancer case count within the family but diluted overall effect of hereditary contribution to disease given the cancer cases weren't all on the same side of the family.

Literature review [NO WORDS LIMIT] [Only for graduate students]

This section should contain a discussion of research papers related to the visualizations used in your website. NOTICE, here you shall talk about research papers focusing on visualization. Do not include research papers related to the domain problem your project is based on. In this section you can also refer to material we discussed in specific lectures during this class.

Visualization 1

Maps have a long history and evolution within the realm of visualization and despite our experience producing them, they remain a challenging medium within which to present data. The newest layer of map making that holds equal share promise and pitfall is the ability to add interactivity.[7] The addition of interactivity to map making effectively empowers the user to more effectively apply the data visualised by the map to further their purpose. The downside is interactivity might not be used at all. In fact, poor interactive might even become a detriment to the use of the map even as a static resource. Interactive maps are at their best when the user and the designer seem to meet in the middle. Interestingly, this draws on many concepts found in the field of usability and user experience. It also overlaps with many key concepts familiar to those running a startup. Essentially the interactivity feature of the map needs to meet the users needs — hence the name 'user-centered' when describing this approach. If the interactivity feature doesn't solve the users problem the way they want it solved, they aren't going to 'buy' the solution i.e. there is no product market fit. In the startup world this is equivalent to developing a product that solves a problem no one cares about — a fast way to go out of business.

This was a hard problem to solve in the interactive map produced in this project. The filters that mattered to the user for this particular dataset produced challenges for the color scale and representation of the data on the map. Those filters make sense to a genetic counselor and represent key groups of patients worth visualizing, but because of the relatively limited size of the data set, different thresholds had to be set for the range of values depending on the filter the user applied. This resulted in dramatically different ranges being applied to the same colors in the color scale across different groups of the data. For example, in Salt Lake City, Utah, over 50 probands responded overall, but only 3 reported a positive genetic mutation. However, for these two different filtered sets of data, that county had the same dark blue color because it represented the max end of the range for both sets. This gives the impression that all respondents in Salt Lake City, Utah county reported positive gene mutations if the range is not appropriately consulted for each filtered set.

Despite this challenge it was still valuable for the data to be tied to a physical location on the map, a key function of cartography in general.[8] This was key to understanding the distribution of respondents across different states and counties in the U.S. For example, we discovered the disproportionate participation in Salt Lake City, Utah and across South Carolina while also observing lack of response in several states including California, Arizona, Colorado, Arkansas, and Alabama. Proband FHx tied to geographical location gives us insight and clues into the demographics, socioeconomics, and environmental circumstances and factors.

An additional challenge that had to be taken into consideration was what to exclude from the map and its interactivity. This is a one of four issues outlined by Annette M. Kim in making maps more impactful to users.[9] There were more filters that could have been included in the map produced in this project but the data associated had sufficiently low counts to warrant them of little value visualized on the map. Ms Kim argues that these types of omissions are critical to make as they may inadvertently render important aspects of the data invisible to users. This can be a blind spot for map developers and hinders the goal of delivering a market fit for the user mentioned previously. Additionally, it prevents users from drawing incorrect conclusions or making potentially misleading assumptions about the data and its presentation.

Visualization 2

Bar charts are widely prevalent in data visualization. The stacked bar chart is a variation of the bar chart that is most suited to visualizing counts in contrast to box plots which are better suited for visualizing distributions.[10] This means the stacked bar chart is perfect for representing counts of probands who failed to meet criteria but still had cancer in their family along with the breakdown of cancer types per cancer count. Interestingly, in a study of visualizing proportions, they found that despite stacked bar charts were superior to doughnut and pie charts at performing a task, the majority of their participants (75%) preferred pie charts for their aesthetics, and just under half of their participants (44%) preferred pie or doughnut charts for their ease of understanding. The irony lies in the fact that the pie chart and (by virtue of essentially being the same chart) the doughnut chart, represent a class of visualizations that have received perhaps more criticism than any other class of visualizations.[11,12] Apart from the ease of misleading users by manipulating them in 3D, bonafide 2D pie charts can still be challenging to interpret because humans in general are bad at comparing angles. At best this makes pie charts generally poor for displaying data and at worst purposefully misleading. The Stacked bar chart displays things in lines (that can be viewed as areas) that are much easier to evaluate by comparative length (or width and height). As discussed in class, these metrics are much easier for the human brain to compare and decrease the potential for misleading audiences.

Conclusion

The visualizations studied in this project, interactive maps and stacked bar charts, provide a sweeping overall view of the data with reasonable interaction for the target audience and meaningful view into a specific subset of the data respectively. While the interactive map may not perfectly meet the needs of the user and leaves something to be desired in meeting the needs of the layperson, it still ties information to geographical location and provides important insight into the data at a high level. The stacked bar chart was the best choice for delving deeper into the subset of data associated with cancer running in the family but not meeting guideline criteria.

References

1. Guttmacher, A. E., Collins, F. F. S. & Carmona, R. H. The family history-more important than ever. *New England Journal of Medicine* **351**, 2333–2336 (2004).
2. Vogel, T. J., Stoops, K., Bennett, R. L., Miller, M. & Swisher, E. M. A self-administered family history questionnaire improves identification of women who warrant referral to genetic counseling for hereditary cancer risk. *Gynecol. Oncol.* **125**, 693–698 (2012).
3. Hampel, H. *et al.* A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment. *Genet. Med.* **17**, 70–87 (2015).
4. Daly, M. B. *et al.* Genetic/Familial High-Risk Assessment: Breast and Ovarian, Version 2.2015. *J. Natl. Compr. Canc. Netw.* **14**, 153–162 (2016).
5. Provenzale, D. *et al.* Genetic/Familial High-Risk Assessment: Colorectal Version 1.2016, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc. Netw.* **14**, 1010–1030 (2016).
6. Welch, B. M. *et al.* Using a Chatbot to Assess Hereditary Cancer Risk. *JCO Clin Cancer Inform* **4**, 787–793 (2020).
7. Roth, R. E. *et al.* User studies in cartography: opportunities for empirical research on interactive maps and visualizations. *International Journal of Cartography* **3**, 61–89 (2017).
8. Montello, D. R., Fabrikant, S. I. & Davies, C. Cognitive perspectives on cartography and other geographic information visualizations. in *Handbook of behavioral and cognitive geography* (Edward Elgar Publishing, 2018).
9. Kim, A. M. Critical cartography 2.0: From 'participatory mapping' to authored visualizations of power and people. *Landsc. Urban Plan.* **142**, 215–225 (2015).
10. Streit, M. & Gehlenborg, N. Bar charts and box plots. *Nat. Methods* **11**, 117 (2014).
11. Randle, T. What do you mean I'm not supposed to use Pie Charts?! https://www.geckoboard.com/blog/pie-charts/ (2015).
12. Few, S. & Edge, P. Save the pies for dessert. *Visual Business Intelligence Newsletter* 1–14 (2007).