HIV Epidemiology Project

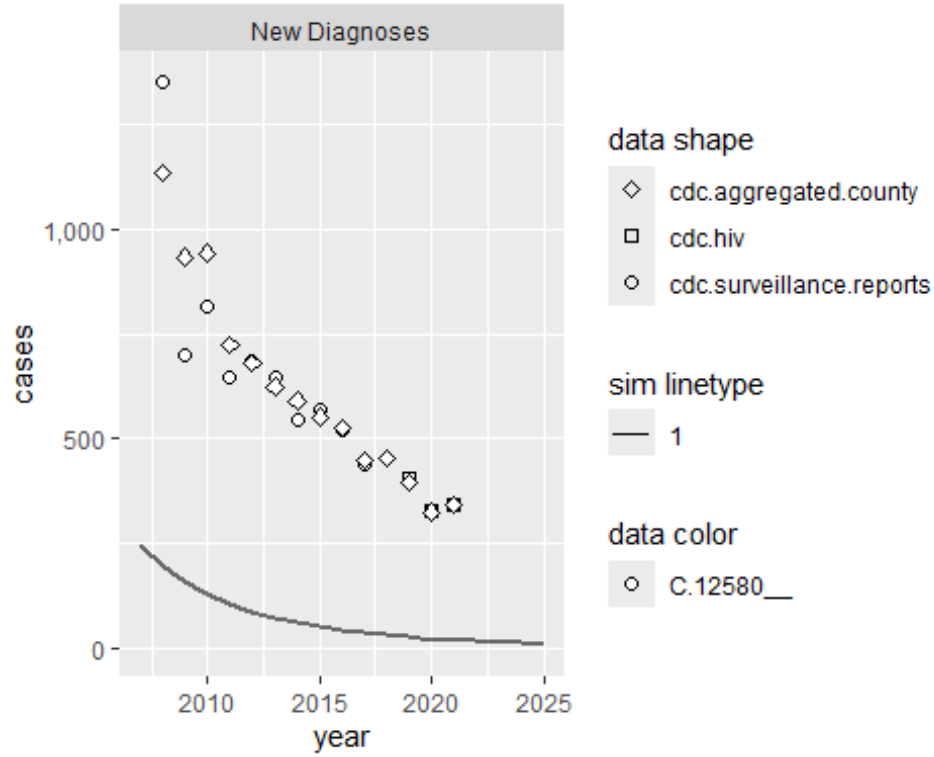
Mohammed Ba-Aoum

2024-09-23

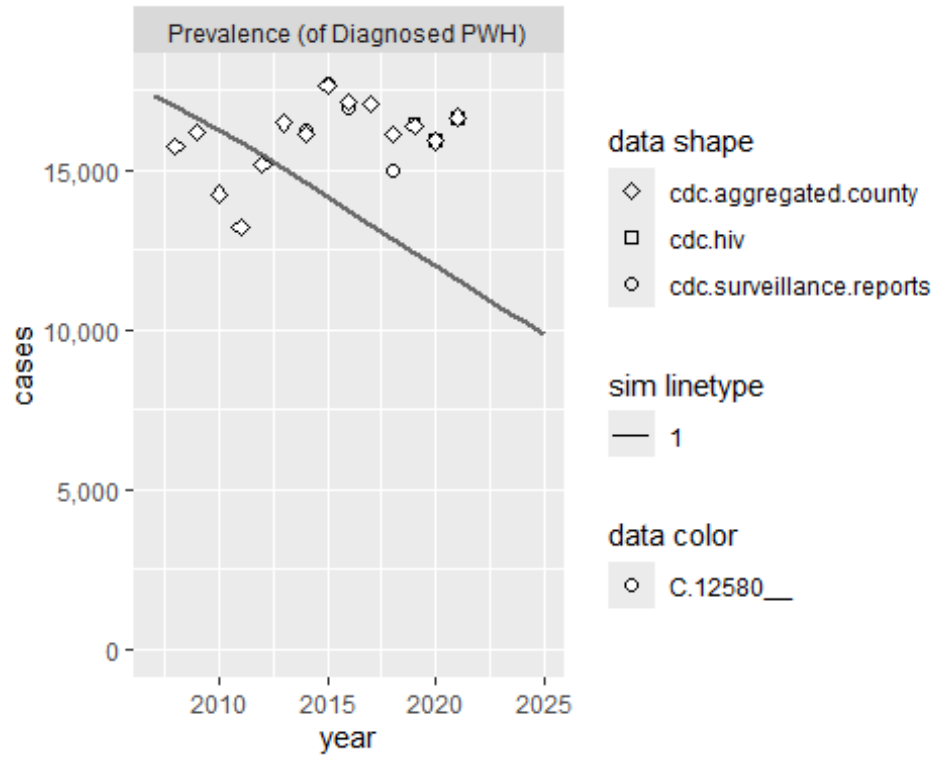
PART 1: Code Setup and Implementation

```
##------------------------------------------------------------------##
##-- PART 1: Code Setup and Implementation                       --##
##------------------------------------------------------------------##

# Source the script that installs the required packages
source('first_time_setup/install_packages.R')

## Skipping install of 'distributions' from a github remote, the SHA1 (808b78
e8) has not changed since last install.
##   Use `force = TRUE` to force installation

## Skipping install of 'bayesian.simulations' from a github remote, the SHA1
(1da1851e) has not changed since last install.
##   Use `force = TRUE` to force installation

## Skipping install of 'locations' from a github remote, the SHA1 (914af72e)
has not changed since last install.
##   Use `force = TRUE` to force installation

source('applications/EHE/ehe_specification.R')

## Loading Surveillance Manager (may take a minute or two)...Done!
## Loading Census Manager (may take a minute or two)...Done!

# Create the model engine for a given location (c.12580 is the code for Balti
more City) (this may take a few moments)
engine = create.jheem.engine(version = 'ehe', location = 'c.12580', end.year=
2025, max.run.time.seconds = 10)
# Load a set of parameters (set at default values)
params = suppressWarnings(get.medians(EHE.PARAMETERS.PRIOR))
# Set the global transmission rate to equal 0.01
params['global.trate'] = 0.01

# Using the model engine, run and save a single simulation using the transmis
sion rate you loaded above (this may take a few moments)
sim = engine$run(parameters = params)

# Visualize and describe the simulation fit for projected "new HIV diagnosis"
and "prevalence of diagnosed HIV" against CDC's reported data:
# 1. New diagnoses
simplot(sim, outcomes = "new", dimension.values = list(year = 2007:2025))
```

```
## Warning: No shared levels found between `names(values)` of the manual scal
e and the
## data's fill values.
```




```
# 2. Diagnosed prevalence
simplot(sim, outcomes = "diagnosed.prevalence", dimension.values = list(year
= 2007:2025))

## Warning: No shared levels found between `names(values)` of the manual scal
e and the
## data's fill values.
```



