MathWorks Math Modeling Challenge 2019

Academy for Science and Design–

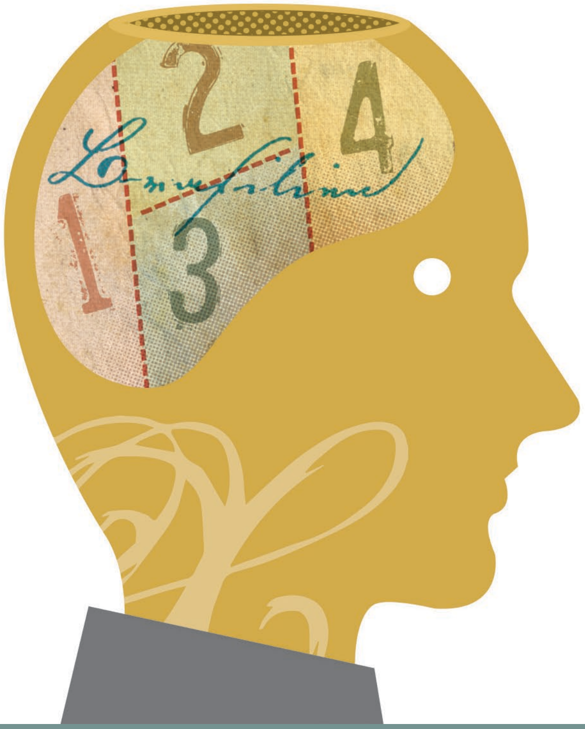
Team #11801 Nashua, New Hampshire

Coach: Karen Legault

Students: Denver Blake, Daniel Bujno, Ian Coolidge, Frederick Lee, Nathan Yeung

MathWorks Math Modeling Challenge Finalist

$5,000 Team Prize



Executive Summary

Substance use and abuse is a national problem that greatly affects the physical and mental health of users. The users of these substances can be as young as middle school students, and substance use can be detrimental to their health in the long-term. The financial and non-financial consequences of substance abuse negatively affect societal regulations and taxes on the national level to control or restrict the consumption of such substances.

To create a mathematical model that predicts the spread of nicotine due to vaping, we first considered the different types of models that could be used before choosing to model the situation with a logistic curve. After considering all the components of the logistic equation (carrying capacity, rate of growth, time), we used the MATLAB Curve Fitting feature to find the two parameters of the logistic model. We found that the spread of nicotine was logistic with a carrying capacity of 24.09%. We observed that the main difference between e-cigarette and cigarette usage growth was that e-cigarette usage grew logistically, while cigarette usage peaked and then declined.

To establish the risk factors associated with the likelihood of a given individual using a certain substance, we looked into various outside datasets to get risk factor values depending on income level and individual sex. Our final mathematical model utilized three main parameters: the average of the two risk factor scores, the number of addicted friends, and the inherent ease of influence of the substance. With this mathematical model, we developed a simulation that dynamically models each student's potential to become addicted to a given drug and used this simulation to make predictions. The simulation discovered that in a typical high school of 300 students, 128 students will try nicotine, 179 students will try alcohol, 105 students will try marijuana, and 22 students will try opioids. It is important to keep in mind that a single student can use many different type of substances.

Our team decided to use the value of a Substance Damage (SD) score from 0 to 100 to rank the relative impact of each substance. The SD score is the sum of the Health Damage score, determined by life expectancy reduction, and the Financial Damage score, calculated by dividing the cost of attaining the drug by income. The final SD scores show alcohol as being most dangerous, with cigarettes and marijuana following in danger level, opioids following as fairly dangerous, and vaping being least dangerous of all.

All three of the questions stressed the idea of the growth of drug use in schools, specifically the substances of nicotine, marijuana, cigarettes, alcohol, vaping, and opioids. The solutions to this challenge showed that the growth of drug use in schools will continue to grow as long as students are present to influence their friends. This frightening idea shows that stricter regulations and actions must be taken at the local, state, and national levels to stop the widespread substance abuse problem.

Global Assumptions

1. The government will not create additional regulations or loosen regulations over drugs in the future of our models.
    a. *Justification:* Such regulations would have a large impact on drug usage. For our models to be consistent, we cannot take into account changes in regulation.

Part 1: The Vape Awakens

1.1 Restatement of Problem

We are asked to create a model that predicts the spread of nicotine use due to vaping over the next 10 years as well as compare its growth to that of cigarettes.

1.2 Local Assumptions

1. Only youth and young adults will be considered for this model.
    a. *Justification:* The usage of e-cigarettes for the older population is largely correlated to other tobacco use; they are often used to replace traditional cigarettes. Thus, to consider the spread of e-cigarettes, we can focus on the younger population.
2. The problem statement asks us to predict the spread of nicotine use due to vaping over the next 10 years. We interpret "the next 10 years" as the time from 2019 to 2029.

