

**A Simple Multi-Stage Model of Urinary Tract Infection
Occurrence After Coitarche**

Barbara Rocci Stephen M. Pollock
Department of Industrial & Operations Engineering
University of Michigan Ann Arbor, MI 48109

Betsy Foxman
Department of Epidemiology
University of Michigan Ann Arbor, MI 48109

Technical Report 97-05
May 1997

CUMR 4764

BAP 1630

Barbara Rocci, Stephen M. Pollock
Betsy Foxman

Urinary tract infections - Epidemiology
mathematical models

A SIMPLE MULTI-STAGE MODEL OF URINARY TRACT INFECTION OCCURRENCE AFTER COITARCHE

Barbara Rocci, Stephen M. Pollock, Betsy Foxman

About half of all women will experience a urinary tract infection (UTI) by age 30; approximately 20 - 25% of these will develop a second infection within 6 months of the first one; nearly 5% of women will experience multiple recurring infections. The etiology of UTI's is not well-understood, and there is evidence that at least some may be sexually transmitted. Mathematical models can provide insight into the transmission dynamics of these infections, as well as help to evaluate the possible effects of prevention strategies. One such model is presented here.

MODEL DESCRIPTION

In the basic model, a woman is tracked at the start of sexual activity (coitarche) and can progress through a number of possible stages, as shown in Figure 1. Thus the model assumes a woman can be in any one of the following states at any given time:

- a) State S_t , $t = 1, 2, \dots, 14$, where $t =$ years after coitarche with no UTI
- b) State M_j , $j = 1, 2, \dots, 6$, where $j =$ months after first UTI with no second UTI
- c) State "UTI1" = occurrence of first UTI
- d) State "UTI2" = occurrence of second UTI
- e) State "No UTI 1 in 14 yrs" = no first UTI within 14 years of coitarche
- f) State "No UTI 2 in 6 mos" = no second UTI within 6 months of first UTI

From state S_t , there is a probability $h(t)$ per month of moving into state "UTI1". From state M_j there is a probability $m(j)$ per month of moving into state "UTI2". Let $T =$ year (after coitarche) of first UTI, $J =$ month after first UTI of second UTI. Then these probabilities are, by definition, the "hazards":

$$\begin{aligned} h(t) &= P(T = t | T > t - 1) & t = 1, 2, \dots, 14 \\ m(j) &= P(J = j | J > j - 1) & j = 1, 2, \dots, 6 \end{aligned}$$

The model also includes an aging process. For each month i in year t after coitarche,

$$p_i(t) = P(S_t \rightarrow S_{(t+1)} \text{ in month } i) = \begin{cases} 0 & i = 1, 2, \dots, 11 \\ 1 & i = 12 \end{cases}$$

The transition to year $t+1$ occurs at the end of month 12 in year t . This is illustrated in the detailed view of state S_2 in Figure 1.

Parameter Estimation - First UTI

We can estimate $h(t)$ using data from Foxman, *et al.* (unpublished), in which the time from coitarche to first UTI was recorded for 210 women. Since all the women in this study developed a UTI, $h(t)$ cannot be estimated directly. The longest time from coitarche to first UTI in the study was 14 years. Thus we first estimate the probability $g(t) = P(T = t | t - 1 < T \leq 14)$, then remove the condition that $T \leq 14$ to obtain an estimate of $h(t)$. Kaplan-Meier estimates of $g(t)$ were first calculated. This non-parametric technique involves recording the number of women at risk of UTI and

