# Calibration of Stochastic Models Using Likelihood Functions

In statistical modeling, one of the primary challenges is calibrating a model to accurately represent real-world phenomena, especially when the true outcomes are not directly observable. In stochastic models, this task often involves estimating the model parameters that maximize the likelihood of observing the data, given the model's assumptions. The process of model calibration ensures that the model’s predictions align as closely as possible with observed data, enabling more accurate inferences and predictions.

This section explores the general theory behind constructing likelihood functions in stochastic modeling. Specifically, we will focus on how likelihood functions can be used to bridge the gap between observed data, true but unobservable outcomes, and the model's simulated predictions. By accounting for errors and uncertainties in each of these components, we can develop a robust framework that helps to quantify and incorporate these uncertainties into the calibration process.

## Motivating Example: Disease Dynamics Model

In this example, we are evaluating the performance of a disease dynamics model by comparing projected number of diagnosed cases against the observed data. To assess the model's accuracy, we will consider two different scenarios as follow:

* In **Scenario 1**, the target is to match the number of diagnosed disease cases over a year, where the actual observed number is 6 cases, and the model predicts 4 cases. The absolute error in this case is 6 - 4 = 2 cases, and the relative error is calculated as 2/6 = 33%. At first glance, the absolute error might not seem terribly large, given that the model is missing only 2 cases. However, when considering the relative error of 33%, we see that this is a substantial portion of the target value.
* In **Scenario 2**, the target number of diagnosed cases is 600 cases, and the model prediction stands at 500 cases. In this case, the absolute error is 100 cases, which seems larger than in Scenario 1. However, the relative error is 100/600 = 16.7%, which is smaller than the 33% error in the first scenario. This suggests that after taking the larger scale of the numbers into account, despite the larger absolute error, the model’s prediction is more accurate. Thus, the model appears to perform better in this scenario in relative terms, even though the absolute error is larger.

**Incorporating Uncertainty**:

To make a more informed decision about the model's acceptability, we must also consider how the **uncertainty** in the **model projections** and **observed data** could influence our judgment about what constitutes a ‘good fit’.

1. **Uncertainty in observed data** can arise from several sources, such as poor data quality, reporting inconsistencies, and/or measurement error. The underlying level of uncertainty in the data allows us to assess the fit of the model in light of that uncertainty. In theory, the more uncertainty is incorporated into the underlying data inputs, the more cautiously we interpret deviations between the model and the observed data—acknowledging that some of the discrepancy may be due to noise rather than model misspecification.

We will now explore two assumptions about the measurement error in the observed data to explore how they affect our evaluation of the model’s predictions.

* In **Scenario 1**, if we assume that the observed 6 cases in scenario1 have a **10% measurement error**, this means the true number of cases could realistically fall within a range of 5.4 to 6.6 (since 10% of 6 is 0.6). In this case, the model’s prediction of 4 cases is not within the uncertainty range of the observed data, which is 5.4 to 6.6. The absolute error of 2 cases might seem small, but given the narrow range of uncertainty (±0.6), this difference could be considered too large to accept the model’s performance. Therefore, in this scenario, we might reject the model or at least consider improving it.
* However, if we assume a **50% measurement error** in the observed 6 cases, the true number of cases could realistically fall between 3 and 9 (since 50% of 6 is 3). In this case, the model’s prediction of 4 cases is comfortably within the range of 3 to 9. Given the wide uncertainty range, the absolute error of 2 cases is much more acceptable. Therefore, with 50% measurement error, the model’s prediction would likely be considered acceptable.

1. **Uncertainty in model projections** stems from the stochastic nature of simulation models, where random seeds are used to generate a different array of events in each run. This variability introduces random fluctuations in the model’s outcomes. To quantify and assess this stochastic variation, we can run multiple simulations (replicates) of the model and compute summary statistics such as the mean and standard deviation. In our example, we will consider two scenarios where the standard deviation (sigma) is evaluated at 10% and 50% of the predicted mean.

* In **Scenario 1**, if we assume a standard deviation of 10% around the model's predicted mean of 4 diagnosed cases, we can quantify the uncertainty in the model’s output. If the simulated outcomes follow a normal distribution, we can use this standard deviation to calculate a confidence interval for the model's predictions using the following formula:

Where is the predicted mean, is the Z-value for the desired confidence level (e.g., for 95% confidence, Zα/2≈1.96), σ is the standard deviation, and n is the sample size.

* Assuming a 95% confidence level, we can compute the confidence interval of 4 ± 0.8 cases for new diagnoses. Since the target of 6 diagnosed cases falls outside the model's 95% confidence interval (3.2 to 4.8), it indicates that the model's prediction doesn't adequately capture the observed data. While the model is close to the target, its uncertainty range suggests it's not a good fit, as the observed value is higher than the predicted range.
* However, if we assume a standard deviation of 60% around the model's predicted mean of 4 diagnosed cases, the 95% confidence level will range from 1.6 and 6.4 diagnosed cases. In this case, the target of 6 diagnosed cases falls at the upper end of the model's 95% confidence interval. The uncertainty in the model's prediction (as indicated by the wider confidence interval) suggests that, while the model may not be an exact match, it is a more acceptable fit given the larger variability in the predictions.
* Furthermore, we can use this information to calculate the percentage of times that model predicts the target of 6 diagnosed cases within a fix margin of error (+-1) as follow:

;

* suggesting that the model predicts between **5 and 7 diagnosed cases approximately 23.2%** of the time.

In conclusion, the decision to accept or reject a model based on its prediction accuracy is not solely determined by the mean behavior but must also consider the uncertainty in the observed data and model projections. These examples prompt some key questions about calibration that we aim to discuss in the following sections:

* What metrics should be used to assess the model’s fit to observed data?
* How should uncertainty in observed data (e.g., due to measurement error or reporting biases) be quantified and incorporated into calibration?
* How can model uncertainty (e.g., structural assumptions, parameter uncertainty) be accounted for in assessing model fit?
* What threshold or criteria define an “acceptable” fit given the level of uncertainty?
* Should all data points be weighted equally during calibration, or should some carry more weight based on reliability or relevance?
* How can we distinguish between model misspecification and data noise when fit is poor?
* What role does expert judgment play in interpreting calibration results under uncertainty?
* How sensitive are calibration outcomes to assumptions about data quality or error distribution?
* What tools or frameworks (e.g., Bayesian calibration, probabilistic sensitivity analysis) are appropriate for incorporating uncertainty in calibration?

## A Likelihood Based Inference Approach to Calibration

In the context of model calibration, a likelihood-based inference approach involves adjusting the model parameters to maximize the likelihood function, which measures how well the model explains the observed data. Grounded in statistical theory, this approach is widely used in both frequentist and Bayesian frameworks, and involves three key steps as follow:

### Step 1) Definition of a likelihood function

Likelihood functions provide a generalizable framework to incorporate uncertainties in data and model predictions by quantifying the probability of observing the given data under different model assumptions and parameter values. In mathematical terms this is expressed as follows:

=

where represents the observed data (measured or collected), θ represents the model parameters of interest, is the probability (for discrete data) or probability density (for continues data) of observing of observing given , is the likelihood function and is the maximum likelihood estimate of the parameters.

In **stochastic models**, where outputs are random variables, the likelihood function is constructed using the **probability distribution of the model’s outputs conditional on model parameters**. As such, the goal of likelihood-based inference is to find the parameter values θ that maximize this likelihood function—i.e., the parameter set that makes the observed data most probable under the modeI.

Let:

Then:

=

where Y is the stochastic model output (e.g., a random variable), is the probability distribution of model output y conditional on input values .

#### Incorporating uncertainty in observed data into likelihood

In most modeling contexts, the true outcome is not directly observable. Instead, we rely on observed data to infer the truth. Differences between the true values and the observed data can arise from various sources of error. In epidemiological data, potential sources of error include:

* **Sampling Variability**: Data may be drawn from different sample populations or regions, which may not perfectly represent the entire population.
* **Reporting Inconsistencies**: Data from different sources (e.g., national vs. local health departments) may not always align due to differences in reporting standards or practices.
* **Measurement Error**: Errors may occur during data collection or recording, such as incorrect diagnoses or misclassification of disease stages.
* **Model Assumptions**: Assumptions made in projecting or calculating the data, such as assumptions about disease progression, can also introduce error.