1.3 Variables

List of Variables for Logistic Model

Symbol	**Definition**	**Units**
K	Carrying Capacity	Percent
P_0	Initial Population	Percent
r	Rate of Growth	1/Years (Frequency)
t	Time (Starting from 2011)	Years

Caption: This table lists all the variables and parameters present in the general form of a logistic model.

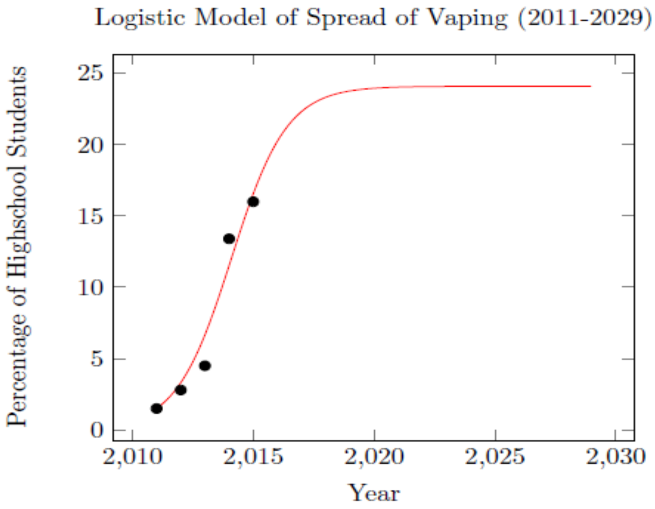
1.4 Results

The first part of the problem asks us to mathematically model the spread of nicotine use due to vaping for the next 10 years. Our team believed that the best model for this spread would be a logistic model due to its two main characteristics: exponential growth and a limit at some fixed capacity. One can imagine that, like bacteria growth in a petri dish, the use of nicotine through vaping will grow exponentially. Once one student/person decides to start vaping, their use will influence their social circle or surrounding people to start vaping as well. As more students/people vape from that initial individual, the rate at which people start vaping will grow and the number of people who vape will increase exponentially. However, just like many real-world exponential growth situations, the number of people who vape will not continue to grow indefinitely. The number of people who vape must reach a carrying capacity once resources become limited. The logistic model is perfect for our team as it takes into consideration these two factors. The general form for a logistic equation is

$$P(t) = \frac{K\,P_0}{P_0 + (K - P_0)\,e^{-r(t-t_0)}}$$

with variable names and definitions as indicated above. Using the MATLAB Curve Fitting tool, we fitted a logistic curve of this form to data on e-cigarette usage in high schoolers from 2011 to 2015, thus finding the following results. As one can see, the MATLAB Curve Fitting found the carrying capacity of the data to be 24.09% of high school students with an annual growth rate of 0.875.

$$f(t) = \frac{1.5 \cdot 24.09}{1.5 + (24.09 - 1.5)\,e^{-0.875(t-2011)}}$$



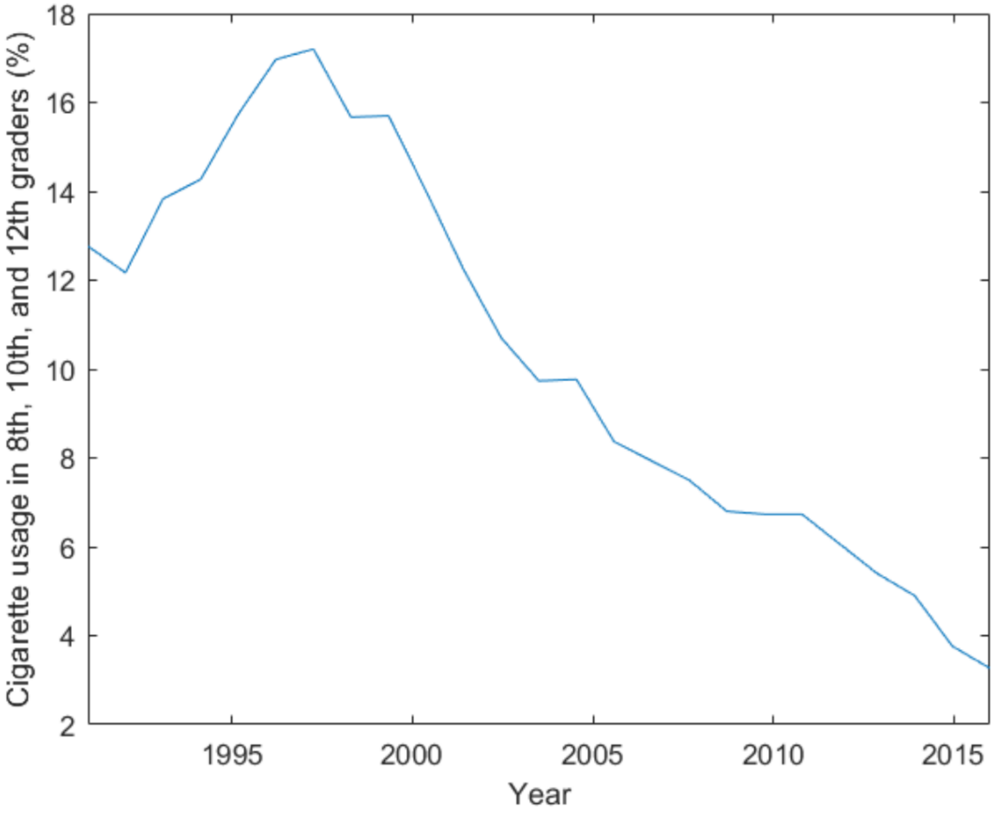
Caption: This graph displays the logistic model of nicotine through vape use growth. As indicated by the graph, the usage seems to increase exponentially before reaching a carrying capacity of 24.09.