Comments on Part 1

Plot 1: New Diagnoses

The first plot compares the simulated trend of new HIV diagnoses with CDC-reported data from 2007 to 2025. While the model captures the overall downward trend observed in the data, it shows a more consistent decline in new diagnoses compared to the CDC data, which features an initial peak followed by a gradual decline. This suggests that the model consistently underestimates new diagnoses, especially in the earlier years (2007-2015). This misalignment could be due to an overly conservative global transmission rate or unaccounted factors such as behavioral changes, prevention efforts, or variations in testing rates. To improve the fit, it may be necessary to re-calibrate these parameters and incorporate additional factors that contributed to the observed peaks in early diagnoses.

Plot 2: Diagnosed Prevalence of HIV

The second plot illustrates the diagnosed HIV prevalence, comparing model simulations to CDC-reported data. The model predicts a continuous decline, while the observed data shows a slower, more stable trend from 2012 to 2020. This suggests the model might overestimate intervention impacts or reductions in transmission. The mismatch, particularly notable in later years, points to potential inaccuracies in assumptions about treatment, care retention, or mortality rates. To enhance model alignment with observed

data, a review and adjustment of assumptions such as testing, care linkage, and mortality are recommended.

Part 2 Code Implementation and Analysis

```
##-------------------------------------------------------------------##
##-- Part 2: Code Implementation and Analysis                     --##
##-------------------------------------------------------------------##

# Set the R seed to 1234 to ensure reproducibility
set.seed(1234)

# Sample three different values of global.trate from a uniform distribution b
etween 0 and 0.05
global_trate_samples <- runif(3, min = 0, max = 0.05)

# Print the sampled global.trate values
print("Sampled global.trate values:")
```

```
## [1] "Sampled global.trate values:"
```

```
print(global_trate_samples)
```

```
## [1] 0.005685171 0.031114970 0.030463737
```

```
# Initialize an empty list to store simulations
simulations <- list()

# Run simulations for each sampled value of global.trate
for (i in 1:3) {
  # Update the parameters with the sampled global.trate value
  params['global.trate'] <- global_trate_samples[i]

  # Run the simulation using the updated parameters
  sim <- engine$run(parameters = params)

  # Store the simulation in the list
  simulations[[i]] <- sim

  # Print a message indicating the simulation was completed
  print(paste("Completed simulation", i, "with global.trate =", round(global_
trate_samples[i], 4)))
}
```