These potential errors in capturing and reporting observed data can be translated into measures of uncertainty, which are incorporated into the likelihood model. To account for this uncertainty, we assume the observed data  as being drawn from a probability distribution centered around the true (but not observed) value with an error term capturing the observation noise or uncertainty. The relationship can be expressed as:

∣()

where () is the probability distribution (e.g., Poisson, normal) used to model the likelihood of the observed data and  is the variance or error term characterizing the uncertainty in observations.

**Estimating the observation error:** Epidemiological surveillance data are often presented with uncertainty ranges, such as confidence intervals, to account for the variability or unknown factors involved in data collection. These ranges provide a way to measure and communicate the level of uncertainty in the observed data. When confidence intervals are reported, we can estimate the observation error by using the half-width of the interval. Assuming that this interval is approximately symmetric and based on a normal distribution, the standard deviation can be estimated as:

Where and are the upper and lower bounds of the confidence interval, is the z-score corresponding to the desired confidence level (e.g., z=1.96 for 95% a confidence interval).

In cases where confidence intervals are not reported, we can estimate the potential uncertainty in the underlying data using various methods. One such technique, particularly in the context of epidemiological surveillance reports, involves looking for comparable estimates of the value of interest from multiple sources. While these sources may not align perfectly, the variations in their reported values can be used to characterize the uncertainty in the data.

**Example:** Reported number of annual HIV diagnosis is one of the key calibration targets for HIV transmission models aiming to represent the true (but not observed) number of new infections that arise in a year. HIV diagnosis data are primarily collected through national and local surveillance systems. In the US, health care providers and laboratories are required to report confirmed HIV diagnoses to the state and local health departments, which then share this data with the CDC. CDC compiles and standardizes this information, ensuring consistency, and performs quality control before publishing it. Given the observed nature of this data, estimates are not accompanied by confidence intervals. in such case, we can rely on multiple sources reporting this data to estimate the “true” number.

Specifically, we rely on capturing the HIV diagnoses data from two sources: 1) CDC AtlasPlus (formerly NCHHSTP Atlas) is a platforms that CDC uses to disseminate HIV/STI data. AtlasPlus is an interactive tool that provides access to HIV diagnoses data (among other outcomes) over time and at different levels of stratifications (by age, race, location, etc), 2) local health department publish annual HIV surveillance summaries. Difference in reported data from these sources can arise due to different reasons, such as duplications, reporting completeness, timeline, and case definition. If we rely on the CDC reported data as the more reliable data, we can model data from local health department as being drawn from a distribution centered on the CDC estimate, with an error term capturing the discrepancies:

Where is the number of diagnoses reported by a local health department, is the CDC’s estimated (treated as the reference value), and captures the discrepancy between these two sources.

hile local health departments may report similar data. In such cases, we can assess the likelihood that the data from each source reflects the "true" number of diagnoses. If we assume that the CDC data is more reliable, we can model the observed data from other sources as being drawn from a distribution centered around the CDC's estimate, with σ2 capturing the variability or error between the sources.

## Linking Observe to Simulated Data

To construct a likelihood function, we aim to describe the probability of observing the true data based on simulated outcomes. In the case of count data, we typically assume the underlying distribution is Poisson. The relationship between the true data and the simulated outcome from a Poisson distribution is given by:

y′∣M=Poisson(M)

where M is both the mean and variance of the Poisson distribution.

## Impact of Sample Size and Error on Likelihood in the Poisson Distribution

The Poisson distribution inherently weights deviations from the mean based on its value. This is important when interpreting errors. Let’s consider two examples to illustrate this:

* **Example 1**: If the target number of diagnosed cases is M=6, then:σ2=6 and σ≈2.

In such case, the relative error (error ratio) is σ/M=2/6≈33%

* **Example 2**: If the target number of diagnosed cases is M=600, then:σ2=600 and σ≈24.5

The relative error in this case is σ/M=24.5/600≈4%

As the number of events (or cases) increases, the relative error decreases, which is a characteristic of the Poisson distribution: estimating larger numbers of events comes with less uncertainty compared to smaller numbers. When the true value is low, the relative error is larger, indicating greater uncertainty in the estimate, and the likelihood function is more sensitive to differences between the observed and true counts. In contrast, when the true value is high, the relative error is smaller, meaning there is less uncertainty, and the likelihood function is less sensitive to deviations between the observed and true counts.

### Approximating Poisson with Normal Distribution

For large values of M, the Poisson distribution becomes highly skewed and heavy-tailed, making it computationally inefficient and difficult to calculate probabilities or likelihoods. This is particularly problematic when using the Poisson probability mass function (PMF) for large values of M.

**Central Limit Theorem and Normal Approximation:** As the mean M increases, the Poisson distribution becomes more symmetric and resembles a normal distribution. This is a well-known result from probability theory and is useful for simplifying computations when M is large.

According to the **Central Limit Theorem (CLT)**:

Y′∼Poisson(M)≈N(M,M)

where N(M,M) denotes a normal distribution with mean M and variance M. This approximation becomes more accurate as M increases.

The normal approximation to the Poisson distribution is generally valid when MM is sufficiently large. A common rule of thumb is that the approximation works well when M≥10. For smaller values of M, the Poisson distribution remains more skewed, and the normal approximation may not accurately reflect the characteristics of the data. In these cases, it is better to use the Poisson distribution directly.

## 9. Putting It All Together

Combining the concepts of error in the observed data and likelihood construction, we can summarize the key relationships:

The relationship between observed data y and true values y′ can be stated as:

y∣y′∼N(y′,σ2)

The relationship between the true values y′ and simulated data M can be stated as:

y′∣M∼N(M,M)

Using this, the likelihood function given M and observed data y can be stated as:

L(θ,M∣y)=P(y∣θ,M)∼N(M,σ2+M)

This framework allows for constructing likelihood functions that capture the relationship between the model parameters and simulated outcomes (θ,M), and observed data (y), accounting for errors and uncertainties in the underlying data and the simulation process.

## Step 3) Estimation of Parameters

Use optimization techniques to find the parameter values that maximize the likelihood function—i.e., those that make the observed data most probable under the model. This process yields the maximum likelihood estimates (MLEs) of the parameters.

# ESTIMATING LIKELIHOOD ERRORS

Overall, the likelihood is represented by a joint likelihood with seven components by data type. Each likelihood follows a multivariate normal distribution (as an approximation to a binomial distribution) where the observed data is centered at some true value (unknown) with some level of measurement error; and the true, unknown value is centered at the model-generated estimate with some level of model error.

* Measurement error is determined based on the source of the data type . Model error assumes a normal approximation to a binomial distribution, where the variance of the model estimates equals the mean estimated value. Both measurement error/bias and model estimates are allowed to be correlated: we allow measurement error/bias to be correlated over time (e.g., the error in reported incidence for 2009 is correlated to the error in reported incidence for 2010), and model estimates to be correlated due to overlapping strata
* Correlations are represented using an estimate of standard deviation and a correlation matrix with either an auto regressive or compound symmetry structure. An auto regressive structure assumes correlations between one year apart are stronger than those 10 years apart, for example, and is used for data over longer periods of time. A compound symmetry structure assumes the correlation between one year apart is the same as the correlation between 10 years apart and is used for data over shorter periods of time.”

## Theory

The two main assumptions allow us to model the uncertainty in both the reported data and the simulated outputs.

**Assumption 1: Measurement Error in Reported Data**

his assumption states that the reported values, ​, are an approximation of the "true" values, , with some measurement error. This is modeled as a multivariate normal (MVN) distribution:

Where

* is the reported value (e.g., CDC reported number of diagnosis),
* is the “true” underlying value
* is the measurement error covariance matrix

However, the complication is that data is often not reported at the finest granularity across all dimensions (e.g., age and race combinations), but instead aggregated over certain dimensions (e.g., total counts by age or by race, but not both). This creates the need for a transformation to account for the aggregated reporting.

Example: Consider four groups:

* Young Black (YB)
* Old Black (OB)
* Young White (YW)
* Old White (OW)

The true number of diagnoses is represented as while the reported data is aggregated along the margins (age and race separately), so: where  and are aggregated over age (young and old), and ​ and  are aggregated over race (black and white).

