Investigating the Factors Influencing the Mental Health Scores of New York City Residents

1. Introduction

Mental illnesses stand as pervasive health concerns in the United States, affecting a substantial portion of the population. According to the World Health Organization, over 20% of adults contend with mental illnesses [1]. The Oxford Martin School further highlights the alarming prevalence, estimating that 1 in 3 women and 1 in 5 men will experience major depression at some point in their lives [2]. Adolescents, aged 13 to 18, also grapple with significantly impairing mental conditions, with over one-fifth affected [3]. The profound impact of poor mental health extends across emotional, social, occupational, and physical dimensions. Left unaddressed, these challenges can exacerbate existing conditions or contribute to the development of additional issues

Societal perspectives have undergone a significant shift, witnessing a heightened recognition and acceptance of mental health issues. In today's society, there is a greater openness to acknowledging conditions like depression and anxiety, leading to increased engagement with healthcare professionals and a proactive pursuit of suitable treatment [4]. As the global spotlight on mental health intensifies, understanding the physical and behavioral factors influencing an individual's mental well-being becomes paramount. This study seeks to unravel prevalent challenges and potential determinants shaping the mental health landscape of New York City residents.

1.1. Data Overview

The 2020 New York City Community Health Survey dataset (NYCHS) stands out for its comprehensive coverage of diverse health and behavioral characteristics, incorporating data from a wide spectrum of survey participants. With a total of 134 columns (excluding meta-data columns) and 8781 data entries, the NYCHS offers a rich exploration of health factors. Notably, the city of New York's remarkable diversity in ethnicities, nationalities, and socioeconomic statuses significantly contributes to the variance. This cross-sectional telephone survey involves approximately 10,000 randomly selected adults aged 18 and older from all five boroughs of New York City. All data collected is self-reported, offering a valuable snapshot of health-related trends in this dynamic urban landscape.

1.2. Research Objective

The objective of this study is to investigate the primary predictors influencing the mental health of respondents who participated in this survey and predicting a respondent's mental health class based on these predictors.

2. Method

2.1. Data Preparation and Cleaning

A common shortcoming in long surveys like the NYCHS (which consists of 133 questions) is the high number of missing values. The range of missing values per column variable within the dataset ranges from as low as 0%, to as high as 80%. As it is a common practice in data analysis to remove columns with a high proportion of missing values, column variables that consisted of more than 5% of missing data were dropped from the analysis. In addition, metadata columns that included irrelevant data, such as unique identifier numbers, stratification variables and references to previous surveys, were deleted.

Post-deletion, column variables that consisted of less than 5% of missing values remained. Given that only 10 column variables had no missing values, the decision was made not to eliminate columns with minor missing values. Deleting such columns would overlook historically proven characteristics, acknowledged by qualified therapists/psychiatrists to be directly linked to mental health. To address this, missing values in column variables with less than 5% were imputed, utilizing the column median for continuous variables (to avoid non-integer outcomes from using the mean) and the column mode for categorical columns. Of the overall 134 predictor column variables initially in the NYCHS, only 78 remained after the preparation and cleaning process.

2.2. Measuring Mental Health and The Kessler-6 Index Score

The "Kessler-6 index score" (K6) variable within the NYCHS originates from the Kessler Psychological Distress Scale, a widely recognized tool for quantifying non-specific psychological distress. Designed to gauge the emotional states of individuals, the K6 comprises six questions addressing aspects such as feelings of nervousness, hopelessness, and restlessness—core indicators of psychological distress. Respondents use a 5-point scale, with 0 denoting "none of the time" and 5 indicating "all of the time," to express the intensity of their experiences. The cumulative responses to these questions contribute to the computation of the K6 value. Importantly, a lower K6 score is indicative of poorer mental health and conversely, a higher K6 score suggests enhanced mental well-being.