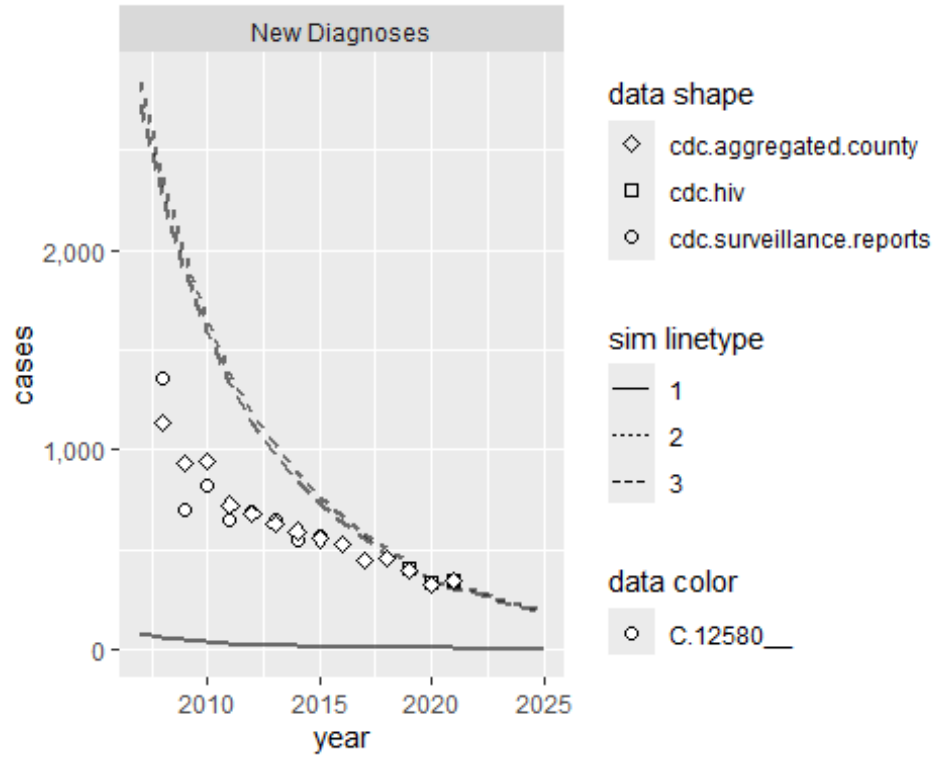
```
## [1] "Completed simulation 1 with global.trate = 0.0057"
## [1] "Completed simulation 2 with global.trate = 0.0311"
## [1] "Completed simulation 3 with global.trate = 0.0305"
```

```r
# Visualize and compare the outcomes of the three simulations

# Plotting new HIV diagnoses for the three simulations
print("Visualizing new HIV diagnoses for the three simulations...")
```

```
## [1] "Visualizing new HIV diagnoses for the three simulations..."
```

```r
# 1. New diagnoses
simplot(
  simulations[[1]],
  simulations[[2]],
  simulations[[3]],
  outcomes = "new",
  dimension.values = list(year = 2007:2025)
)
```

```
## Warning: No shared levels found between `names(values)` of the manual scal
e and the
## data's fill values.
```

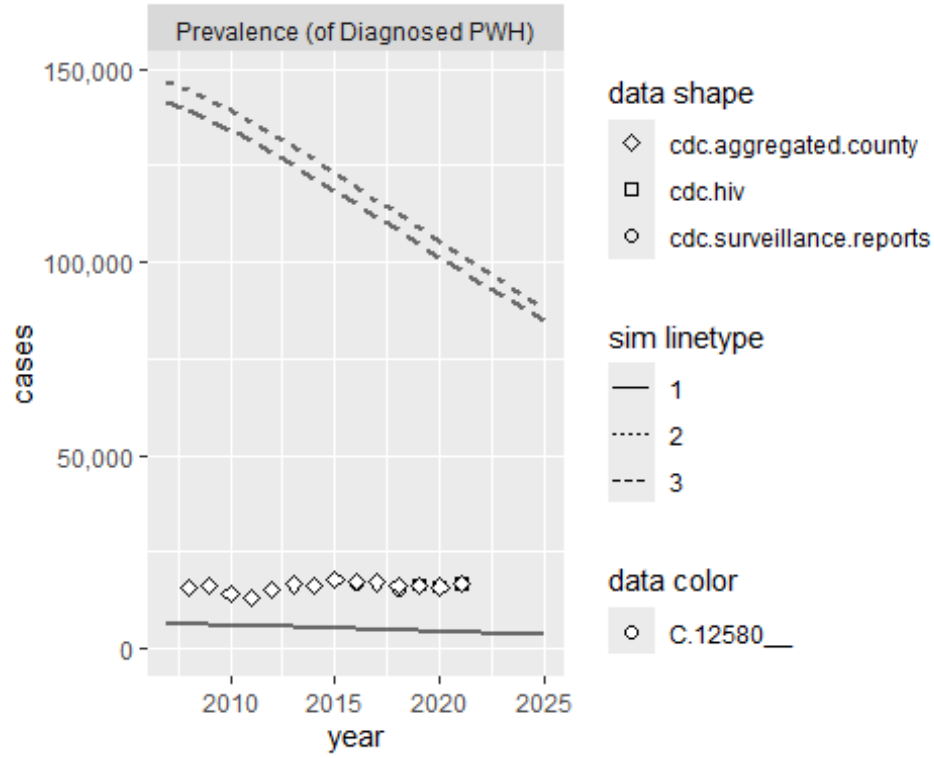



```r
# Plotting the prevalence of diagnosed HIV for the three simulations
print("Visualizing prevalence of diagnosed HIV for the three simulations...")
```

```
## [1] "Visualizing prevalence of diagnosed HIV for the three simulations..."
```

```r
# 2. Diagnosed prevalence
simplot(
  simulations[[1]],
```

```
  simulations[[2]],
  simulations[[3]],
  outcomes = "diagnosed.prevalence",
  dimension.values = list(year = 2007:2025)
)

## Warning: No shared levels found between `names(values)` of the manual scal
e and the
## data's fill values.
```



Comments on part 2:

Plot 1: New Diagnoses

Simulations 2 and 3 show a steep decline in new diagnoses over time, with the simulated lines starting much higher and declining more rapidly than the observed data. Simulation 1, on the other hand, shows a slight decline before becoming almost stable. The observed data from CDC sources, however, shows a more gradual decline compared to Simulations 2 and 3 and consistently higher values than Simulation 1 throughout the period, particularly in the early years (2007-2015). This suggests that the models in Simulations 2 and 3 overestimate new diagnoses, while Simulation 1 underestimates them based on the global trend. This indicates that the optimal value of global.trate would be between 0.0057 and 0.0311. In this plot, Simulations 2 and 3 appear to have a better fit than Simulation 1.