To transform the true data into the reported data format, we use a transformation matrix M, which maps the more detailed data  into the coarser, aggregated data ​. The matrix MM would look like:

**Summary of Assumption 1:**Thus, the reported data ​ is modeled as: Where M is the transformation matrix that maps the true, detailed data to the reported, aggregated data .

**Assumption 2: Simulation Error in Model Output**

The second assumption deals with the simulated values from the model. We assume that the simulated values, , approximate the true underlying values data , but with some simulation error . This is also modeled as a multivariate normal distribution: 

Where:

* : The simulated value from the model.
* : The simulation error covariance matrix.

**Combining Assumptions 1 and 2**

When we combine these two assumptions, we aim to model the reported data ​ as a function of the simulated values , considering both the measurement error and the simulation error

This equation tells us that the reported data ​ follows a multivariate normal distribution with a mean  (the simulated values transformed into the reported data format) and a combined error covariance , which accounts for both measurement error and simulation error.

## Components Required for Full Error Estimation:

To estimate the total error between the simulated data  and the reported data ​​, we need three components:

### Transformation Matrix M:

This maps the detailed simulated data into the aggregated form of the reported data. The matrix M is constructed based on the comparison of dimensions between the simulated and reported data (as mentioned, this is estimated using ontology codes or other metadata).

### Simulation Error Γ:

This error represents the uncertainty in the simulated values. For frequency outcomes (like counts of new diagnoses), the assumption is that the simulation error follows a Poisson distribution, where the variance equals the mean.

This reflects the fact that the variance for each outcome is proportional to the mean value of the simulation for that outcome.

**Why Poisson?** The Poisson distribution is typically used to model the number of occurrences of an event within a fixed period or population size. A key property of the Poisson distribution is that the mean and variance are equal: Var(y)=E(y)

This is useful when dealing with frequency data, especially when there is a mean-variance relationship. For example, if the simulated number of new diagnoses for a particular group is expected to be , then the standard deviation is also  ​, which matches the behavior of real-world count data where the variability grows with the magnitude of the counts.

The Poisson distribution assumes that successive draws (e.g., counts of new diagnoses from one year to the next) are independent of each other. This assumption is reasonable in many epidemiological models where the occurrence of new cases in one year or group doesn't directly affect the number of cases in another year or group.

## Measurement Error Σ:

This captures the uncertainty in the reported data, such as inaccuracies in CDC-reported diagnoses. Σ is typically estimated from external data or through expert judgment.

It’s easier to break covariance matrix into its two main components—**standard deviations** and the **correlation matrix**

### Component1: Standard Deviation:

The  for each dimension gives us an idea of the error for that dimension

We can estimate SD in multiple ways:

**Option1) Use of External Data (Direct Estimation):** Some external datasets or validation studies report measurement errors and we can translate that directly to SD

For example, the Census reports a 3% error in their population size estimates

Should we assume , and therefore ? The problem with this approach is that if or , the error remains the same

We could assume ? This translate to , which is the coefficient of variation.

We could also assume , and say if q= 0.5, this means:

Todd has tried these combinations and concluded that CV was the best one for population count.

**Option 2) Use of historical data from the same source:** this is true for situation where historical data is corrected in a future publication and we have 2 versions of the same data. For example, CDC published new HIV diagnosis counts each year but the values that are later reported in the HIV atlas are different from the original reports.

See calculating\_error\_terms\_for\_ehe\_likelihoods.R for an example

### Component2) The correlation matrix

Correlations are derived from the covariance matrix by normalizing each covariance by the product of the standard deviations of the related variables. This provides a standardized measure of the relationship between errors across different dimensions.

The correlation captures how measurement errors in one dimension are related to one another. For example:

* If the reported values are inaccurate in **year 1**, how likely are they to also be inaccurate in **year 2** and **year 3**?
* If the values are incorrect for the **Black race**, to what extent are they also incorrect for **White** or **Hispanic** populations?

In practice, we typically assume that measurement error correlations exist **over time** (i.e., errors in different years are correlated), but not across other strata such as **age, race, or sex**.

There are 2 typical formats for correlations that are applicable here:

1. **Autoregressive (AR) Correlation:** In an AR(1) process, correlation between consecutive years (or time points) decreases exponentially as the time gap increases.

Where ρ is the correlation coefficient between successive time points.

1. **Compound Symmetry (CS):** All pairs of observations within the same group (e.g., race, age group) have the same correlation. This is often used for longitudinal data or repeated measures within a group. Constant correlation across all pairs of observations in the same group.

Where ρ is the constant correlation between all pairs of observations.

In reality, a hybrid approach is often needed for modeling error correlations, depending on the data structure and available information. For example, when we have **limited data points** (e.g., 10 years of observations), a **compound symmetry (CS)** model is frequently used. This assumes that the error is similarly distributed across all data points. In such cases, if we know that the reported values were off by 10% in 2000, it would be reasonable to assume that they were also off by around 10% in 2010. This simplifies the correlation structure, making it easier to model with limited data.

However, when working with a **larger dataset** spanning a **longer time horizon**, an **autoregressive (AR)** model may be more appropriate. This model assumes that errors in consecutive time periods are more strongly correlated, but this correlation weakens over time. For example, if the new diagnosis data in 2000 were off by 10%, an AR model would suggest that errors in the following years (e.g., 2001, 2002) are similarly off, but the impact would diminish over longer periods (e.g., by 2010).

# Capturing Misclassification Error for Syphilis Diagnosis

In this model, we aim to simulate the classification process for diagnosing early latent (EL) and late latent (LL) syphilis, while considering the misclassification errors that may arise during diagnosis.

We define the following parameters:

* μ1: Simulated number of early latent (EL) syphilis diagnoses.
* μ2: Simulated number of late latent (LL) syphilis diagnoses.

Let z1 and z2 represent the true number of EL and LL syphilis diagnoses, respectively.

Assuming the simulation model is unbiased, the true number of diagnoses can be approximated using a Poisson distribution. Specifically, the true number of diagnoses follow a Poisson distribution with predicted means μ1 and μ2 as follows:

* Z1∼Poisson(μ1)
* Z2∼Poisson(μ2)

This means that the number of diagnoses for each group (EL and LL) is randomly distributed and occurs independently, with a constant average rate of occurrence, as described by the Poisson distribution. If the model is unbiased, the expected (average) number of simulated diagnoses will closely match the true population values. The Poisson distribution is well-suited for modeling rare events like syphilis diagnoses because its mean equals its Variance and assumes that events occur randomly and independently.

## Misclassification Probabilities

In this model, we assume that the probability of correctly classifying each stage of syphilis (EL and LL) is known:

* p1: The probability that an EL case is correctly classified as EL.
* p2: The probability that an LL case is correctly classified as LL.

We define the following:

* m1: The number of EL diagnoses correctly classified as EL.
* m2: The number of LL diagnoses correctly classified as LL.

Given these definitions, we model the number of correct classifications for each stage using a binomial distribution:

* m1∼Binomial(z1,p1)
* m2∼Binomial(z2,p2)

The number of EL diagnoses correctly classified as EL follows a binomial distribution, where z1 is the true number of EL diagnoses and p1 is the probability of correctly classifying an EL case. Similarly, the number of LL diagnoses correctly classified as LL follows a binomial distribution, where z2 is the true number of LL diagnoses and p2 is the probability of correctly classifying an LL case.

This setup allows us to model the number of correct classifications while accounting for the misclassification probabilities of each stage.

## Observed Diagnoses and Reporting

Let O1 and O2 represent the observed numbers of reported early latent (EL) and late latent (LL) syphilis diagnoses, respectively, accounting for misclassification.

Based on the earlier definitions, we can express the observed diagnoses as follows:

o1 (Observed EL cases): The true number of EL cases correctly classified as EL (m1) plus the true number of LL cases misclassified as EL (z2 - m2):

* o1=m1+(z2−m2)

o2 (Observed LL cases): The true number of LL cases correctly classified as L (m2) plus the true number of EL cases misclassified as LL (z1 - m1):

* o2=m2+(z1−m1)

### Reporting in Surveillance

The final reported values in the surveillance report for each stage are assumed to follow a normal distribution with means o1 and o2, respectively, and a common measurement error Variance, σ². Specifically:

* y1∼N(o1,σ2)
* y2​∼N(o2​,σ2)

This final model accounts for both misclassification in the diagnosis process and the measurement error in reporting syphilis cases in surveillance systems.

