Jared Stuber

IS 682

April 16, 2022

**Pattern discovery, validation, and online experiments: a methodology for discovering television shows for public health announcements: A Summary and Critique**

This study describes Public Health Announcements (PHA’s) and the necessity of their distribution to the public. These announcements are critical to the health and awareness of the public at large, yet the distribution of the announcements is often limited and sporadic in contrast to television programming and advertisements that promote behaviors and products counter to the PHA’s subject matter. The authors of this study utilized a database of U.S. television viewership, known as the Nielsen Database, and developed a novel data analysis methodology designed to discover and display television viewership patterns to determine optimal times, programs, and stations to present these PHA’s to the widest intended audience of them. For instance, determining which shows would be best utilized to feature anti-smoking PHA’s either within the program or as a commercial during a break in the designated show. This novel method was constructed using data from viewership pattern discovery of high dimensional data, non-parametric testing to validate findings, and data from experiments conducted on Facebook. The authors of the study claim that the idea to design such a process is based on, in part, a generic hypothesis that claims advertisements and product placements within television programs are associated with specific behaviors counter to the goals of the PHA’s, and that these associations are not accidental in their placement. (i.e., advertisers and television programmers are aware of the association and its benefit in revenue generation)

A three-step process was utilized and began with Nielsen’s database of television viewership across the U.S. The database was used to determine which television programs are associated with risky behaviors, such as binge drinking or smoking, and/or specific health conditions.

This data is determined as accurate at the household level, not at an individual level. Viewership was recorded from both broadcast and cable stations by the metric of total daily viewership, recorded in minutes. Furthermore, data at the household level is only available with prior household consent. This consent requirement further reduces available data for the model. The authors of the study note that some initial findings of the high dimensional data gathered from television viewership are more likely to correlate with the target variable of risky behavior or health condition data provided by the state by nothing more than pure chance.

It is noted that the popularity scores of television shows may accidentally match different health statistics in various states, and that correlation does not always equal causation.

According to the authors, there is only one case of risky behavior, smoking, in which there is a direct correlation, supported by sufficient evidence, in which on-screen smoking when viewed by young people, increases the risky behavior of smoking among that demographic.

It is also noted within the study that though a limited number of direct correlations between show viewership and adverse health conditions or risky behaviors occur, there are several factors within the model that have the potential to generate indirect correlations. These indirect correlations can result from many unknown, or latent variables and/or demographics unknown to the study’s authors. Furthermore, the authors suggest that these simple latent correlations may also prove beneficial in identifying different television shows for targeted PHAs. It is also important to note that the initial findings of the prescribed methodology generated correlations between show viewership and STDs, the authors chose to omit those findings during the experiment implementation phase of this study.

For the television show data, a total of 600 popular shows, as of 2016, across the nation were utilized to create a baseline. The rating of top 600 came from a compilation of popularity scores obtained from IMDB, TV Guide, Nielsen, and TV.com. The shows covered a wide range of categories such as drama, comedy, reality, sports, news, and politics to ensure the largest and most accurate dataset.

This data set is contrasted by the availability of public health statistics in a more focused format, at the aggregate level. To work with the differences in available data, the authors first determined the popularity at the aggregate level of the designated TV shows in each category. The so-called “popularity score” is determined by calculating the household-level show viewership by addressing a specific aspect of a show’s overall popularity within a given state and assigning a designated score when the same metrics are assigned to all the shows, and then placed in a ranked position. The overall computations resulted in a total of 2,400 popularity scores by state. This was determined by multiplying the original 600 shows by 4 state-level popularity scores, per state.

The model used was implemented on this first-level data by using a single split binary regression tree to locate and display patterns and correlations between overall health conditions and behaviors. During this selection, the regression model will select a given shows popularity score and its relation to a health condition by using one of the four metrics of popularity and utilizes whichever outcome proves most useful in showcasing the correlation between the given show and given health condition, by reducing and choosing the lowest root mean square error (RMSE). When completed, the model will ultimately update a total of 4,800 times throughout the initial process. (600 overall shows and 8 conditions) It should be noted, however, that the robustness of a show is not dependent on any arbitrary RMSE cutoff, individually, if the same cutoff is used in the original determination and in the cutoff version of calculating an overall score.