Plot 2: Diagnosed Prevalence of HIV

The second plot compares the trends in the prevalence of diagnosed HIV cases for the three simulations against CDC-reported data. The observed data indicates a relatively stable prevalence of diagnosed cases from 2007 to 2025, with slight fluctuations. Simulation 1 shows similar behavior and captures the trend accurately.

In contrast, Simulations 2 and 3 predict a consistent and substantial decline in prevalence over the same period. The simulations differ slightly in their trajectories, with Simulation 3 showing a slightly more significant decline, reflecting its higher sampled global.trate value. This suggests that increased transmission rates lead to greater reductions in prevalence within the model's framework, likely due to assumptions around treatment uptake, retention in care, and mortality rates.

However, the declining trends in Simulations 2 and 3 diverge significantly from the observed data, indicating a potential overestimation of intervention impacts or an underestimation of ongoing transmission and retention challenges. To better align the model with real-world data, adjustments to factors like care linkage, retention, and mortality dynamics should be considered, along with further calibration of transmission rates and other key parameters.

Which Simulation Provides the Best Fit?

Among the three simulations, none perfectly aligns with the observed data for new diagnoses or diagnosed prevalence. In the first plot (new diagnoses), Simulation 3 appears to have a better fit. However, Simulation 1, which shows the slowest decline among the three, seems closer to the observed trends than Simulations 2 and 3, especially in the second plot (prevalence).

Despite this, the fit is still far from ideal in both plots, highlighting the need for further calibration and refinement. The underestimation of new diagnoses and the overestimation of prevalence decline suggest that key model assumptions, such as the effectiveness of interventions, transmission rates, and care dynamics, may need to be revisited. Further fine-tuning of these parameters, potentially through targeted calibration efforts and integration of additional data, will be essential for improving the model's accuracy and better reflecting the observed patterns in CDC data.

Part 3 Model Calibration

```
##------------------------------------------------------------------##
##-- Part 3: Model Calibration                                   --##
##------------------------------------------------------------------##

# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

# Extract the calibration targets for new diagnoses and prevalence of diagnos
ed HIV
new.diagnosis.target <- SURVEILLANCE.MANAGER$pull(outcome = "diagnoses", loca
tion = "C.12580", source = "cdc.aggregated.county")
diagnosed.prevalence.target <- SURVEILLANCE.MANAGER$pull(outcome = "diagnosed
.prevalence", location = "C.12580", source = "cdc.aggregated.county")

# Define the years to match for extraction
years_to_match <- 2008:2021

# Extract simulated values and ensure lengths match the target data
new.diagnoses.sim <- sim$get("new", year = years_to_match)
diagnosed.prevalence.sim <- sim$get("diagnosed.prevalence", year = years_to_m
atch)



# Function to calculate the goodness-of-fit measure (RMSE) for a given global
.trate
calculate_fit <- function(global_trate) {
  # Update the global.trate parameter in the simulation
  params['global.trate'] <- global_trate

  # Run the simulation with the updated parameter
  sim <- engine$run(parameters = params)

  # Extract the simulated values
  new.diagnoses.sim <- sim$get("new", year = years_to_match)
  diagnosed.prevalence.sim <- sim$get("diagnosed.prevalence", year = years_to
_match)



  # Calculate RMSE for new diagnoses
  rmse_new <- sqrt(mean((new.diagnoses.sim - new.diagnosis.target)^2, na.rm =
TRUE))

  # Calculate RMSE for diagnosed prevalence
  rmse_prevalence <- sqrt(mean((diagnosed.prevalence.sim - diagnosed.prevalen
ce.target)^2, na.rm = TRUE))
```

```r
  # Combine the RMSE values
  total_rmse <- rmse_new + rmse_prevalence

  # Return the total RMSE as the measure of fit
  return(total_rmse)
}

# Run the optimization to find the best global.trate
print("Starting optimization...")
```

```
## [1] "Starting optimization..."
```

```r
result <- optim(par = 0.01, fn = calculate_fit, method = "L-BFGS-B", lower =
0, upper = 0.1)

# Extract the optimal global.trate
best_global_trate <- result$par

# Run the simulation one more time with the best-fit global.trate
params['global.trate'] <- best_global_trate
best_fit_sim <- engine$run(parameters = params)

# Extract the simulated values for visualization
best_new_diagnoses <- best_fit_sim$get("new", year = years_to_match)
best_diagnosed_prevalence <- best_fit_sim$get("diagnosed.prevalence", year =
years_to_match)

# Summary of results
print(paste("The best-fit global.trate is:", round(best_global_trate, 4)))
```

```
## [1] "The best-fit global.trate is: 0.0105"
```

```r
print(paste("RMSE for new diagnoses with best-fit global.trate:", round(calcu
late_fit(best_global_trate), 2)))
```

```
## [1] "RMSE for new diagnoses with best-fit global.trate: 3182"
```