## Calculating the Conditional Likelihood Function

To capture the likelihood of observing the reported syphilis diagnoses, we need to account for five key components that reflect both the true values and the misclassification errors. The components are as follows:

**Expected Value of Y1 & Y2**

We start by calculating the expected number of observed early latent (EL) diagnoses, **E(y1 | p1, z1)**, which is the expected number of **m1** (true EL cases correctly classified) plus the number of true LL cases misclassified as EL (**z2 - m2**).

E(y1∣p1,z1)=E(m1+z2−m2)=p1z1+z2−p2z2=p1z1+(1−p2)z2

And similarly we can conclude:

E(y2∣p2,z2)=E(m2+z1−m1)=p2z2+z1−p1z1=p2z2+(1−p1)z1

**Variance of Y1 & Y2**

Next, we calculate the Variance of the observed EL diagnoses, **Var(y1 |**∣p1,p2,z1,z2**)**. This Variance accounts for the Variance due to misclassification and measurement error. We break it down as follows:

* Var(y1∣p1,p2,z1,z2)=Var(y1∣o1)=Var(o1∣p1,z1)+σ2

Here, **σ²** is the measurement error Variance.

Expanding the Variance of **o1**:

* Var(o1∣p1,z1)=Var(m1+z2−m2)=Var(m1)+Var(z2)+Var(m2)+2Cov(m1,z2)−2Cov(m1,m2)−2Cov(m1,z2)

Since m1 and m2 follow a binomial distributions, we can follow:

* + Var(m1​)=z1​p1​(1−p1​)
  + Var(m2)= z2p2(1-p2)

Because z2 is a constant, **Var(z2)** = **Cov(m1, z2)** = **Cov(m2, z2)** = 0, And because m1 and m2 are independent, **Cov(m1, m2)** =0

So it follows:

* + Var(y1∣p1,p2,z1,z2)=z1p1(1-p1)+ z2p2(1-p2) +σ2

Using a similar process, we can compute:

* + Var(y2∣p1,p2,z1,z2)=z1p1(1-p1)+ z2p2(1-p2) +σ2

**Covariance y1 and y2**

Cov(y1,y2|p1,p2,z1,z2)= Cov(m1+z2−m2,m2+z1−m1)

Using the properties of Covariance:

Cov(Y1,Y2)=Cov(m1,m2)+Cov(m1,z1)−Cov(m1,m1)+Cov(z2,m2)+Cov(z2,z1)−Cov(z2,m1)-Cov(m2,m2)-Cov(m2,z1)+Cov(m2,m1)

=-Cov(m1,m1) −Cov(m2,m2)

= -Var(m1) – Var(m2) = -z1p1(1-p1) – z2p2(1-p2)

**Conditional Likelihood**: The likelihood function is constructed by combining the conditional expectation and Variance components calculated above, as well as incorporating the measurement error. The likelihood will typically involve normal distributions due to the assumption that the observed data (Y1, Y2) are normally distributed with their respective means and Variances.

The final likelihood function can be represented as:

TBD

## Numerical example

Peterman, et. al (2005) review records in 6 US jurisdictions to determine if reported cases met the Centers for Disease Control and Prevention case definitions. Table 1 reports the stage of syphilis as classified by the sites and reclassified by the authors using the CDC case definition. Assuming the reclassified cases represent the ‘true’ distribution of stages, we use this data to calculate the number of records correctly-classified and those misclassified by stage and estimate the misclassification error for the EL and LL/Unknown stages

SHIELD> inputs> input\_misclassification\_error.R

# incorporating incompelte reporting into MSA estiamtes

The Shield model includes the top 40 U.S. Metropolitan Statistical Areas (MSAs) with the highest syphilis diagnosis burden from 2020 to 2025. Each MSA consists of multiple counties, with surveillance data primarily reported at the county level, segmented by age, sex, race, and year. To build aggregate estimates for the MSAs modeled in Shield, we consolidate the county-level data. For case counts (e.g., number of diagnoses), this is typically done by weighting county-level estimates according to population size and simply adding the data. For proportions (e.g., proportion receiving PrEP), we weight the county-level estimates by population density, which represents the denominator for the given proportion (e.g., the number of people eligible for PrEP in each county).

In some cases, estimates for certain counties within an MSA may be missing. In such cases, we have generally relied on two strategies in the past:

1. **Limiting MSA-level estimates to those capturing 100% of counties:** For example, if a county-level estimate is missing, the MSA estimate is set to "NA" (not available).
2. **Allowing MSA-level estimates to represent any available county-level estimate:** In this approach, the MSA-level estimate is calculated based on the available county-level data, even if some counties are missing estimates.

We now consider a third option to report a **completeness index**, representing the **proportion of the population captured among counties with available data**. This approach will provide an estimate of how much of the population within an MSA is accounted for, based on the counties with available data, offering a more nuanced understanding of data completeness in cases where some county-level estimates are missing.

We can use this index to incorporate additional error that may be introduced into our aggregate estimate by incorporating it into our likelihood function. This index represents the additional uncertainty in the MSA-level estimate due to missing county-level data. If the index is 100%, it indicates that the entire population within the MSA is captured by the available data, meaning there is no additional uncertainty. As the index reduces toward 0%, the uncertainty increases, reflecting the unknown characteristics of the population missing from the available data in that MSA. This helps us quantify and account for the uncertainty associated with incomplete data in the modeling process.

Let's assume rr is the quantity of interest at the MSA level, with several counties. The estimate for rr is derived as a weighted mean of county-level estimates, with data on both known and unknown counties.

r=wknown⋅rknown+wunknown⋅runknown

Where:

rknownis the estimate of r from counties with available data.

runknown​ is the estimate of r for the counties with missing data (often assumed to be unknown or imputed).

wknown​ is the weight for counties with available data.

wunknown​ is the weight for counties with missing data.

We now compute the expected value and standard deviation of r based on known and unknown values.

**Expected Value of r:** To compute the **expected value** of r, we can express the **conditional expectation** of r, given that we know known​ for the counties with available data.

E(r∣rknown)=wknown⋅E(rknown∣rknown)+wunknown⋅E(runknown∣rknown)

Where:

* E(rknown∣rknown)=rknown​, because for the counties with available data, the expected value of r is just the observed value.
* E(runknown∣rknown)=rknown, assuming that the expected value of r for the unknown counties can be approximated by the estimate from the known counties (or we might assume some form of similarity).

So, the expected value simplifies to:

E(r∣rknown)=wknown⋅rknown+wunknown⋅rknown =rknown​

Thus, the expected value of r at the MSA level is simply known​, which is the weighted average of the county-level estimates, with the same value for both known and unknown counties

**Sandard Deviation of r:** The **standard deviation** of r would depend on the variance of the known and unknown county-level estimates.

Var(r)=wknown2⋅Var(rknown|rknown)+wunknown2⋅Var(runknown|rknown)+Cov(rknown,runknown|rknown)

Where:

* Var(rknown|rknown)=0); the variance of known values is zero because the estimate is already given for the known counties
* Var(runknown|rknown) is the variance of the unknown counties.
* Cov(rknown,runknown|rknown) =0; the covariance between the known and unknown values is zero, assuming independence between the two

Substituting the assumptions:

Var(r)=wunknown2​⋅Var(runknown​∣rknown​)

# CALIBRATION TARGETS

## Demographic model

* **Basic Likelihood:** This applies when data at the desired location level is available.
* **Handling Data Gaps:** At times, calibration data may not be available for a specific location of interest. For example, if we need to calibrate the proportion of awareness in Baltimore but lack local data, we could use data from Maryland and incorporate some uncertainty to account for differences between these locations.

**Targets for the Demographic Model:**

1. **Population Size:** Available from 2010 to 2020, broken down by age, sex, and race.
2. **Deaths:** Total count available for two periods—2001 to 2010 and 2011 to 2020.
3. **Births:** Available from 2007 to 2023, broken down by the age and race of mothers.