Once this first step was concluded, the second step utilized a series of nonparametric tests on the data gathered from the database and its correlations to risky behavior and/or health conditions. The nonparametric tests were designed to determine the strength and robustness of the initial correlation findings. The authors believed that if discovered data patterns resulted or appeared to be accidental, then those results should then appear across all fifty states (i.e., if a risky behavior such as binge drinking was identified, and the same process on the model was shuffled and returned, then a similar number of shows should be returned, so long as their RMSE’s end up falling below an arbitrary cutoff score level.) The authors claim that the potential for a show to be determined as informative is overall lower than what was originally produced before any shuffle occurred. They further conclude that these correlations are not then random, but rather suggest genuine correlations between given health conditions and specific show viewership popularities. Also, the overall model was run a total of 1,000 times per specific health condition resulting in an overall simulation process of 4,800,000 models for each show, all randomly shuffled to produce a standard benchmark of shows that have the potential to inform viewers of the risks of a given behavior or health condition. Findings from this approach show however that less than five total shows are informative, further increasing the likelihood that much of the original data on the shows gathered was beneficial and considered to be “non-accidental” and of ultimate value to the authors of the study and governmental agencies. Similarly, though, in shows in which drug overdoses were calculated regarding their correlation with specific television shows, several programs were discovered with very similar RMSEs, with most of the findings resulting in a skew to the right in their distribution, as the regression trees are only roots at this point in the model. These findings are determined to be equal to the standard deviation of a given cutoff rate and below any given, but arbitrary cutoff rate.

The third and final step in the data analysis came directly from Facebook’s own split testing platform, which was utilized to test the patterns discovered in the first two steps of the methodology. First, a PHA was presented in the form of a Facebook advertisement to a random group of perceived fans of a television program, compared to the delivery of the same or similar advertisement PHA to a selection of fans of a “control” show chosen at random, more responses to the PHA were expected from the targeted groups of bans based on their preference for a specific “test” show, as long as the health condition the authors were targeting was not in fact, accidental in the models’ findings. Specific PHA’s for every identified health condition and risk behavior were designed and distributed to a target audience. This approach allowed the authors to target specific audience sets, with no overlap, with the same PHA’s to determine their overall effectiveness. A follow-up link was integrated into each PHA to determine how different targeted and control groups responded to the messaging within them. For the implementation of the Facebook advertisements, only the top 10 shows were used for the targeted and controlled distribution of the PHA’s. To determine the top 10, the strongest association with a specific behavior and/or health condition and a show was utilized. For the control portion of this top 10 strategy, a random group of identified fabs was used to validate the findings. The initial experiments resulted in the design of a model that used 10 shows, with 5 specific conditions identified each, and that ran for the entirety of a week on Facebook to determine their effectiveness. It should be noted, however, that these testing platforms from Facebook were experimental themselves. Also of note is the fact that the awareness methods utilized and tested in this study were viewed by 1.5 million American adults. The authors of this study claim this methodology can be applied, repeatedly, so long as new data is introduced to the model. Furthermore, the authors suggest that findings from 2016 viewership data when entered into this model identified several television shows that were associated with, their popularity, health conditions, and specific risk behaviors. The definitions of risk behaviors are provided and are defined within the study as sexually transmitted diseases (STDs), smoking, binge drinking, obesity, diabetes, and drug overdoses. The data for such risk behaviors were provided to the study’s authors from state health agencies to provide reliable statistics at the state level for the year 2016.

The results of this experiment show that 34 of 50 total experiments produced higher clickthrough rates (CTRs) from fans of shows determined to be correlated with risk behaviors and/or health risks. In 10 of these experiments, there was a significant difference between the correlation between identified shows and their fans and those of the control groups. However, there were 11 shows with control audiences in which the experimental data showed higher CTRs. It should be noted that each of these findings is determined to have little to no statistical significance within the study. The findings also show a significantly less expensive delivery method of PHA’s to identified and targeted groups. Specifically, the cost of a single PHA was determined to be $0.61 cheaper on average when a PHA was shown to identify fans of a specific television program, with the most significant reduction in single-cost advertisements determined to be PHA’s aiming to reduce binge drinking, at a total reduction of $1.02 less expensive on average when the PHA is delivered to fans of specific programs determined to be of high significance by the model. The authors claim these findings will prove to be very beneficial to state and local agencies seeking ways to better target and distribute their PHA’s to a wider audience.