Categorical variables are often more interpretable than continuous ones. For instance, understanding that a person falls into the “Very Good”, or one of the other class categories is more intuitive than interpreting a specific continuous score. In addition, binning in general allows us to leverage several decision making and classification based machine learning algorithms. The subsequent phase of our analysis entailed the creation of subclasses for the K6 column. Five classes were created, each covering an equal interval length of the K6 values that range from 0 to 24, as depicted in Figure 1. Consequently, the original K6 column was substituted for the categorized K6 column (K6 class). The distribution of the entries within the NYCHS by K6 class is illustrated in Figure 2**.**

2.3. Understanding Variables directly influencing K6 score class

The overall goal of our analysis is to understand which factors significantly influence mental health, and use these factors to predict the mental health class of the NYCHS respondents. In order to understand these characteristics and traits, the first step is to isolate some of them from the 168 column variables in the dataset. This was done by implementing reverse selection, with K6 class as the independent variable and the oher 77 column variables as the dependent variables.

2.4. Predicting a survey respondent's K6 score class

Once we obtained a better understanding of the predictor columns that directly influence the K6 class, the next major step in our analysis was to predict a respondent's mental health class using these predictors. To achieve this, we trained a classification model with the training data set, iterating through different combinations of hyperparameters and noting accuracy scores to determine the optimum model. The last step involved using the trained model to predict the testing data. These steps of training and testing are repeated across the following classification algorithms:

a. Multiple Logistic Regression (MLR)
b. K-Nearest Neighbors (KNN)
c. Random Forests (RF)
d. Support Vector Machines (SVM)

To determine the optimal model, cross-validation is employed to assess the accuracy score on the testing data predictions, simulating the model's performance on previously unseen data.

3. Results

To identify the predictor columns with a direct impact on K6 classification, we employed a backward selection approach. This involved testing multiple linear regression models, each with K6 as the dependent variable and utilizing various combinations of the 78 column variables remaining after data cleaning and preprocessing. The performance of these models was cross-validated using the Akaike Information Criterion (AIC) to identify the optimum model.

From the selected optimum model, predictor columns influencing the K6 class were determined based on p-value scores less than 0.05. The magnitude and direction of these effects were described by the coefficient values. In Figure 3, we document the 30 predictor column variables associated with p-values less than 0.05, providing insight into the key features influencing the K6 classification.

After finalizing the 30 predictor columns directly affecting the K6 class, the next step involved training machine learning algorithms (MLR, KNN, RF, SVM) with the training data, using only these predictor columns. Hyperparameters were varied during training to obtain the optimum set. The optimized model was then used to predict the classification of the testing data, noting the accuracy score of these predictions as the cross validation metric to compare the different classification algorithms' performance on unseen data. The results of each of the optimized algorithms' accuracy scores on the testing data are illustrated in Figure 4. The model with lowest

accuracy score of 58% was the k-nearest neighbor implementation, and the random forest model had the highest accuracy score of 67%.

The optimized random forests model stands out as the most effective in predicting K6 classes based on 30 predictor columns selected through backward selection. This optimized model reported an accuracy score of 67% and an AUC (Area Under the Receiver Operating Characteristic Curve) value of 0.56, which is greater than 0.5, indicating a performance better than that of a naive model. Figure 5 displays the confusion matrix depicting the model's predictions on the testing set. Notably, the illustration emphasizes the model's enhanced proficiency in predicting positive mental health classes (Good and Very Good), accompanied by a diminished accuracy in forecasting negative mental health cases (Bad and Critical).

4. Discussion

Our research revealed 30 factors that influenced respondents' mental health classification, as illustrated in Figure 3. These factors include finance and income levels, general health status, amount of daily physical exercise, the number of children aged seventeen or younger within the household, possession of medical insurance, and several others, all of which reported p-value scores < 0.05. Concurrently, studies in the United States (Sandra Mary Travasso et al., 2014) have similarly identified associations between mental health issues and factors such as increased poverty levels, poor health conditions, the number of young children within the household, heightened stress levels, and increased alcohol consumption [5].

The predictor columns that impact K6 classification span diverse aspects, with notable categorization based on dietary and personal consumption. Specifically, key predictors in this category encompass "fruitveg20," "avgsugarperday20," "nsodasugarperday20," and "daysalc30." These variables quantify the average daily intake of fruits and vegetables, the average consumption of sugar-sweetened drinks, the daily ingestion of sugar-sweetened sodas, and the quantity of standard alcoholic drinks consumed per week, respectively. A predominant trend among these predictors is the prevalence of negative coefficient values. This trend suggests a correlation between certain dietary behaviors, such as increased fruit/vegetable consumption and heightened sugar-sweetened beverage intake, with favorable mental health outcomes. In contrast, the positive coefficient associated with weekly standard alcoholic drink consumption indicates a link between elevated alcohol intake and diminished mental well-being.