1.5 Comparison to Cigarette Growth

The second part of the question asks us to analyze and compare the growth of vapes to the growth of cigarettes. Shown below is a graph of percent cigarette usage for 8th, 10th, and 12th graders created in MATLAB from data provided by the U.S. Department of Health and Human Services[2].

Looking at the percent cigarette usage graph, we can see that during the period of 1991 to 1999, the percent cigarette usage grew almost linearly with different slopes to a peak of almost 17% before it started to decrease substantially. Comparing this to the graph of percent of high school students using vapes, the logistic model had the percent vape usage growing exponentially during the period of 2011 to 2018 before reaching its carrying capacity of 24.09%. The main difference between the growth of e-cigarettes and cigarettes is that e-cigarettes grew exponentially, while cigarettes grew linearly with different slopes with time. Another difference is that cigarette usage began to decrease after 1999, but the e-cigarette model levels off to a carrying capacity instead of decreasing in usage.

Cigarette Usage of Students from 1991-2016



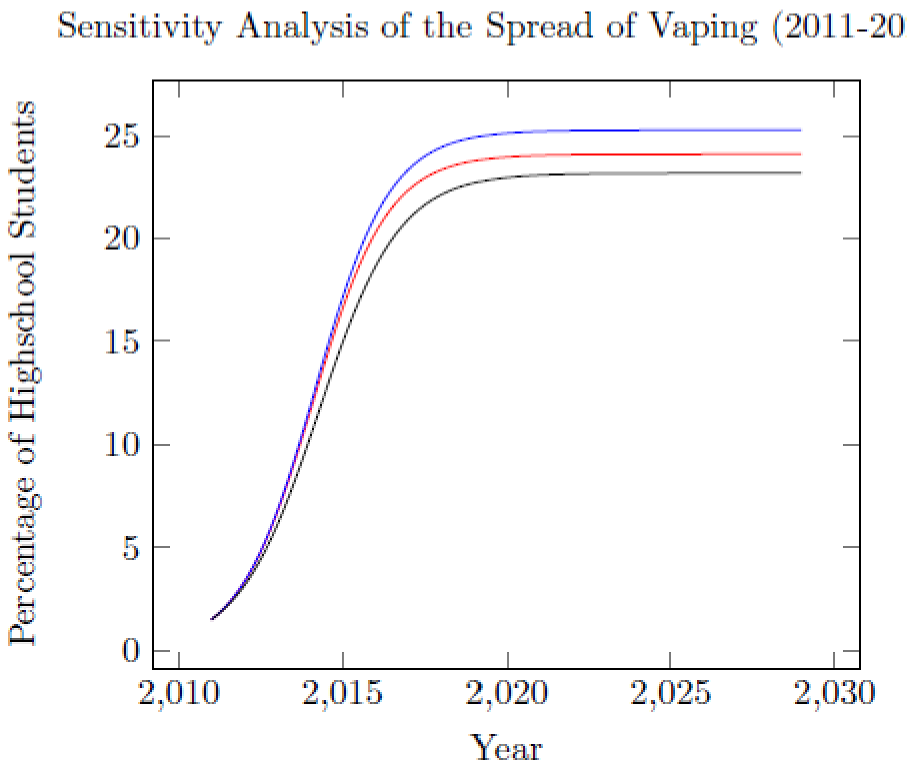
Caption: The graph shows cigarette usage percentages for 8th, 10th, and 12th graders over a long period of time. One can see that cigarette usage increases linearly with differing slopes and then starts decreasing after reaching a maximum.

1.6 Validation

To validate our model, we compared the percentage of high school students who vaped in 2018 given by the model with the real percentage of high school students provided by the Centers for Disease Control and Prevention. The Centers for Disease Control and Prevention found that the rate of high school students using e-cigarettes was at an all-time high at about 20.8% in 2018[6]. This is in line with our result of about 23.32% from the model given t = 2018. This percentage is quite accurate with a relatively small population but will become less valid with a greater population.

Our RMSE of the model has a value of 1.71%, which shows that our model is off by about that much on average for interpolated points, as well as extrapolated points if we assume that our model is sufficient.