# SHIELD components

The specification file provides the blueprint for the core model. definition of compartments and all the transitions in the model. The figure shows how it connects to other scripts required for running an engine test.

A diagram of a system

Description automatically generated with medium confidence

**Fig: High-level review of the SHIELD engine and scripts**

The specification file is organized as follow:

## Initial setup:

**SHIELD.SPECIFICATION:** defining the blueprint of compartments and various strata (age, race, sex, etc).

* The model represents the U.S. population through three distinct sex and sexual behavior categories: **heterosexual\_male**(males engaging in heterosexual sex), **msm** (men who have sex with men), and **female** (females, which may include both heterosexual and bisexual behaviors). These categories are further divided by race and ethnicity into three groups: **Black**, **Hispanic**, and **Other**, allowing for the analysis of differences in risk behaviors and infection transmission dynamics by demographic factors. The population is also subdivided into 11 age groups: 0-15 years, 16-20 years, 21-25 years, 26-30 years, 31-35 years, 36-40 years, 41-45 years, 46-50 years, 51-55 years, 56-65 years, and 65+ years. These age categories enable the model to track changes in risk and disease progression over the course of an individual's life.
* The population is further categorized into two main groups: **infected** and **uninfected**. The infected group is divided into two primary categories: the **continuum of infection**, which includes individuals who are either **undiagnosed** or **diagnosed but untreated**, and the **disease stages**, which represent the progression of infection through different phases: **ps** (primary or early stage), **el** (early latent stage), **ll** (late latent stage), and **ter** (terminal or advanced stage). This distinction allows for the modeling of the progression of infection from undiagnosed to untreated to advanced stages of disease.
* For the **uninfected** population, the model defines two key states: **susceptible**, representing individuals who are at risk of infection but have not yet been diagnosed, and **diagnosed.treated**, representing individuals who are diagnosed and receiving treatment for conditions related to HIV/STI. This helps differentiate between at-risk individuals and those who are under care, capturing both the uninfected and treated segments of the population.
* The model starts in 1940 and projects the population through to 2040, providing a dynamic view of infection dynamics, treatment, and transmission risks across time. This time span allows the model to simulate long-term trends and shifts in the population’s health status, helping to understand the impact of historical and future interventions on public health.

A diagram of a patient's process

Description automatically generated

## Simulation modules

The specification file includes several modules that define the composition of the initial population, as well as the various transitions that occur over time. These transitions are crucial for modeling the dynamics of the population and the spread of syphilis, with each module describing a different aspect of the population's health, behavior, and demographic changes.

* **Initial Population Composition:** The initial population composition is defined based on demographic characteristics such as age, sex, race/ethnicity, and sexual behavior. This underlying population is informed by U.S. census data, with full demographic composition only available from 2010 onward. As such, we assume a fixed population size and composition up to the year 2010. Capturing earlier periods prior to 2010 helps us reflect the syphilis epidemiology more closely, while after 2010, we allow for the natural evolution of population size and composition.
* **Births and Deaths:** The model captures demographic transitions related to births and deaths, which are key factors in shaping the population's size and structure over time. Births are modeled based on fertility rates, while deaths occur due to a combination of natural causes and disease progression. Deaths in the infected population, particularly those in the late latent and terminal stages of syphilis, are critical for modeling disease outcomes. These transitions ensure that the model reflects the reality of population turnover, with individuals entering and exiting the population through both natural processes and disease-related mortality.
* **Differential Aging Rates:** These rates are included in the model to reflect how the population ages at different rates across various demographic groups. Aging is modeled as a gradual process, where individuals transition between age categories over time. The rate of aging may vary by race or health status. This factor is crucial for capturing the evolution of the population's composition over time
* **Sexual Transmission:** One of the primary transitions modeled is sexual transmission, which describes how syphilis and other STIs are transmitted between individuals. This includes transitions between the uninfected and infected states (e.g. from susceptible to undiagnosed or diagnosed.untreated, and vice versa). Sexual transmission is influenced by behavior (e.g., whether individuals engage in heterosexual or MSM sexual activities) and demographic factors, including age and race/ethnicity. This dynamic allows the model to capture how infection spreads through the population over time, particularly within different sexual networks and risk groups.
* **Syphilis Continuum of Care:** The syphilis continuum of care defines the transitions that individuals with syphilis undergo as they progress through the stages of infection and treatment. This includes transitions from undiagnosed to diagnosed, from diagnosed to treated, and from untreated individuals progressing through various disease stages, including primary/early, early latent, late latent, and terminal. The model tracks these transitions to simulate the impact of treatment and diagnosis on the course of the disease, as well as the overall burden of syphilis in the population. This structure is critical for assessing how interventions at different points in the continuum—such as testing, treatment initiation, and follow-up—can affect disease outcomes and reduce transmission.

Together, these transitions provide a dynamic framework for modeling the spread of syphilis and other STIs in the population, as well as the health outcomes of different demographic groups. By capturing sexual transmission, births, deaths, aging, and the continuum of care, the model simulates the long-term effects of public health interventions and demographic changes. This helps in offering insights into how to reduce syphilis prevalence and improve population health outcomes over time

## Parameters

The specification file includes instructions for describing transitions between different compartments in the model, which are often represented by rates. In the simplest case, the rate of an event is fixed over time and informed by literature. However, in more advanced scenarios, the rate can be dynamic: changing over time and differentiated by factors such as age, sex, race, or other variables (i.e., it has a functional form). This allows the model to capture more complex behaviors and interactions within the population.

JHEEM uses a distinct construct for defining parameters in the model:

* **Elements:** These are *scalar* values or *functional forms*. They represent the basic building blocks of the model and can be constants or dynamic functions that describe various processes (e.g., infection rates, mortality rates). Elements can be used directly or combined to define more complex quantities.
* **Quantities:** These are more complex constructs that can represent combinations of elements and other quantities. A quantity can be equivalent to another quantity or an element, or it can be an expression or function of other quantities. Quantities are used to inform the model's compartments and the transitions between them, allowing the model to capture relationships between different variables in a more nuanced way.

Together, elements and quantities provide a flexible framework for modeling the dynamic relationships between different factors, enabling the specification of rates, transitions, and compartments that reflect real-world epidemiological processes.

## Outputs

There are two primary methods to capture outputs in the model: **Compartment Outputs** and **Transition Outputs**. These methods allow for tracking different aspects of the model's dynamics and assessing the impact of various processes on the population.

1. **Compartment Outputs:** Compartment outputs refer to the frequency values of specific compartments, such as the population size in each compartment. These outputs can be captured in different ways depending on the desired level of detail:
   1. **track.point.outcome():** This method captures a static outcome at a specific moment in time. For example, it could be used to track the number of individuals in a specific compartment at the start or end of a year (Jan 1st). This provides a snapshot of the population at a given point.
   2. **track.integrated.outcome():** This method integrates the point estimates over a specific time period, such as a year (Jan 1st to Dec 31st). It is more useful for calibration as it accounts for the average values across the entire timeframe rather than just a single moment. This method is preferred when looking for a more comprehensive measure of a compartment's population size or other attributes across time.

track.point.outcome(…, name='point.population', …)

track.integrated.outcome(…, name='population',

value.to.integrate = 'point.population',….)

1. **Transition Outputs:** Transition outputs capture the movement or event rates between compartments, representing the dynamics of individuals moving from one state to another (e.g., from susceptible to infected, or from diagnosed to treated). These outputs provide insight into how the population is changing over time due to various events and processes.
   1. **track.transition():** This method captures dynamic outcomes, indicating the movement of individuals between specific starting and ending compartments. It provides information on the flow between compartments, such as how many individuals transition from an undiagnosed to a diagnosed state over a specific period.
   2. **track.dynamic.outcome():** This method captures outcomes that account for individuals entering the model without specifying a starting or ending compartment. It is useful for tracking events that involve the population entering the model (e.g., through birth or migration) or when individuals do not fit neatly into defined starting or ending states.
   3. **track.cumulative.outcome():** This method sums multiple dynamic outcomes over time, providing an overall measure of the cumulative effect of transitions across different compartments. It is particularly useful for capturing the total impact of specific processes or events, such as the total number of individuals who have been diagnosed with syphilis over a set period.

track.**dynamic**.outcome(SHIELD.SPECIFICATION,

name='births.from',…)

track.**cumulative**.outcome(SHIELD.SPECIFICATION,

name='fertility.rate',

value=expression(births.from/population),….)