The primary indicators associated with the financial dimension encompass "delaypayrent," "rodentsstreet," and "proudneigh." These variables represent the frequency of delayed rent payments, the presence of rodents on neighborhood streets, and the respondent's pride in their current residential area, respectively. Both the frequency of delayed rent payments and the visibility of rodents on neighborhood streets exhibit positive coefficients, indicating a correlation with adverse mental health outcomes. This aligns with findings from Travasoo et al., underscoring the association between heightened poverty levels and poor mental health because delayed rent payments and rodent infestation are factors commonly associated in low-income neighborhoods. Conversely, the respondent's pride in their neighborhood carries a

positive coefficient, suggesting a connection to unfavorable mental health; However, the magnitude of this coefficient is 0.14, suggesting a weak correlation and that the sign could have easily changed given a slightly different set of predictors. Additionally, the predictors "imputed_neighpovgroup4," "imputed_povertygroup," and "imputed_pov200," all representing diverse aspects of household income levels, feature negative coefficients. These coefficients imply that increased household incomes correlate with positive mental health, reinforcing the notion that decreased income levels correspond to adverse mental health—an assertion consistent with the findings of Travasoo et al. regarding the link between increased poverty levels and poor mental health.

Indicators related to the medical and health dimension encompass "generalhealth," "insuredgateway20," and "pcp20," representing the respondent's self-reported general health, the availability of personal medical insurance, and the presence of a personal general practitioner, respectively. These column variables exhibit positive coefficients of approximately 0.53, 0.38, and 0.20, implying an association with poor mental health. However, a detailed examination of the dataset metadata reveals that these variables are reverse scaled. For instance, a value of 1 in "pcp" indicates the availability of a personal doctor, while 2 represents the opposite. Consequently, this analysis suggests that increased access to medical resources is correlated with good mental health. Furthermore, the "didntgetcare20" predictor column takes on a value of 1 for respondents who needed treatment in the last twelve months but did not receive it, and a value of 2 for the opposite scenario. The negative coefficient of -0.70 suggests a link between respondents who never required medical assistance and did not receive it with good mental health. Conversely, respondents who needed treatment but did not receive it are associated with poor mental health.

Analyzing the predictor columns related to gender and demographic data revealed that individuals identifying as female are more prone to experiencing poor mental health compared to their male counterparts. The data indicates a higher prevalence of mental health issues among adults aged between 21 and 64, as indicated by the negative coefficient values in the "agegroup5" and "age21up" columns, which denote the respondent's age range. Additionally, the positive coefficient on the "birthsex" columns, where 1 indicates male and 2 indicates female birth sex, signifies the impact of being female on K6 classification. It's important to note that the survey assumes a binary gender system, and non-binary genders are not within the scope of this analysis.

In the results of the backward selection process, specific predictor columns have been pinpointed, raising speculation that these columns may not have a direct impact on the K6 class but could potentially be indicative of causal relationships. Notably, the "difficultdailyact" column, which gauges whether respondents struggle with waking up and performing daily tasks, holds a positive coefficient of 1.19, making it the second-highest magnitude coefficient according to Figure 3. This implies that the difficulty in executing routine tasks may not merely be a predictor but a potential causal factor prevalent among respondents experiencing poor mental health. Likewise, columns such as "nspd" (non-specific psychological distress) and "mhtreat20_all" (active participation in counseling services) report negative coefficient values of -9.45 and -1.67,

respectively, with both using the value 1 to indicate "Yes" and 2 to indicate "No." The negative coefficients suggest that these factors are predictors influencing the K6 class, where a negative response (value of 2) is directly correlated with poor mental health. However, it's crucial to consider the alternative perspective that non-specific psychological distress and engagement in counseling services might also be viewed as causal effects prevalent among respondents with compromised mental health. This nuanced interpretation underscores the complexity of the relationship between these predictors and mental health outcomes.