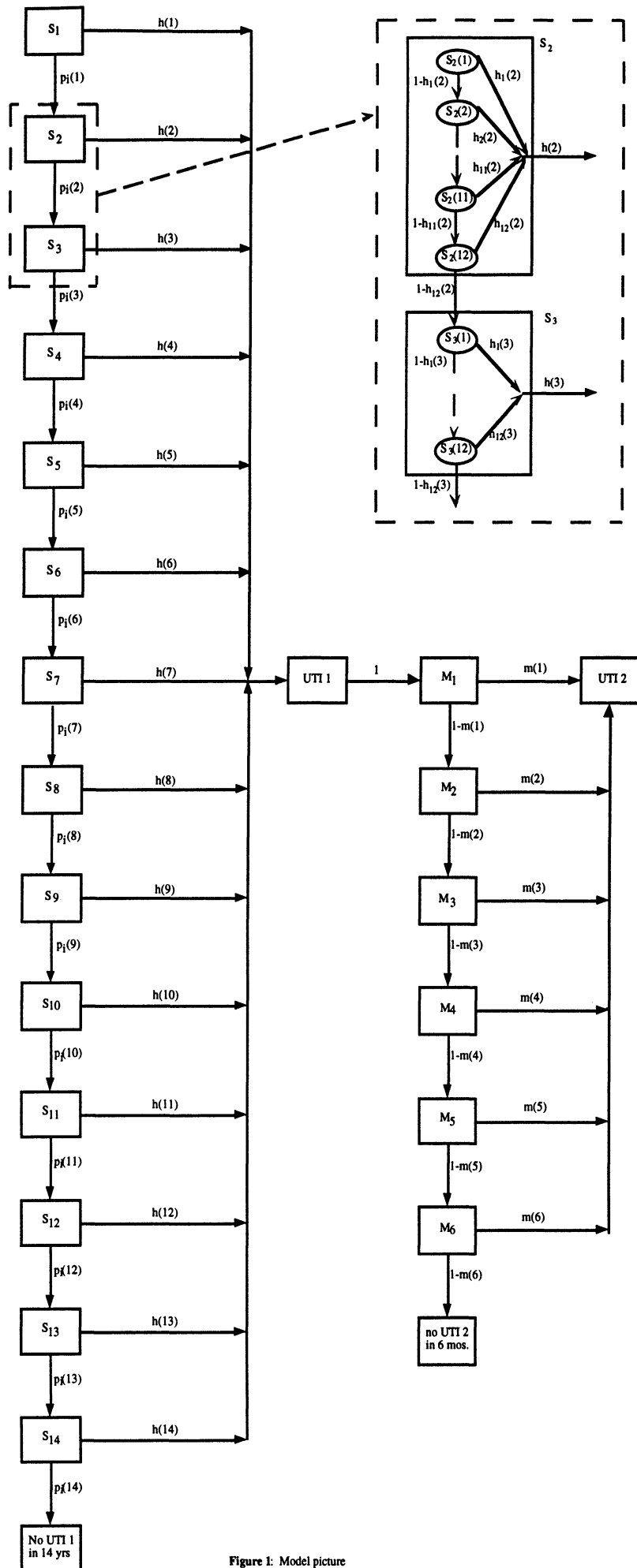


Figure 1: Model picture

the number developing a UTI during each year, then estimating the overall probability of surviving up to t . From the data, we have, for $t = 1, 2, \dots, 14$:

$$\begin{aligned} n_t &= \text{number of women for whom } T > t - 1 \\ d_t &= \text{number of women for whom } T = t \end{aligned}$$

Detailed calculations are given in Appendix A. The resulting estimates are shown by the squares in Figure 2. These results suggest that $g(t)$ might reasonably take the form

$$g(t) = \begin{cases} at & t \leq b \\ ab & t \geq b \end{cases} \quad (1)$$

Using this representation, maximum likelihood parameters \hat{a} and \hat{b} were calculated (see Appendix B), yielding $\hat{a} = 0.18$ and $\hat{b} = 4$ years. The solid line in Figure 2 shows the resulting $\hat{g}(t)$.