For dynamic transitions that change over time (e.g., testing), the anchor points are coded at the beginning of the year (e.g., if transmission changes from 2000 to 2020, these dates represent jan 1st of those years)

By using these output tracking methods, the model can provide valuable insights into both the state of the population at different times (compartment outputs) and the dynamics of transitions between compartments (transition outputs). These outputs are essential for understanding the long-term effects of interventions and changes in the model's parameters.

# Initial population

The SHIELD model begins in 1940. However, since census data on age, race, and sex is only available back to 2010, we need to select a baseline year before 2010 to set the population composition. Without data to inform earlier years, modeling dynamic changes would not add value here.

**Fixed Strata Year**: Birth data begins in 2007, and deaths are reported in two time frames, modeled as static parameters. This makes 2007 a suitable year for setting initial strata sizes. Thus, we start the model in 1940 but keep the population size and demographic composition fixed at 2007 levels until we introduce dynamic changes over time.

**Initial Population**: We model three subgroups based on the intersection of sex and HIV acquisition risk, including females, heterosexual males, and males who have sex with males (MSM). Census data provides the population sizes for females and males. To further segment the male population, we estimate the proportion of males who are MSM.

**Determining MSM Population Size**:

**Proportion of MSM in the Male Population**: We have two key data sources for estimating the MSM population size:

* **County-Level Data from Emory**: This provides estimates of the total proportion of males who are MSM, available at the county level in the Surveillance Manager. The basic approach is to compute a weighted average of these county-level proportions based on each county’s population size. This would give us an MSA-level estimate that we could apply uniformly across age and race groups. However, this method overlooks potential racial heterogeneity.
* **MSA-Level Data from BRFSS**: The Behavioral Risk Factor Surveillance System (BRFSS) provides race-specific proportions of males reporting same-sex behavior at the MSA level.

**Race-Specific MSM Risk Adjustment:** Using the BRFSS data, we first apply race-specific MSM proportions to the male population by race in each county to estimate an initial count of MSM males.

We then compare this initial estimate to the county-level MSM proportion from Emory and adjust the race-specific proportions to better fit the Emory data. This method ensures that the final estimates align with both the overall MSM proportion from Emory and the racial distribution from BRFSS, capturing both racial heterogeneity and county-level variation.

**Adjusting Race Categories**: There's a slight complication in using BRFSS data because the race categories differ from those in the Census. Here’s how we handle the discrepancies:

* **Flattening Race and Ethnicity**: We combine race and ethnicity into a single race category, resulting in five groups:

1. Hispanic
2. American Indian or Alaska Native
3. Asian or Pacific Islander
4. Black
5. White

This structure aligns more closely with Census data, where Hispanic is treated as an ethnicity, and we create a single "Asian or Pacific Islander" category to match the available Census groupings.

* **Combining Asian & Pacific Islander Groups**: In the BRFSS, Asians and Native Hawaiians/Other Pacific Islanders are reported separately. Instead of taking a simple average, which assumes equal population sizes, we apply a weighted combination based on an expected population ratio of 9:1 (Asian to Pacific Islander). This gives:

*’Asian or Pacific Islander’=0.9×(proportion MSM for Asian)+0.1×(proportion MSM for Native Hawaiian/Other Pacific Islander)*

This approach provides a more accurate reflection of the population composition in the "Asian or Pacific Islander" group.

# DEMOGRAPHIC MODELING

We model births + immigration into and deaths + emigration out of each “compartment”

* Migration (immigration and emigration) doesn’t depend on disease state, so it doesn’t change the disease prevalence in each compartment

## Births:

Birth rate= number of births/ population size

Fertility rate= number of births/number of female in childbearing ages

* JHEEM has used “birth rate” (although it’s called fertility rate in the code).
* SHIELD will use “fertility rate” to capture vertical transmissions more accurately

## Deaths

There are two components: general mortality and disease-specific mortality rates

Jheem uses state specific death rates, called metro-deaths, these are not reported directly.

To estimate the national-level or MSA-level mortality rate, we extract the counties within each MSA, map these counties to their corresponding states, and then take a weighted average of the state-level death rates to approximate the MSA-level mortality rate

Steps:

1. Extract county level death rates
2. Extract county level population count
3. Compute county weights as proportion of their population relative to the state or national population
4. Compute weighted average death rate

## Aging

In a compartmental model, the aging process represents individuals transitioning from one age group to the next as they age. The rate at which individuals "age out" of one compartment and enter the next is governed by the aging rate, which is often calculated as the inverse of the duration individuals typically remain in that age group. When individuals in a particular compartment reach the upper age limit, they transition to the next age group at this rate. For example, for each age group i (e.g., 0-15), the aging rate ai​ can be calculated by:

ai=1/duration in age group > 1/15 per year

**Issue**: Fix aging rates don’t allow for the calibration of models to historical or future trends in aging. These models don’t account for shifts in life expectancy, fertility, or mortality, and could result in inaccurate projections of the age distribution.

**Solution:** To address the limitations of fixed aging rates, you can incorporate **differential aging rates** that **vary by age group and time**

### Time-Dependent Aging Rates:

Define aging rates ai​(t) for each age group that can change over time. Rather than assuming a constant rate, aging rates can be modeled as functions of time to reflect evolving demographic factors like healthcare advancements or lifestyle changes.

1. **Number of parameters**? We need to determine the number of parameters required to model differential aging rates effectively. In our model:
   1. We have 11 age groups, requiring 10 independent aging rates. The model also considers 3 races and 3 sexes, totaling 90 parameters if we include interactions.
2. **Dynamic Aging Rate Modeling?** Should changes be linear, or should we use a spline-based approach? Generally, we use splines with knots in 2010 and 2020:
   1. 2010 is our base year for population calibration.
   2. 2020 serves as a recent benchmark to fine-tune projections for the baseline year (2024).
3. **Differentiating by Infection Status?** Whether to differentiate aging rates for infected versus uninfected populations depends on data availability.
   1. Uninfected Population: Census data on population age structure helps us calibrate aging among uninfected individuals.
   2. HIV-Positive Population: Estimated HIV prevalence by age allows us to capture aging patterns among the infected population.
   3. Syphilis: No prevalence estimates are available, so we lack the data to differentiate aging rates for syphilis.

**Aging Rates in Shield Model**: In the Shield model, we do not differentiate aging rates by infection type. To provide most flexibility, we include 90 parameters for all compartments \* 2knots=180 independent parameters.

# SEXUAL MIXING

## AGE Mixing Model Overview

The model for mixing considers three key components: age, sex, and race.

In general, we model contacts between the following groups:

* MSM (men who have sex with men) with other MSM, heterosexual men, or women.
* Heterosexual men with MSM, other heterosexual men, or women.
* Women with MSM or heterosexual men.

## Age Mixing Model

For the age component, Todd utilized data from a study in Australia that reported the ages of each pair of partners. Based on these data, the following model was assumed to represent differences in partner ages:

Age difference=N(μ,σ) where

μ=B0+B1×a

σ=L0+L1×a

This gives us:

Diff(a)=N(B0+B1×a,L0+L1×a)

Here, a represents the age of the individual, and B0,B1, L0, and L1 are coefficients estimated from the data.

Using this model, we estimate three separate age models for females, heterosexual men, and MSM. These are stored in PAIRING.INPUT.MANAGER$sex.age.models.

### Age Mixing Matrices

The next step is to compute the age mixing matrices, which represent the *proportion of contacts that occur between different age groups*. Since we have categorical age groups, we need to determine the proportion of the population that falls into specific age years. For example, within the age group 25-34, we calculate the proportion of individuals who are 25, 26, 27, and so on.

In a simplified case, we can assume a uniform distribution across ages within an age group, meaning that each year represents 1 tenth of the total population for that group. We can then estimate the proportion of contacts from a "mixture normal distribution."