In the context of predicting the K6 class of a respondent, the optimum model is the random forests model which has a 67% accuracy when predicting the testing data set. Figure 5 displays the confusion matrix depicting this model's predictions on the testing set. Notably, the illustration emphasizes the model's enhanced proficiency in predicting positive mental health classes (Good and Very Good), accompanied by a diminished accuracy in forecasting negative mental health cases (Bad and Critical). This discrepancy could stem from the uneven distribution of entries across the dataset. As illustrated in Figure 5, showcasing the K6 distribution by class, a notable 87% of all entities fall within the Good and Very Good K6 classes, while a mere 2.61% and 0.6% of entries are attributed to the Bad and Critical classes. With only a small percentage of entries belonging to the Bad and Critical categories, the model may not have had sufficient data to learn and generalize patterns effectively for both these subclasses.

5. Conclusion

This study aimed to uncover the relationship between mental health and different characteristics among residents of New York City. It revealed 30 characteristics, ranging from socioeconomic indicators to lifestyle choices, providing valuable insights into the factors influencing an individual's mental well-being in urban areas. The findings demonstrated that socioeconomic factors such as increased income levels, access to medical services and counseling, regular physical exercise, physical stature, were positively correlated with mental health. Conversely, higher alcohol intake, increased poverty levels, frequent consumption of sugar-sweetened beverages, and specific lifestyle choices were identified as negative influencers on mental health. Furthermore, our exploration into predictive modeling highlighted the superiority of the random forest model among the four models trained. This model demonstrated an impressive 66% accuracy in predicting the mental health classes of respondents in previously unseen data.

6. Limitations

During the data preparation and pre-processing, missing values were imputed using the column mean and mode. This imputation altered the variance across and might have introduced some bias into our dataset. The removal of the columns containing significant missing values may have led to the loss of valuable information and correlations, leaving a potentially biased subset of the data. It is possible that during the removal process, we may have removed some mental and behavioral characteristics that have been proven qualified therapists/psychiatrists to be directly linked to mental health. Instead of removing significantly missing columns, imputing

missing values using more sophisticated methods might have allowed us to handle missing data patterns more effectively.

Additionally, the selected features we adopted as the predictor columns directly affecting the K6 class may not be the most informative subset. A backward selection considers several multiple linear regressions and cross validates them based on their respective AIC scores. During this process, it is impossible that the backward selection considered every single possible combination of the 77 predictor variables. Therefore, the selected features we adopted as the predictor columns directly affecting the K6 class may not be the most informative subset because the most informative subset might have been one of the combinations that the backward selection overlooked.

Appendix

Figure 1: K6 Classes and Intervals

K6 Class	Class Description	K6 Range	% Of NYCHS entries
1	Very Good	0 - 4	62%
2	Good	5 – 9	25%
3	Moderate	10 – 14	9%
4	Bad	15 – 19	3%
5	Critical	20 - 24	1%