Comment on Part 3-1

Best-Fit Global Transmission Rate and RMSE

The best fit Global trate:0.0105

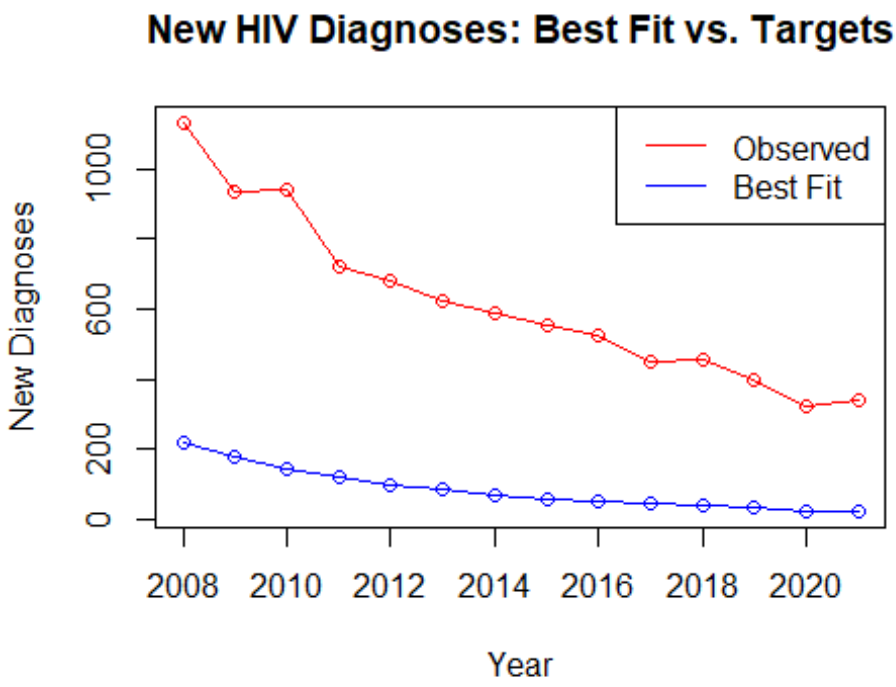
RSEM: 3182

The optimization process identified the best-fit global transmission rate (global.trate) as 0.0105, which was consistent even when the upper bound was increased from 0.05 to 0.1.

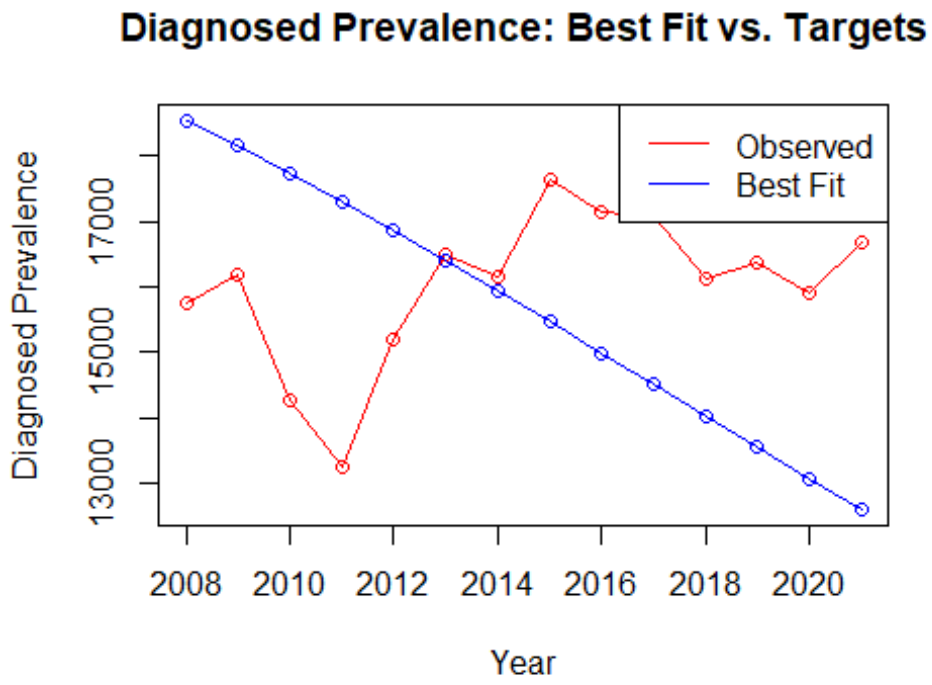
This indicates that the optimizer consistently converges to this value within the defined parameter space, suggesting that the true optimal transmission rate lies near this point. RMSE for new diagnoses with this best-fit transmission rate is 3182.03, which reflects the model's overall fit to the observed data. Despite varying bounds, the RMSE value did not significantly improve, indicating that further changes to global.trate are unlikely to yield a better fit without additional adjustments to other model parameters.