We also performed sensitivity analysis on our logistic model on nicotine spread among high school students. By using the same data points but varying the percentage of students vaping by increasing by 10% and decreasing by 10%, as shown by the blue and black function below, we compared the result to the original logistic model as red. As shown below, the model is robust in the presence of uncertainty.



1.7 Strengths and Weaknesses

Strengths

- Simplicity
  - A strength of this model is that it is purely simple in its function form. The final model has one input which is time since 2011 and has one simple output that gives the percentage of high school students who use nicotine through vapes.

Weaknesses

- Accountability of Events
  - We cannot account for events that may alter percent vape use by high school students. For instance, if the U.S. government were to ban all use in 2020, our model would not be able to account for this event and rapid decrease in vape use.
- Low Sample Size
  - There was an extreme lack of data related to vaping due to its recent rise and lack of data collected further than the yearly level. The modeling and logistic function could be more accurate with a larger sample size.

Part 2: Above or Below the Influence?

2.1 Restatement of Problem

We created a model that simulates the likelihood that a given individual will use a given substance by taking into account social influence, characteristic traits, and the drug itself. We then demonstrated the effectiveness of our model by predicting the number of students among a class of 300 who will use nicotine, marijuana, alcohol, and un-prescribed opioids.

2.2 Local Assumptions

1. Any student can be part of any social circle with equal probability.
    a. *Justification:* It is unnecessary to account for the idea of similar people assimilating into their own cliques. Furthermore, our simulation will run many times to ensure that all different clique sizes and types of people in cliques will be accounted for.
2. Cliques have approximately 6 people, and the number of relationships in a school can be calculated based on this number.
    a. *Justification:* The PMC estimates that the average clique size is 5.72 with a range of 3 to 18.[5] This rounds up to 6.
3. Race is an insignificant factor compared to others in relation to drug usage.
    a. *Justification:* Race is strongly correlated with income, so it is not independent from other factors. In addition, race has not been proven to be a significant factor in drug usage, and since races are fairly diverse among themselves, we believe it is safe to ignore race for this situation.
4. Income level is directly correlated with quality of living conditions, including factors such as neighborhood, number and quality of parents, and child abuse.
    a. *Justification:* Discounting outliers, higher income generally corresponds to better stability as well as better access to goods, services, and location; more money almost always means more freedom.
5. Usage of one substance will not in generally affect usage of another substance.
    a. *Justification:* We were unable to find strong data discussing the correlation of usage for each pair of substances.
6. Overdose mortality is ignored by our model.
    a. *Justification:* Overdose mortality is rare enough and has only a localized effect at most, so this factor is insignificant.

2.3 Variables and Constants

List of Variables and Constants for Model of Pr(Y)

Symbol	**Definition**	**Units**
S	Drug Use Relative to Sex (Male/Female)	Percent Probability
I	Drug Use Relative to Income Level (≤20K, 21K≤40K, ≥41K)	Percent Probability
N	Number of Friends	Friends
n	Number of Addicted Friends	Friends
Pr(Y)	Total Probability to Start Substance Abuse	Percent Probability
X	Ease of Influence of Each Substance	Unitless Constant
G	Percentage of 10th Graders Using Drugs	Percent

Caption: These are the variables needed and used for the mathematical model that predicts the number of people using an illicit substance over the course of a school year.

2.4 Mathematical Model

In our model, we consider four main factors or metrics to significantly affect the likelihood that a given individual will use a certain substance: sex, income level, and number of friends. Using data from various sources indicated in our footnotes, we can construct this table of factors that, when manipulated according to a specific function, can be used to find the likelihood that a given individual will use a specific substance. We found the drug use percentage increase per year by taking the difference between the drug use percentage, and dividing it by the average change in time, which was 3 years. This was calculated for both drugs and all three income brackets for which data was available.

List of Risk Factor Scores

	Alcohol	**Nicotine**	**Marijuana**	**Opioid**
G[1]	37.8%	32.3%	27.5%	6.9%
Income Level[8]				
≤20K	6.933%	3.433%	*	*
21K≤40K	9.567%	4.833%	*	*
≥41K	10.667%	5.3%	*	*

Sex[4]				
Male	14.9%	6.2%	5.8%	3.3%
Female	17.0%	7.2%	6.2%	4.1%

Caption: These are the risk factor scores compiled from various data sources as well as computed if raw values were given instead of a rate. Some values are missing due to the fact that there was no data found for this information.
*No data available.