Figure 2: Distribution of Dataset Values by K6 Class

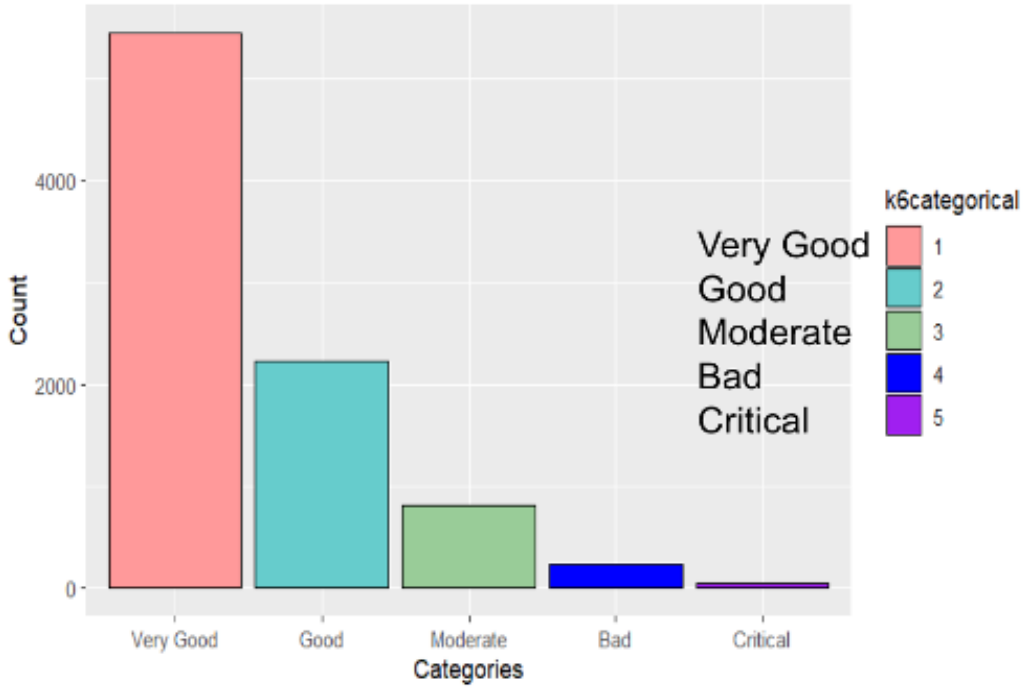


Figure 3: Predictor Column Variables Affecting K6 Classification (p-values ≤ 0.05)

Column Title	Explanation	Coefficient	P-value
mood11	Needed Treatment but did not Receive	-1.344248949	1.783631e-13
newrace	Race Identified with	-0.225239924	1.170806e-02
newrace6	Race Parents Identified with	0.1950923866	1.839947e-02
agegroup5	Age Range	-0.4810573304	2.698538e-04
age21up	Age 21 +? (Yes/No)	-0.4798210438	1.799366e-02
imputed_neighpovgroup4_	Annual Household Income	-0.146358079	3.362378e-04
imputed_povertygroup	Annual Household Income 200% > FPL (Yes/No)	-0.1130978421	2.434674e-02
imputed_pov200	Annual Income Categorical	-0.2987134951	4.300525e-02
generalhealth	General Health Estimation	0.5389584137	2.993751e-47
insuredgateway20	Has any form of Medical Insurance	0.3815287601	3.876927e-02
insure5	Type of Medical Insurance had	-0.1320247176	2.102000e-02
pcp20	Do you have a Personal Doctor?	0.2040180970	1.333709e-02
medplace	Where do you get Medical Assistance	-0.0089662179	1.333709e-02
didntgetcare20	Needed Medical Aid in last 12 months and didn't get	-0.7001360776	5.108623e-10
everasthma	Ever had Asthma	-0.3409932890	9.267611e-04
nspd	Non-specific Psychological distress	-9.4578763399	0.000000e+00
mhtreat20_all	Received Counselling or Prescription Medication	-1.6704921381	3.326861e-25
delaypayrent	Delayed Housing Payment in the last 12 months	0.7423656981	1.653609e-13
rodentsstreet	Seen Rodents on your street in the last 12 months	0.4998978185	1.263404e-10
proudneigh	Proud of the Neighborhood that you live in	-0.1481811006	2.007274e-03
fruightveg20	Daily Intake of Fruits and Vegetables	-0.167691449	-2.015382
weightall	Height without Shoes	-0.5231725837	1.865649e-02
avgsugarperday20	Average Number of Sugar Drinks Daily	-0.1253204682	2.993751e-47
nsodasugarperday20	How Often you Drink Sugar Sweetened Drinks	-0.1682373698	4.449017e-03
cyclingfreq	Frequency of Cycling Activities	-0.0682309725	2.682401e-02
difficultdailyact	Difficulty Getting up and performing required Duties	1.1985374810	5.634867e-25
daysalc30	Average Number of Standard Drinks per day	0.0333884101	2.627521e-10
insultipv	Verbal Abuse from a Partner/Spouse	-0.7748397251	3.270968e-09
child	Number of Children Younger than 17	0.3464826795	6.188038e-03
birthsex	Birth Sex Assigned	0.252343224	6.362917e-04