While a lower RMSE generally indicates a better model fit, what qualifies as a "good" RMSE depends on the context, including the scale of the data, the specific application, the units of measurement, and the modeling objectives. However, RMSE is not always the most appropriate measure to validate a model performance , particularly because it assumes that errors are normally distributed. If the residuals are skewed or not symmetrically spread around zero, RMSE might fail to accurately reflect the model's true performance. Thus, it is necessary to combine statistical analysis against histroical data with expert judgments when evaluating a model performance.

```
# Plotting the fit of the best simulation against the targets
plot(years_to_match, new.diagnosis.target, type = "o", col = "red", ylim = ra
nge(c(new.diagnosis.target, best_new_diagnoses)),
     xlab = "Year", ylab = "New Diagnoses", main = "New HIV Diagnoses: Best F
it vs. Targets")
lines(years_to_match, best_new_diagnoses, type = "o", col = "blue")
legend("topright", legend = c("Observed", "Best Fit"), col = c("red", "blue")
, lty = 1)
```



```r
plot(years_to_match, diagnosed.prevalence.target, type = "o", col = "red", yl
im = range(c(diagnosed.prevalence.target, best_diagnosed_prevalence)),
     xlab = "Year", ylab = "Diagnosed Prevalence", main = "Diagnosed Prevalen
ce: Best Fit vs. Targets")
lines(years_to_match, best_diagnosed_prevalence, type = "o", col = "blue")
legend("topright", legend = c("Observed", "Best Fit"), col = c("red", "blue")
, lty = 1)
```



Comment on Plot Analysis for part 3:

Plot of New HIV Diagnoses

The observed data (in red) shows a significant decline in new diagnoses over the years, with noticeable fluctuations. However, the model consistently underestimates the number of new diagnoses throughout the entire period. Although the best-fit line captures the overall downward trend, it fails to align with the magnitude and variability seen in the observed data, particularly during periods when diagnoses are relatively high.

This discrepancy suggests that, despite calibrating the transmission rate, other factors influencing new diagnoses—such as intervention coverage, behavioral changes, or other model parameters—may need further adjustment.

Plot of Diagnosed Prevalence

The observed data (in red) exhibits considerable variability, with peaks and troughs, while the simulated prevalence shows a smoother, steady decline. The best-fit simulation (in blue) captures the overall downward trend but does not align well with the fluctuations observed in the real-world data.

The observed prevalence initially starts below the model's estimates but intersects around 2013, after which the model's predictions fall consistently below the observed data. This misalignment suggests that the model may not fully capture the complex interactions that influence diagnosed prevalence, such as testing rates, treatment uptake, or changes in the population at risk over time.

Part 4: Data Extraction and Analysis

```
##-------------------------------------------------------------------------
---------------------------------##
##-- Part 4: Data Extraction and Analysis
--##
##-------------------------------------------------------------------------
---------------------------------

# Load necessary libraries

library(ggplot2)


# Task 1: Extract the Number of New Diagnoses
# Extracting new diagnoses with full subgroup dimensions
new.diagnoses <- sim$get("new", keep.dimensions = c("location", "year", "age"
, "race", "sex", "risk"))


##Task2: Describe the dimensions and names of the dimensions of the Extracted
Objects

print("Structure of new.diagnoses with subgroup dimensions:")

## [1] "Structure of new.diagnoses with subgroup dimensions:"

str(new.diagnoses)

##  num [1, 1:56, 1:5, 1:3, 1:3, 1:3, 1] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 7
##   ..$ location: chr "C.12580"
##   ..$ year    : chr [1:56] "1970" "1971" "1972" "1973" ...
##   ..$ age     : chr [1:5] "13-24 years" "25-34 years" "35-44 years" "45-54
```

```
years" ...
##   ..$ race    : chr [1:3] "black" "hispanic" "other"
##   ..$ sex     : chr [1:3] "heterosexual_male" "msm" "female"
##   ..$ risk    : chr [1:3] "never_IDU" "active_IDU" "IDU_in_remission"
##   ..$ sim     : chr "1"

print("Dimensions and names:")

## [1] "Dimensions and names:"

print(dim(new.diagnoses))

## location     year      age     race      sex     risk      sim
##        1       56        5        3        3        3        1

print(dimnames(new.diagnoses))

## $location
## [1] "C.12580"
##
## $year
##  [1] "1970" "1971" "1972" "1973" "1974" "1975" "1976" "1977" "1978" "1979"
## [11] "1980" "1981" "1982" "1983" "1984" "1985" "1986" "1987" "1988" "1989"
## [21] "1990" "1991" "1992" "1993" "1994" "1995" "1996" "1997" "1998" "1999"
## [31] "2000" "2001" "2002" "2003" "2004" "2005" "2006" "2007" "2008" "2009"
## [41] "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017" "2018" "2019"
## [51] "2020" "2021" "2022" "2023" "2024" "2025"
##
## $age
## [1] "13-24 years" "25-34 years" "35-44 years" "45-54 years" "55+ years"
##
## $race
## [1] "black"    "hispanic" "other"
##
## $sex
## [1] "heterosexual_male" "msm"               "female"
##
## $risk
## [1] "never_IDU"        "active_IDU"       "IDU_in_remission"
##
## $sim
## [1] "1"

#Part4 continue
# Task 3: Filter Objects for Specific Subgroups
# a. Filter the new.diagnoses to focus on the subgroup: 13-24 years old, Blac
k, MSM, and never IDU.

# Extract the correct indices based on the dimension names obtained from your
data
location_index <- 1  # Only one location available: "C.12580"
```