Our team believed that the first step in finding a correct model for the probability of a given individual using a substance was to find all the important factors, which are sex and income level. We combined the features by finding the mean of the factors to find the probability by only considering sex and income level.

$$Pr\left(Y\right) = \frac{S+I}{2}$$

Nonetheless, this equation fails to consider the idea of social influence from multiple friends to try a certain substance. To account for this, our team decided to multiply the above equation by *n* or the number of addicted friends. If an individual has more friends that use a substance, they are more likely to be influenced to also try and use that specific substance.

$$Pr\left(Y\right) = n\frac{S+I}{2}$$

This equation is stronger but also fails to consider the differing levels of ease for starting to use the different types of substances. For instance, we must consider that it is easier to start drinking than to start doing drugs. The rate of influence of each substance, X, is calculated by using the data from[3] that said 55% of alcohol users were influenced by peers and 70% of smokers were influenced by peers. We assume that the alcohol influence is similar to smoking at 70%. Using the percentage represented by P multiplied by G, the total percentage of averaged students affected by the substances, we get the percent of people who use the substance due to peers. That is divided by the percent of people who didn’t use a substance without outside influence to get the ease of influence of a substance.

$$X = \frac{G \cdot P}{1-(G\cdot(1-P))} = \frac{\text{Percent of People Who Use a Substance Due to Peers}}{\text{Percent of People Who Didn't Use a Substance Without Outside Influence}}$$

Finally, the last equation is then divided by 12 to get the monthly probability of using the substance. The final probability model equation is shown below.

$$Pr\left(Y\right)=\frac{1}{12}Xn\left(\frac{S+I}{2}\right)$$

2.5 Simulation and Results

In order to accurately model the time-dependent nature of this problem, we developed a simulation that dynamically models each student's potential to become addicted to a given drug on a monthly basis. Our team felt that it was crucial to develop a dynamic simulation due to the ever-changing nature of social relationships. Due to ease of use and prior programming knowledge, we chose to develop this simulation using Python and NumPy. In order to organize our code and facilitate testing, we used a Jupyter notebook.

The purpose of the simulation code is to compute an expected final number of students who would be exposed to an illicit substance by the end of senior year. The code provides multiple useful features, including the ability to randomly generate metrics for students in a senior class and randomly generate a given number of relationships between these students. Additional functions allow the user to run individual or bulk simulations of a given high school's senior year, giving results that can be validated against real-life data.

Perhaps the code's most useful feature is its modularity, which allows the programmer to enter new substances or metrics with ease as long as they are given data. This made it significantly simpler for our team to work with various factors and their relationship to various substances. The results of the code were highly accurate against NIH data, as is shown below in the Validity section.

The table below shows the average of 5 runs for the model for a class of 300 high school seniors with varying characteristics with the following substances: nicotine, marijuana, alcohol, and nonprescription opioids.

Simulation Results

Substance	Model	Percent
Nicotine (Vaping)	128.4	42.8%
Alcohol	178.8	59.6%
Marijuana	104.7	34.9%
Opioid	22.2	7.4%

Caption: Simulation results after averaging the results of 5 runs for a class of 300 high school students with varying characteristics. The total is found by multiplying the simulation percentages by the 300 students.

2.5.1 NumPy Builtins

While writing our code, we made use of various functions within the NumPy library. These functions, including numpy.random.choice and numpy.mean, were used to speed up our code and make matrix and vector processing more straightforward.

2.6 Validity

To validate our model, we compared the average percentage of seniors who used the substances we found to the NIH average percentage of seniors who used the substance. The table below shows the average of 5 runs for the model for a class of 300 high school seniors with varying characteristics with the substances nicotine, marijuana, alcohol, and nonprescription opioids, and also shows the NIH average and percent error.

Percent Error between NIH Data and Model Data

	Model	NIH Data[17]	Percent Error
Nicotine (Vaping)	42.8%	37.3%	14.7%
Alcohol	59.6%	53.3%	11.8%
Marijuana	34.9%	35.9%	2.8%
Opioid	7.4%	7.8%	5.1%

Caption: This is a table listing the NIH compiled data versus the data and percentages that our simulations found. The model percentages are the average values of 5 simulation runs for a class of 300 high school seniors with varying traits. The data also lists the percent error between our model data and the NIH true data.

The results of our problem are similar to the NIH data with the percent error ranging from 2.8% to 14.7%, but a little off due to limited time. There are other factors involved in an individual using a substance that will require more investigation that could account for the error.

2.7 Strengths and Weaknesses

Strengths

- Dynamic Evolution
    - Our model took into account evolution of time, rather than providing a static prediction based on the given variables. By using a simulation that took into account changing social circles monthly, our model captured the essence of human relationships and friendships as these relationships relate to the spread of drug use.
- Modularity
    - Our model was highly modular; if additional data were presented, we would be able to adapt the model to utilize the additional data with relative ease.

Weaknesses

- Correlation of Variables
    - Significant correlations existed between the model parameters which may lead to overcounting of certain features. For example, the data suggesting that 6.2% of males and 7.2% of females are susceptible to nicotine abuse may have a correlation with other variables, despite our model not decoupling them. This lack of decoupling could ultimately lead to a slightly less accurate model.
- Lack of Data
    - Due to an inability to find data about certain metrics, we were unable to utilize as many factors in our model as we would like. Factors such as advertising exposure and genetics may play a role in drug use which this model would not account for.

Part 3: Ripples

3.1 Restatement of Problem

We are asked to model the impact of substance abuse, both financially and non-financially. We are then asked to rank the substances from problem #2.