Figure 4: Model Performance on Testing Data

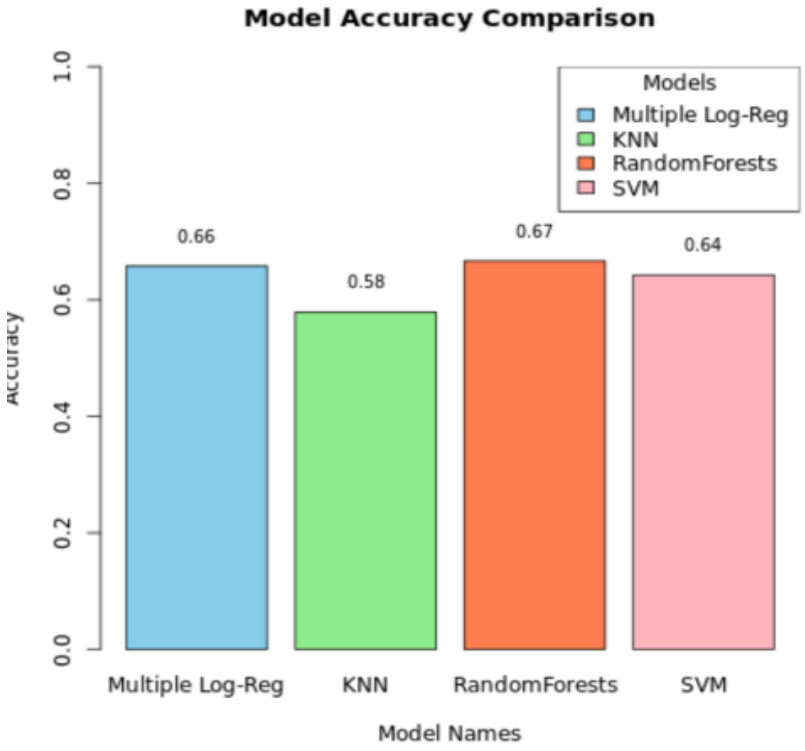
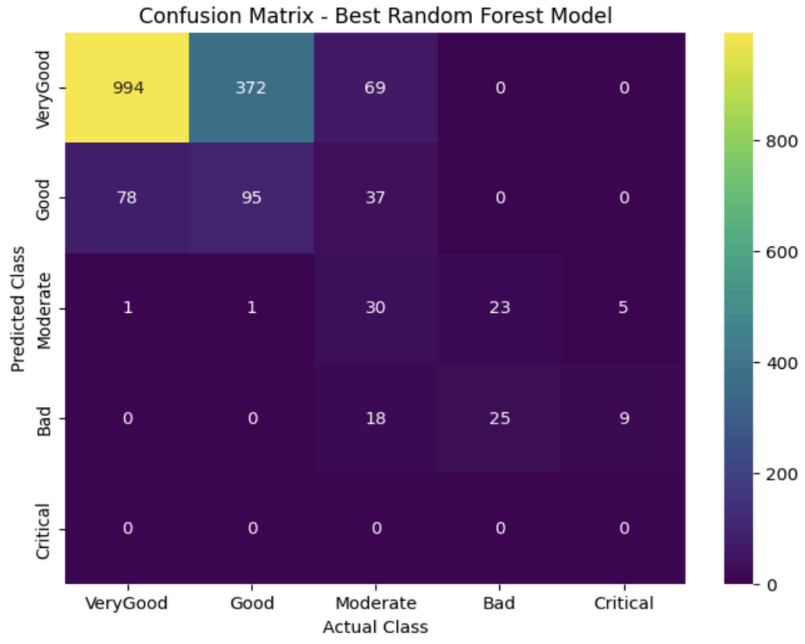


Figure 5: Optimized Random Forests Model Confusion Matrix



References

[1] https://ourworldindata.org/mental-health
[2] https://ourworldindata.org/mental-health
[3] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide
[4] https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3922014/
https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-10560-y
https://digital.nhs.uk/blog/transformation-blog/2018/the-past-present-and-future-of-innovation-in-mental-healthhttps://www.who.int/health-topics/mental-health#tab=tab_1
https://digital.nhs.uk/blog/transformation-blog/2018/the-past-present-and-future-of-innovation-in-mental-health
https://www.nyc.gov/site/doh/data/data-sets/community-health-survey-methodology.page

https://www.nimh.nih.gov/health/statistics/mental-illness
https://www.nyc.gov/site/doh/data/data-sets/community-health-survey-public-use-data.page
https://search.r-project.org/CRAN/refmans/Rfast2/html/lm.bsreg.html
https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960

[5] https://bmcwomenshealth.biomedcentral.com/articles/10.1186/1472-6874-14-22