For example, to estimate the proportion of contacts for women aged 25-34 that occur with individuals in the 13-24 age group:

P(13−24)=0.1×pnorm(13−24, μ25, σ25)+0.1×pnorm(13−24,μ26,σ26)+…+0.1×pnorm(13−24,μ34,σ34)P(13−24)

In JHEEM, Todd used a more sofisticated method for maping popualtion proportions based on census in each location

* get.heterosexual.male.single.year.age.counts()
* get.female.single.year.age.counts()
* get.msm.single.year.age.counts()

Using this approach, we estimate three age mixing matrices: one for females, one for heterosexual males, and one for MSM.

### Age of Sexual Debut and Availability

Additionally, we must model the reduction in sexual availability for the youngest and oldest age groups. This is handled using the get.sexual.availability() function, which maps changes in sexual availability across ages. The model reflects an increase in sexual activity starting from age 13, reaching 100% at ages 20 to 64, and gradually tapering off until age 85, the final age group.

### Calibration

All the parameters introduced so far are estimated from data and remain fixed. However, we include one additional parameter specifically for calibration—a multiplier applied to the standard deviation in the age model. This calibration parameter adjusts the variability in age assortativity (age.mixing.sd.mult)

* Larger values of the multiplier increase the variability in age differences, resulting in less age assortativity (i.e., individuals tend to partner with others from a wider range of ages).
* Smaller values decrease the variability in age differences, leading to greater age assortativity (i.e., individuals tend to partner with others closer to their own age).

This allows us to fine-tune the model to reflect observed patterns in age mixing.

## Sex Mixing Model

We aim to construct a 3x3 matrix representing the proportion of partnerships between females, heterosexual males, and MSM (men who have sex with men). In this model, only female-female partnerships are excluded, while all other pairings can have a positive value.

**Logic**

Consider the case for females: if there is no sex assortativity, the proportion of female partners who are MSM or heterosexual males is proportional to their population distribution in a given location. For example, if 20% of men in Baltimore are MSM, then females would be expected to have 20% MSM and 80% heterosexual male partners. This implies that the **observed-to-expected (OE) ratio** for MSM partnerships would be equal to 1. However, when there is assortativity (i.e., a preference for partnering within specific groups), the OE ratio will deviate from 1—being either greater than or less than one, depending on the degree of assortativity.

### Estimating proportion of females contacts with msm and male hetrosexuals

We estimate the prior value for the OE ratio from a single study:

name=′oe.female.pairings.with.msm′, value=0.0895(Pathela 2006)

Using this value, we can estimate the proportion of female partnerships that are with MSM or heterosexual males in each location as follows:

* Pmsm=0.089×prop.males.msm / (0.089×prop.males.msm+prop.males.not.msm)
* Phet.male=prop.males.not.msm/ (0.089×prop.males.msm+prop.males.not.msm)

Since these are the only two options, the total must satisfy:

* Pmsm+Phet.male=1

This approach allows us to estimate the proportions of MSM and heterosexual male partnerships for females across different locations.

The oe.female.pairings.with.msm is also used for calibration

### Estimating proportion of male contacts with msm and male heterosexuals

What fraction of hetrosexual males have contact with other men: fraction.heterosexual.male.pairings.with.male, value =0.004

What fraction of msm

## Race mixing model

Similar to sex mixing, this relies on observed to expected proportion of contact between difference racial groups

# SYPHILIS TESTING

#probability of having received a HIV test in the last year (used to approximate syphilis screening rate)

#needs a functional form

#assuming X% of people with genital ulcer/rash will seek care, now the quesiton is what proportion of people develop genital ulcer/rash

# 40% have ulcer, 90% seek care = 36% of people with syphilis will seek care

# duration 4 monts -> rate: 36% = 1- e-rt

# look for time to diagnosis for people who are diagnosed, and the look for proportion not diagnosed

Improving efficiency and equity

* Drone operation: medical supply delivery

US department of transportation, NSF

* Deployment telehealth kiosk

Focus on rural

Network design modeling

Identify the objectives and constraints

* Changes in demand distribution in relation with existing facilities

Hospitals and health centers are clustered at more populated areas

Rural areas lack access

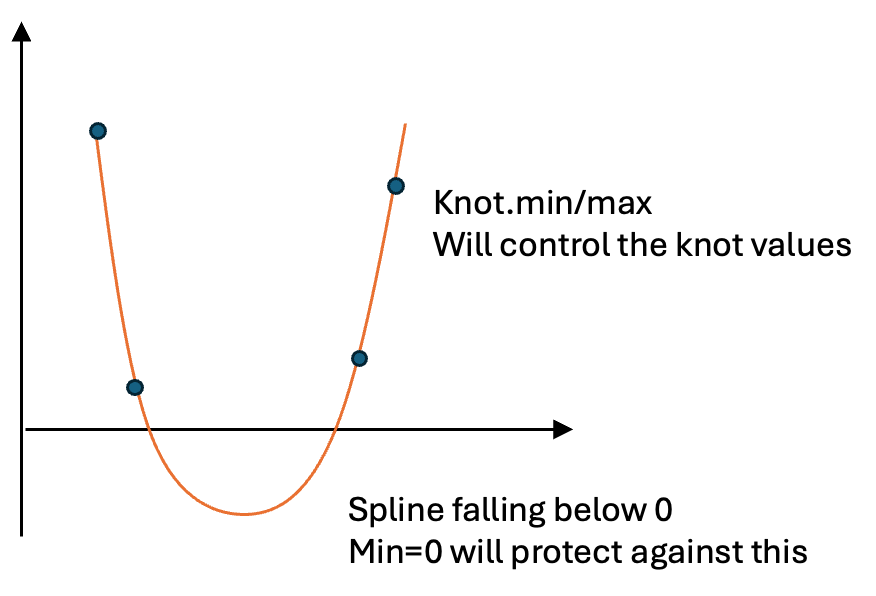
Depending on the current state of the system, we can estimate the demand: any change in system’s condition can change the demand for the future

# SETTING UP PARAMETERS FUNCTIONAL FORMS

## Natural Spline function()

**what are min/max, vs knot.min/knot.max?** Consider an example when we are setting a functional form for a rate parameter. by definition, rate can not fall below 0. There can be a situation where you have 4 knot values greater than zero. If we use an identity link, it will draw a linear projections between the knots and as long as all knots are greater than 0, we don’t have an issue. But if we do use a “log” link, the projected spline between knits could potentially fall below 0 .

* **min/max:** protect against spline values falling outside of range. The minimum and maximum values this functional form can create. The default (NA) sets to the min/max for the link specified (eg, 0-Inf for a log link, 0-1 for a logistic link)
* **knot.min/knot.max:** The min and max values that KNOTS can take before being splined (note, the min and max values the functional form can take, after splining, are determined by parameters min and max)
* **link:** The name of a transformation to the scale at which the SPLINE should apply. The knots are transformed to this scale, the spline is applied, and then the splined values are back-transformed. One of 'identity', 'log', or 'logistic'
* **knot.link:** The name of a transformation to the scale at which the knot values have alphas ADDED. The knots are back-transformed from this scale prior to the spline scale being applied