3.2 Local Assumptions

1. Death is the most costly outcome for any individual.
2. Damage to health costs more than the hospital bill.
    a. *Justification:* Consistent abuse of drugs can deteriorate health beyond the point that can be fixed by a doctor.
3. The only one who loses from abusing substances is the addict.
    a. *Justification:* Sellers and hospitals only gain money from serving addicts. The opportunity cost of additional possible labor for the community is outweighed by the amount of labor already available relative to the quantity of jobs open.
4. The money spent on drugs and hospital trips is less than the income of the addict.
    a. *Justification:* Spending more than one's own income is a route that invariably leads to the inability to legitimately pay for all of their expenses, eventually leading to either bankruptcy or crime. This extremely bad yet unpredictable outcome is difficult to measure, but since it is known to be terrible anyway we omit this possibility from our considerations.
5. Smoking marijuana lowers your life expectancy by the same amount as smoking cigarettes.
    a. *Justification:* There is a lack of data on smoking marijuana. In addition, smoking universally hurts lungs.

3.3 Variables and Constants

List of Variables and Constants for Model of SD

Symbol	**Definition**	**Units**
SD	Substance Damage	Score (Metric Between 0 and 100)
I	Income	Dollars
C_D	Cost of Drugs	Dollars
FD	Financial Damage	Score (Metric Between 0 and 100)

E_N	Normal Life Expectancy	Years
E_D	Affected Life Expectancy	Years
HD	Health Damage	Score (Metric Between 0 and 100)

Caption: These are the variables needed and used for the mathematical model that calculates the Substance Damage from one particular drug based on income cost of drugs, and life expectancy with and without the use of drugs.

3.3 Analysis and Validation

To consider the drug's financial impact, one can find the annual cost of the drug relative to the income of the addict.

In addition, there are a couple of non-financial factors:

1. Mental and physical health
2. Death from suicide or overdose

Both of these factors are considered together in terms of life expectancy.

The model will have an output on a scale of 0 to 100, where 0 is no cost while 100 is the worst possible outcome of death. This output will be called the Substance Damage score (SD).

One component of the SD is the financial input. The total drug cost (C_D) is the sum of the annual drug cost and the annual expected medical expenses resulting directly from drug abuse. This will be divided by annual income (I) and multiplied by 100 to obtain the Financial Damage (FD) score.

$$FD = 100\left(\frac{C_D}{I}\right)$$

This method of calculating FD makes sense. If two drug addicts have the same money left annually after drug cost but one has a higher income than the other, the one with the higher income will have the same amount of money to pay higher expenses.

The other factor in the SD is the Health Damage (HD), which is a function of normal remaining life expectancy (E_N) and life expectancy remaining under the influence of drugs (E_D). This damage has an upper bound of the expression (100 - FD).

$$HD = (100 - FD)\left(\frac{1 - E_D}{E_N}\right)$$

Combining the two is simple:

$$SD = FD + HD$$

The edge case where SD = 100 can be intuitively explained as either the total cost of drugs equaling the income (after all, death costs everything) or the remaining life expectancy equaling 0, either way implying the worst outcome of death.

This model prioritizes financial cost over health cost, with the idea that without being able to pay for your other expenses, worrying about your health will become a luxury.

3.4 Results

The results are calculated through the above formulas and data obtained online, considering smoked marijuana to reduce life expectancy by the same amount as cigarettes due to lack of data and the fact that smoking universally hurts lungs. The results assume that the person is 20 years old, has an annual income of $20,000, and has a total life expectancy of 76.7 as someone born in 1999[7].

Damage Scores for Different Substances

Drug Type/Amount	**Life Expectancy Reduction (1 - E_D)**	**Annual Cost (C_D)**	**Financial Damage Score (FD)**	**Health Damage Score (HD)**	**Substance Damage Score (SD)**
Alcohol	26 years[13]	$4500[10]	22.5	35.53792	58.03792
Cigarettes (3 packs per day)	10 years[14]	$7224[12]	36.12	11.26631	47.38631
Vaping (3 packs per week)	9.5 years[15]	$696[11]	3.48	16.17178	19.65178
Prescription Opioids	14.64 years[16]	$3654[10]	18.27	21.10277	39.37277
Smoked Marijuana (1 cigar per day)	10 years[14]	$7000[10]	35	11.46384	46.46384

Caption: Table that indicates the values of life expectancy reduction, annual cost, Financial Damage score, Health Damage score, and final Substance Damage score for all the different types of substances.

The data here shows a proportionality between annual cost and FD score, as well as scaling of HD score—although vaping has the lowest life expectancy reduction, it still has an HD score significantly greater than that of smoking due to vaping being financially cheap. The final SD scores show alcohol as being the most impactful, with cigarettes and marijuana following in impact level, opioids following as fairly impactful, and vaping being least impactful of all.