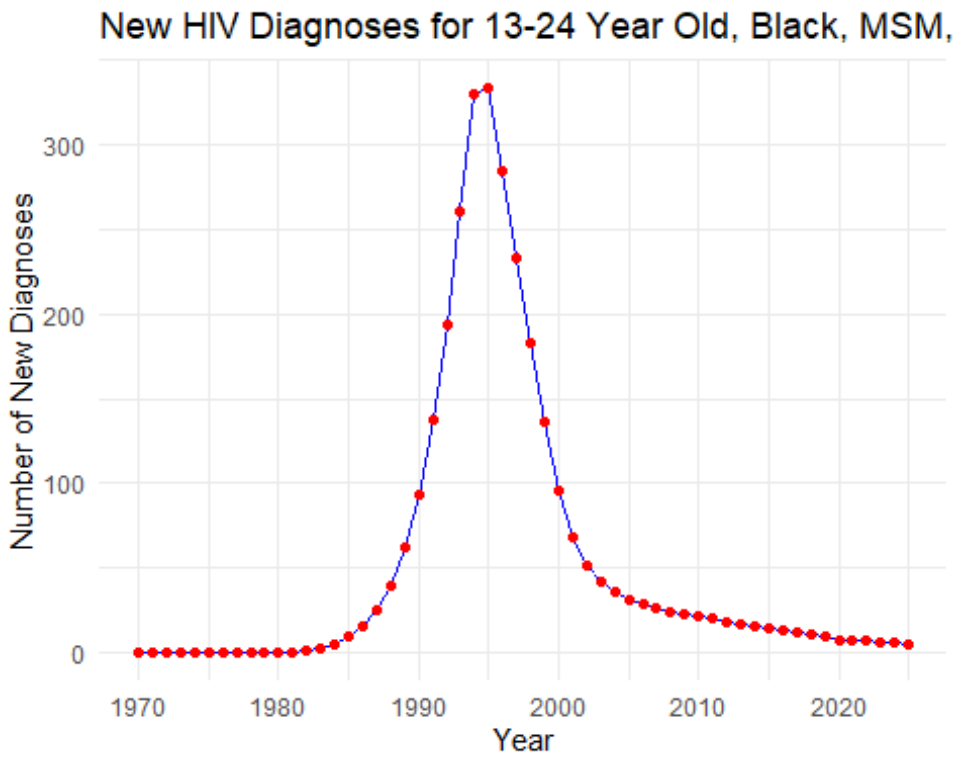
```r
year_index <- 1:dim(new.diagnoses)[2]  # All years available (56 years in tot
al)
age_index <- which(dimnames(new.diagnoses)$age == "13-24 years")  # Index for
"13-24 years"
race_index <- which(dimnames(new.diagnoses)$race == "black")  # Index for "bl
ack"
sex_index <- which(dimnames(new.diagnoses)$sex == "msm")  # Index for "msm"
risk_index <- which(dimnames(new.diagnoses)$risk == "never_IDU")  # Index for
"never_IDU"

# Filter the data for the subgroup of interest
new_filtered <- new.diagnoses[location_index, year_index, age_index, race_ind
ex, sex_index, risk_index, , drop = FALSE]

# Check the filtered data to ensure it is correctly extracted
print("Filtered new diagnoses data for 13-24 years old, Black, MSM, never IDU
:")
```

```
## [1] "Filtered new diagnoses data for 13-24 years old, Black, MSM, never ID
U:"
```

```r
print(new_filtered)
```

```
## , , age = 13-24 years, race = black, sex = msm, risk = never_IDU, sim = 1
##
##          year
## location  1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981
1982
##   C.12580    0    0    0    0    0    0    0    0    0    0    0    0 1.00
4606
##          year
## location     1983     1984     1985     1986    1987    1988     1989
1990
##   C.12580 2.96707 5.260135 9.054225 15.18423 24.7829 39.5303 61.60473 93.6
3623
##          year
## location      1991     1992     1993     1994     1995     1996     1997
##   C.12580 137.2362 193.3761 260.4704 329.5211 333.5718 284.8443 233.2964
##          year
## location      1998     1999     2000     2001     2002  2003     2004
2005
##   C.12580 182.8592 136.0095 95.83905 68.66911 51.92381 41.76 35.52617 31.5
1783
##          year
## location      2006     2007     2008     2009     2010     2011     2012
##   C.12580 28.70402 26.35732 24.31422 22.60338 21.22423 19.88386 18.50501
##          year
## location      2013     2014     2015     2016     2017     2018     2019
##   C.12580 17.15896 15.86835 14.61924 13.35991 12.13489 10.97378 9.882101
##          year
```

```
## location      2020     2021     2022     2023     2024     2025
##   C.12580 7.446766 7.505718 7.061188 6.233047 5.502404 4.857152

# Task 3b: Plot the filtered data over time
# Extract years from the dimension names for plotting
years <- as.numeric(dimnames(new.diagnoses)$year)

# Sum the filtered diagnoses across the subgroup
new_filtered_agg <- as.numeric(new_filtered)

# Plotting the filtered data over time
ggplot(data = data.frame(Year = years, Diagnoses = new_filtered_agg), aes(x =
Year, y = Diagnoses)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "New HIV Diagnoses for 13-24 Year Old, Black, MSM, Never IDU",
       x = "Year", y = "Number of New Diagnoses") +
  theme_minimal()
```



Part 4 Task 3. c comment on plot of New HIV Diagnoses for 13-24 Year Old, Black, MSM, Never IDU:

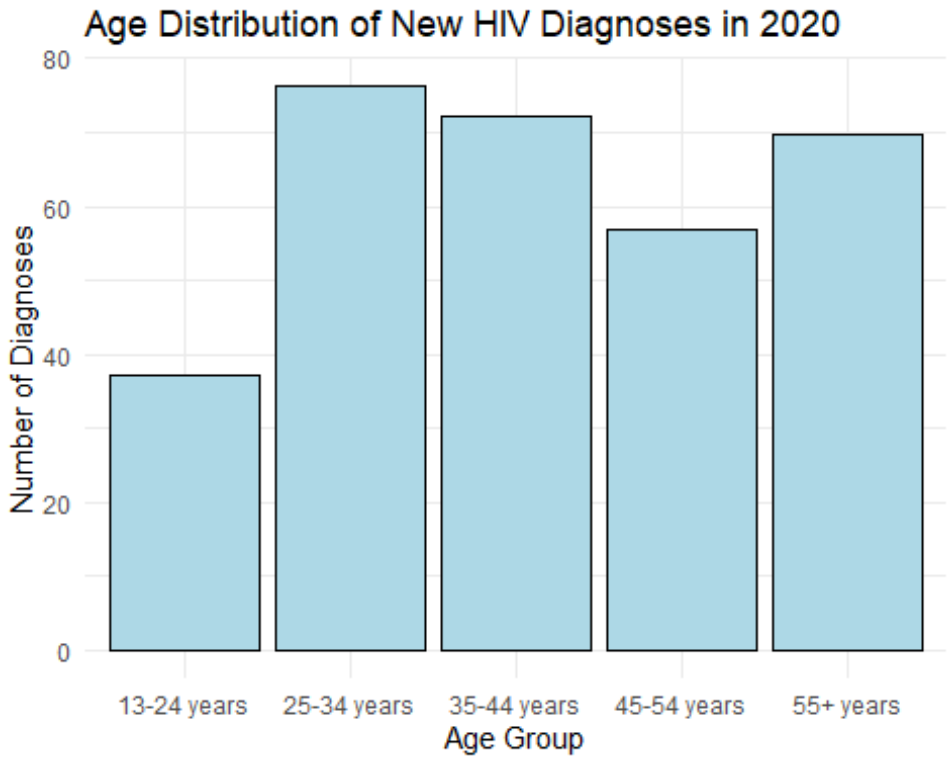
The plot shows a sharp increase in diagnoses during the 1990s, peaking around 1995, followed by a rapid decline in subsequent years. This trend could be influenced by the

broader HIV epidemic dynamics during this period, including high transmission rates before effective interventions became widely accessible.

The sharp decline after the peak suggests the impact of improved prevention efforts such as increased awareness, and access to testing and care. However, the steep nature of the curve may indicate that factors like behavior change or targeted interventions had a strong effect on this specific subgroup.

```
# Task 4: Draw a histogram of the age distribution for new HIV diagnoses in t
he year 2020

# Extract the index for the year 2020
year_2020_index <- which(dimnames(new.diagnoses)$year == "2020")

# Aggregate data across all subgroups for the year 2020, focusing on the age
distribution
# We will sum across all other dimensions except age to get the distribution
of diagnoses by age
# This effectively collapses the array by summing over location, race, sex, r
isk, and sim dimensions
age_distribution_2020 <- apply(new.diagnoses[, year_2020_index, , , , , , dro
p = FALSE], 3, sum)

# Create a data frame for plotting the age distribution
age_labels <- dimnames(new.diagnoses)$age
age_data <- data.frame(Age = age_labels, Diagnoses = age_distribution_2020)


# Calculate the proportion of new diagnoses in 2020 that occurred among 13-24
years old
# Find the index for the "13-24 years" age group
age_13_24_index <- which(age_labels == "13-24 years")
proportion_13_24 <- sum(age_distribution_2020[age_13_24_index]) / sum(age_dis
tribution_2020)

# Print the proportion of new diagnoses among 13-24 years old
print(paste("Proportion of new diagnoses in 2020 among 13-24 years old:", rou
nd(proportion_13_24 * 100, 2), "%"))

## [1] "Proportion of new diagnoses in 2020 among 13-24 years old: 11.93 %"

# Plotting the age distribution using a bar chart
ggplot(age_data, aes(x = Age, y = Diagnoses)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +
  labs(title = "Age Distribution of New HIV Diagnoses in 2020",
       x = "Age Group", y = "Number of Diagnoses") +
  theme_minimal()
```

```
Proportion of new diagnoses in 2020 among 13-24 years old: 11.93 % ( around
12%)
```



Comment on Age Distribution of New HIV Diagnoses in 2020:

The plot highlights that the 25-34 years age group had the highest number of new diagnoses, followed by the 35-44 and 55+ years age groups. The 13-24 years group, which is of specific interest in the previous figure, had the lowest number of new diagnoses among the groups shown.This distribution suggests that HIV prevention and intervention efforts need to be especially focused on young adults and middle-aged individuals, where new infections remain substantial.

The relatively lower proportion (12 %) of new diagnoses among 13-24 years old in 2020 may reflect ongoing prevention successes in younger populations but also underscores the need to sustain and enhance targeted strategies to continue reducing infections in this age group.

These two figures in part 4 highlight the ongoing challenge of HIV prevention and underscore the critical importance of tailored interventions that address the unique dynamics and risk factors associated with different subpopulations and age groups.