Example:

get.syphilis.to.hiv.testing.functional.form = function(specification.metadata){

create.natural.spline.functional.form(knot.times = c("1980"=1980, "1990"=1990, "2000"=2000, "2010"=2010,"2020"=2020),

knot.values=list("1980"=0.8, "1990"=0.8, "2000"=0.8, "2010"=0.8,"2020"=0.8),

knot.min = 0, #knot values can not exceed this range

knot.max = 1,#knot values can not exceed this range

min=0, #projected spline values should remain within this range

max=1, #projected spline values should remain within this range

knots.are.on.transformed.scale = F,

knot.link = "log",

link = "identity" #it's safer to use linear to avoid exponential growth

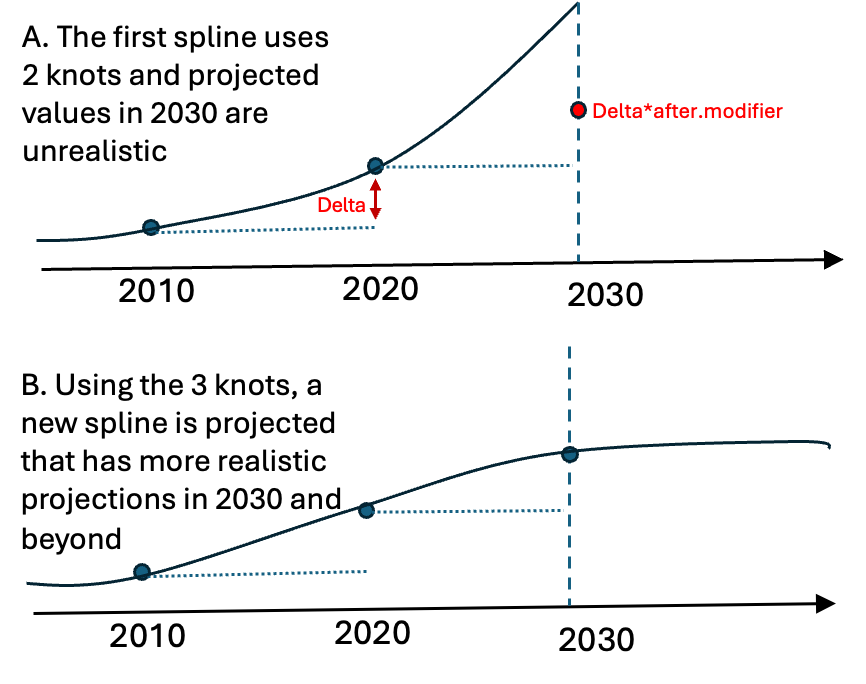
)

**What are Before.time/After.time. and Before.modifier/After.modifier?**

These parameters are used to manage spline projections for periods occurring before the first knot or after the last knot, ensuring that the projections remain realistic.

For example, consider a spline model with two knots at 2010 and 2020. Without adjustments, the projected values for 2030 may be unrealistic and excessively large. To address this, we set the after.time parameter to 2030 and apply an after.modifier of 10% to adjust future projections. Using these parameters, the model creates a new virtual knot in 2030 (red dot) and assigns it a value based on the change observed between 2010 and 2020 (delta), scaled by the modifier value.

The model then fits a new spline using all three knots, resulting in more realistic projections for the period after 2030 (Panel B).



Example:

get.emigration.rates.functional.form <- function(location, specification.metadata, population.years=DEFAULT.MIGRATION.YEAR){

rates = get.emigration.rates(location=location,

specification.metadata = specification.metadata)

create.natural.spline.functional.form(

knot.times = c(time.1 = 2010,time.2 = 2020),

knot.values = list(time.1 = rates, time.2 = rates),

link = "identity",

knot.link = 'log',

min = 0,

after.time = 2030, #this adds a new knot in 2030

after.modifier = 0.1,#sets the value of 2030 knot to 0.1\*changes between 2010-2020

knots.are.on.transformed.scale = F)

}

# Population size

The decennial U.S. Census provides a full count of the population every 10 years (e.g., 2010, 2020). However, population numbers in non-census years are **estimates**, which fall into two main categories:

* **Intercensal Estimates:** These estimates use two census counts (e.g., 2010 and 2020) as anchors and interpolate population changes for the years in between. By incorporating both the start and end points of a decade, intercensal estimates adjust for any errors in the previous census and are generally more accurate.
* **Postcensal Estimates:** These estimates start with the most recent census count (e.g., 2020) and project population changes forward based on demographic data like births, deaths, and migration. Because postcensal estimates lack a second census count for correction, they tend to be less accurate over time as the projection extends farther from the anchor year.

Both types of estimates rely on the **Cohort-Component Method**, which updates the base population by accounting for:

* **Births** and **Deaths**: Data collected from state and local vital statistics offices.
* **Migration**: Estimates of both domestic and international migration derived from administrative records (e.g., IRS, Medicare, and Department of Homeland Security data) and surveys like the American Community Survey (ACS).

In summary, intercensal estimates provide more reliable population figures due to their ability to reconcile errors with two census anchors. Postcensal estimates, while still useful for planning and decision-making, are less precise, especially in years further from the last census.

**CALIBRATION ISSUE:**The challenge we're facing is that the Census Bureau has not yet published the 2010-2020 **intercensal estimates**. The Census Bureau provides the following update on its guidance page:

“*The 2010-2020 Intercensal Estimates of Population by Demographic Characteristics are tentatively scheduled to be released in Fall 2025 (specific timeline for release forthcoming). The Vintage 2020 estimates were modified to account for differences between these estimates and the results of the 2020 Census, resulting in a consistent time series from the 2010 Census to the 2020 Census*.”

This statement explains why there is a noticeable jump in population counts between 2019 and 2020, reflecting adjustments made to align the Vintage 2020 estimates with the results of the 2020 Census. We had hoped the intercensal estimates would be available this year to inform our final calibration. However, with their release now delayed until at least 2025, we’ll need to address the associated uncertainties in a more systematic and principled way.

# Rates and probability of event

1. **Exponential Distribution and Hazard Rate**

The **exponential distribution** describes the time between events in a **Poisson process**, where the event rate is constant over time. In the context of disease progression or failure times, this distribution is commonly used to model the time until an individual develops a disease (the "event").

* **Hazard rate (λ)**: This is the rate at which the event (e.g., disease onset) occurs. It's the inverse of the average time between events. The rate is constant for the exponential distribution.

λ=1 / mean time to event

1. **Cumulative Probability**

The **cumulative probability** of an event occurring by time t is the probability that an individual will develop the disease by time t. In survival analysis, this cumulative probability is often calculated from the **survival function** S(t), which represents the probability of surviving (not developing the disease) up to time t.

* The survival function for an exponential distribution is given by:

S(t)=e−λt

where λ is the hazard rate, and t is time.

* The **cumulative probability** of the event occurring by time t (i.e., the probability of developing the disease) is the complement of the survival function:

P(t)=1−S(t)=1−e−λt

This represents the probability of experiencing the event by time t.

In many epidemiological studies, we encounter the **proportion of patients experiencing an event** by a given time. To use this information in a compartmental model, we often need to **convert this cumulative proportion to an event rate**. Here are two common approaches for doing this:

1. **Approach 1: Assuming Exponential Risk (Constant Hazard Rate)**

We can compute the **hazard rate** using the following formula, which is derived from the **exponential distribution**:

λ=−ln(1−P)​/t

Where:

P is the **cumulative probability** of the event occurring by time t.

λ is the **hazard rate**.

t is the **time** (in years).

This approach assumes that the **rate of event occurrence is constant over time**, and the **probability of the event happening increases non-linearly** as time progresses. This can be approximated using an exponential distribution, where the risk of the event increases as time passes.

1. **Approach 2: Assuming Uniform Risk Over Time (Simplified Model)**

A simpler model assumes that the **event happens uniformly over time**. In this case, we calculate the rate as:

Λ=Pt

Where:

* P is the **cumulative probability** of the event.
* t is the **mean time** to the event.

This approach assumes that the event risk is evenly distributed over time, without accounting for the increasing likelihood of the event over time.

**Comparison:**

These two approaches can often lead to similar results, but they are based on different assumptions about the distribution of risk over time. Here’s an example: 30% of population experienced an event by 30 years of follow up

cumProb=.3

meanTime=50

# **Exponential distribution** considers the increasing likelihood of the event as time progresses and gives a more accurate hazard rate, especially in cases like disease progression or failure rates.

r1= -log(1- cumProb)/meanTime; print(r1)

P1=lapply(c(1:50),function(t) 1-exp(-r1\*t))

## **Simple division** assumes uniform distribution of risk and doesn't account for the compounding risk over time.

r2= cumProb/meanTime; print(r2)

P2=lapply(c(1:50),function(t) 1-exp(-r2\*t))

plot(unlist(P1),type ="l",ylab="Proportion experiencing the event",xlab="Years")

lines(unlist(P2),col="red")

legend("bottomright",legend=c("Assuming Exponential Dist of Risk","Assuming Uniform Dis of Risk"),col=c("black","red"),lty=1)

A graph of a number of years

AI-generated content may be incorrect.

* The exponential distribution model assumes the risk increases over time and gives a more accurate hazard rate in cases like disease progression or failure rates.
* The uniform distribution model is simpler but assumes risk is evenly distributed over time, which doesn't account for the compounding risk as time passes.