3.5 Strengths and Weaknesses

Strengths

- Simplicity and Accuracy
    - This model has the strength that it is simple and transparent while giving a good idea of how dangerous drugs are for you both financially and for your health. By scaling the health factor based on how affordable the drug is, this model focuses on what affects them more.
- Flexibility
    - This model can be adapted to any person based solely on income, age, drug cost, normal life expectancy, and decrease in life expectancy based on drug use. As long as one can find or approximate the drug statistics, one can easily calculate how much of a risk they face.

Weaknesses

- Lack of Accounting for Medical Expenses
    - Medical expenses due to drug abuse can be difficult to calculate. Still, such incidents can take a heavy toll on one’s finances. If we had more time, we would have accounted for these expenses as well for the Financial Damage score.
- Lack of Score Partitioning
    - There are no suggested bounds for high, medium, and low risk levels relative to the type of drug based on this metric. Coming up with these bounds is not easy; the fact that vaping takes away almost 10 years of life but has a score of only around 20 shows that this model might not have the best scaling. This is another issue that we could have covered if more time had been available.

Bibliography

[1] https://m3challenge.siam.org/node/439

[2] https://www.hhs.gov/ash/oah/adolescent-development/substance-use/drugs/tobacco/trends/index.html

[3] https://cf.ltkcdn.net/teens/files/2251-Statistics-on-Peer-Pressure-infographic-v2a.pdf

[4] https://www.sciencedirect.com/science/article/pii/S0376871613003013

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636991/

[6] https://www.cnbc.com/2018/10/22/teen-cigarette-smoking-ticks-up-as-vaping-surges.html

[7] https://www.infoplease.com/life-expectancy-birth-race-and-sex-1930-2010

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1446419/pdf/11111260.pdf

[9] https://e-cigarettes.surgeongeneral.gov/documents/2016_sgr_full_report_non-508.pdf/

[10] https://www.rehabspot.com/treatment/paying-for-rehab/cost-of-addiction/

[11] https://www.electrictobacconist.com/blog/2018/06/how-much-does-it-cost-to-vape/

[12] http://hd.ingham.org/Portals/HD/Home/Documents/eh/Tobacco/Tobacco%20and%20You/Sessions/Group%20I/Chapter%206%20How%20much%20does%20smoking%20cost/chap6_handout.pdf

[13] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402015/

[14] https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/tobacco_related_mortality/index.htm

[15] https://www.bmj.com/company/newsroom/study-estimates-years-of-life-that-could-be-saved-in-us-if-smokers-switched-to-e-cigarettes/

[16] https://www.ncbi.nlm.nih.gov/pubmed/17088224

[17] https://www.drugabuse.gov/drugs-abuse/opioids

Appendix

Note: Horizontal lines in Model 2.5 code indicate separate cells evaluated in the Jupyter notebook. All code in Model 1.5 is written for MATLAB R2018b, and all code in Model 2.5 is written for Python 3.x using NumPy.

Model 1.5 Cigarette Usage

```
years = linspace(1991, 2016, 25);

cigarette_8th = [7.2 7.0 8.3 8.8 9.3 10.4 9.0 8.8 8.1 7.4 5.5 5.1 4.5 4.4 4.0
4.0 3.0 3.1 2.7 2.9 2.4 1.9 1.8 1.4 1.3];
cigarette_10th = [12.6 12.3 14.2 14.6 16.3 18.3 18.0 15.8 15.9 14.0 12.2 10.1
8.9 8.3 7.5 7.6 7.2 5.9 6.3 6.6 5.5 5.0 4.4 3.2 3.0];
cigarette_12th = [18.5 17.2 19.0 19.4 21.6 22.2 24.6 22.4 23.1 20.6 19.0 16.9
15.8 16.6 13.6 12.2 12.3 11.4 11.2 10.7 10.3 9.3 8.5 6.7 5.5];

cigarette_all = mean(cat(1, cigarette_8th, cigarette_10th, cigarette_12th),
1);

plot(years, cigarette_all);
xlabel('Year');
ylabel('Cigarette usage in 8th, 10th, and 12th graders (%)');
xlim([1991 2016]);
```