Some critiques of this study are mentioned by the authors, and have been expanded upon, as well as others not listed by the authors of the study are listed below.

First, the authors of this study state that a total of four variables were to be used in determining the overall decision tree to display correlations between behaviors and/or risks and the viewership of certain television programs. The model utilized to make this determination was essentially free to choose on its own the variable best suited for the parameters specified. With this setup in mind, it should be noted that in the event a single variable or show is removed, the entirety of the decision tree could change, affecting the entirety of the model.

Second, it should be noted that an arbitrary score was assigned to each health condition and risky behavior. Nowhere in the study is this number specified, and was this number the same throughout the study regardless of the variable, or was an entirely different number used for each condition, behavior, and score for the findings to fit within the model? Would changing the arbitrary score change the outcome of the study at any level of significance?

Third, and in line with the first critique, is the issue of the Facebook experimental tests. A listing of the top ten programs was chosen and tested experimentally as described in the study. As stated earlier, the algorithm utilized in this study chose the best-fitting variable in the prescribed parameters provided to it. With the removal or addition of a single variable, would the entire model be affected, therefore affecting the validity of the experiments conducted on Facebook, or at the very least, the control and test show utilized for the experiments?

Fourth, throughout the study, it is claimed that data was gathered throughout the U.S. down to the household level, and only with household consent. Utilizing this metric, it was prematurely determined that all viewership statistics related to specific television programming were accurate and represented adult viewership. A critique of this assumption would be there is no way in determining if the perceived audience of a show was the actual audience of that show. Was it an adult viewing the program, or was it a child watching adult programming? Are there multiple adults in the household, living communally as roommates all with different interests in television programming? Perhaps the television was simply on as background noise or to entertain an animal while the adults of the household were away. This variable is unknown and cannot be determined without further analysis, and therefore weakens the overall validity of the findings of the study.

Fifth, this study claims this data can be utilized by public health agencies for the foreseeable future but fails to accommodate for the fact that cable and broadcast television viewership is decreasing with the advent and wider acceptance of streaming services like Netflix and Hulu. Nowhere in this study is the variable or metric of this reality mentioned or included in the model for PHA purposes.

Sixth, the study claims that this model was designed to determine the best placement for PHA’s to be placed into television programming and claims this placement can be done at a relatively low cost. With that in mind, is this study claiming that these PHA’s be written directly into shows, or that they are aired as commercials during breaks? If the claim supports the latter, the study does not consider how often viewers will ignore commercials or mute their television and/or skip commercials if they are able. This has the potential to further limit the spread of these PHAs. If the authors of the study intend for these PHA’s to be written into specific shows, how can they ascertain or even estimate the financial cost of this suggestion? Will smaller states and counties not be included in these write-ins? Furthermore, what happens if a single show is identified with either risky behavior or health conditions of two different issues such as binge drinking and smoking and each of these issues affects different states in different ways?

Seventh, simply because these PHA’s reached 1.5 million Americans, is there any measure that demonstrates any level of benefit from airing such PHA’s? If not, then is this model beneficial for health agencies? (i.e., simply because a PHA is viewed, it does not necessarily mean that any reduction in risky behavior has occurred.)

The authors recognize many of the shortcomings of this study and openly discuss them in the discussion portion of their paper. The above critique is meant only to expand on those already acknowledged critiques, and to shine a light on additional and potential shortcomings of the study at large.

In conclusion, I found this study incredibly interesting, and it is very easy to see the overall benefit of such efforts. Though some of those efforts may lack in areas, the overall benefit realized through such a campaign is likely to outweigh any costs or drawbacks associated. It would be very beneficial for future iterations of this model to attempt to incorporate remediations for some of the critiques identified within this summary, though the product is still presenting an enormous benefit to local and state agencies. Of the critiques identified, I believe the most important to address would be the development of some metric to determine the actual benefits of viewing the PHA’s, regardless of the behavior or health condition the PHA represented. Should the authors of this study develop such a tool, health agencies could adjust their campaigns in real-time, lowering costs and reaching wider audiences to benefit a larger sector of society. Without this tool, agencies are limited to simply knowing how many people have viewed their PHAs, which is the main drawback of such a program. If this drawback can be addressed, many of the other critiques could be disregarded as the benefits to society will ultimately outweigh or eliminate any smaller critique.