Model 2.5 Simulation

```
import numpy as np
```

```
#This cell contains all data for each drug.
#Additional metrics and drugs can be added in a modular fashion without
changing the internal code.

#List of possible chances for an individual to become addicted to tobacco per
month
tobacco_metrics = {"gender": [0.062 / 12, 0.072 / 12],
                   "income": [0.034 / 12, 0.048 / 12, 0.053 / 12]
                   }

#Chance for an individual to be addicted to tobacco entering junior year
tobacco_chance = 0.323

#Rate of influence for tobacco
tobacco_inf = 0.298
```

```
#These statistics are repeated for the other 3 substances.
alcohol_metrics = {"gender": [0.149 / 12, 0.170 / 12],
                   "income": [0.0693 / 12, 0.09567 / 12, 0.10667 / 12]
                   }
alcohol_chance = 0.378
alcohol_inf = 0.250

marijuana_metrics = {"gender": [0.058 / 12, 0.062 / 12]}
marijuana_chance = 0.275
marijuana_inf = 0.173

opioid_metrics = {"gender": [0.033 / 12, 0.041 / 12]}
opioid_chance = 0.069
opioid_inf = 0.039
```

```
#Return the mean of each probability in a list
def padd(l):
    return np.mean(l)
```

```
#Return metrics for a randomly generated high school senior class
#Parameters:
#    N: size of the class
#    metrics: Metrics to be initialized for each student
#
#Returns:
#    hs: Array of metrics for each student in the class
def generate_high_school(N, metrics):
    hs = np.ndarray((N, len(metrics)))
    for i, metric in enumerate(list(metrics.keys())):
        hs[:,i] = np.random.choice(metrics[metric], size=(N,))
    return hs.T

#Average all metrics in a high school, returning a single influence score for
each student.
#Parameters:
#    hs: High school to be averaged. Obtained by calling generate_high_school
#
#Returns:
#    total: Array of influence scores for each student
def add_metrics(hs):
    N = hs.shape[1]
    total = np.ndarray((N,))
    for i in range(N):
        total[i] = padd(hs[:,i])
    return total

#Generate a matrix of randomized relationships between students in a class.
#Parameters:
#    hs: High school to generate matrix for. Obtained by calling
generate_high_school
#
#Returns:
#    R: NxN symmetric matrix in which a 1 at (i,j) represents a friendship
between two students.
def generate_relationship_matrix(hs):
    N = hs.shape[1]
    R = np.zeros((N, N))

    # The total number of connections, (N / 6) ** 2, is calculated due to the
fact that the size of an average friend group is 5.7, which we rounded up to
6.
    for k in range(int(N / 6) ** 2):
        i = np.random.randint(0, N)
        j = np.random.randint(0, N)

        if R[i][j] == 0:
            R[i][j] = 1
            R[j][i] = 1

    return R
```

```
#Simulate a single set of months during which drug use can spread within a
high school
#Parameters:
#    n_months: number of months over which to run the simulation
#    hs: High school on which to run simulation. Obtained by calling
generate_high_school
#    R: Relationship matrix for the high school. Obtained by calling
generate_relationship_matrix.
#    init_chance: Chance for an individual student to be exposed to the drug
at the beginning of the simulation.
#    influence_rate: Influence rate of an individual student's ability to
influence another student.
#
#Returns:
#    infected: Array of size (N,) in which a 1 represents a student exposed
to the drug.
def simulate(n_months, hs, R, init_chance, influence_rate, debug=False):
    N = hs.shape[1]

    new_R = np.copy(R)
    all_metrics = add_metrics(hs)

    infected = np.zeros((N,))
    for i in range(N):
        if np.random.rand() < init_chance:
            infected[i] = 1

    for month in range(n_months):
        if debug:
            print("Month: " + str(month+1))
            print("Number infected: " + str(np.sum(infected)))

        next_infected = np.zeros((N,))

        for i in range(N):
            if infected[i]:
                p = np.sum(R[:,i] * infected) / np.sum(R[:,i])
                for j in range(N):
                    if R[i][j] == 1 and np.random.rand() < all_metrics[i] * 
inf:
                        next_infected[j] = 1

        infected = np.logical_or(infected, next_infected)

    return infected
```

```
#Run multiple simulations on a single high school to find the average number 
of influenced students.
#Parameters:
#    num_sims: Number of times to run the simulation
#    n_months: Number of months over which to simulate
#    metrics: Metrics to be initialized for each student
#    init_chance: Chance for an individual student to be exposed to the drug 
at the beginning of the simulation.
#    influence_rate: Influence rate of an individual student's ability to 
influence another student.
#
#Returns:
#    total: Average number of influenced students across each simulation.
def run_sims(num_sims, n_months, metrics, init_chance, influence_rate, 
debug=False):
    totals = []
    hs = generate_high_school(300, metrics)
    R = generate_relationship_matrix(hs)

    for _ in range(num_sims):
        sim = simulate(n_months, hs, R, init_chance, inf, debug=debug)
        totals.append(np.sum(sim))

    total = np.mean(totals)
    return total
```

```
#Example cell to run a single simulation for marijuana
hs = generate_high_school(300, marijuana_metrics)
R = generate_relationship_matrix(hs)
simulate(24, hs, R, marijuana_chance, marijuana_inf, debug=True)
```

```
#Cell to run simulations for tobacco
run_sims(5, 24, tobacco_metrics, tobacco_chance, tobacco_inf)
```

```
#Cell to run simulations for alcohol
run_sims(5, 24, alcohol_metrics, alcohol_chance, alcohol_inf)
```

```
#Cell to run simulations for marijuana
run_sims(5, 24, marijuana_metrics, marijuana_chance, marijuana_inf)
```

```
#Cell to run simulations for opioids
run_sims(5, 24, opioid_metrics, opioid_chance, opioid_inf)
```