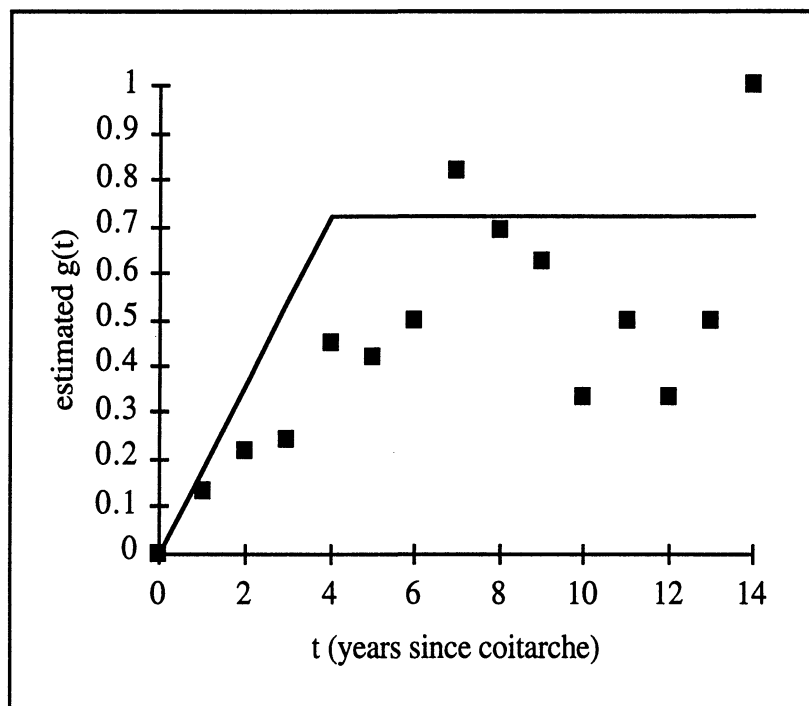


Figure 2: Conditional probability per year after coitarche of developing first UTI (• = K-M estimate, — = parametric estimate using equation (1))

The parametric estimate

$$\hat{g}(t) = \begin{cases} \hat{a}t & t \leq \hat{b} \\ \hat{a}\hat{b} & t > \hat{b} \end{cases} \quad (2)$$

is used in the model. As defined above, $\hat{g}(t)$ is an estimate of the conditional probability of developing the first UTI in year t given that a UTI will occur by year 14 after coitarche. Since a woman will not necessarily experience a UTI within 14 years after initiating sexual activity, it is necessary to calculate the unconditional estimate for use as $\hat{h}(t)$ in the model. Again, letting T be the year of first UTI, we have:

$$\begin{aligned}
\hat{g}(t) &= P(T = t \mid t-1 < T \leq 14) \\
&= \frac{P(T \leq 14, T = t \mid T > t-1)}{P(T \leq 14)} \\
&= \frac{P(T \leq 14 \mid T = t)P(T = t \mid T > t-1)}{P(T \leq 14)}
\end{aligned} \tag{3}$$

Clearly, $P(T \leq 14 \mid T = t) = 1$ for $t < 14$. Thus equation (3) becomes:

$$\hat{g}(t) = P(T = t \mid t-1 < T \leq 14) = \frac{P(T = t \mid T > t-1)}{P(T \leq 14)}.$$

Thus an estimate of $h(t)$ is given by:

$$\hat{h}(t) = P(T = t \mid T > t-1) = P(T = t \mid t-1 < T \leq 14)P(T \leq 14) = \hat{g}(t)P(T \leq 14) \tag{4}$$

Studies suggest that about half of all women experience a UTI by age 30. Assuming that the average age of coitarche is 16, $P(T \leq 14) \cong 0.5$. Thus equation (4) becomes:

$$\hat{h}(t) = (0.5)\hat{g}(t) \tag{5}$$

Equation (5) is used to calculate $\hat{h}(t)$, using the parametric estimate $\hat{g}(t)$ of equation (2).

Parameter Estimation - Second UTI

The model assumes that an infection is treated and cured within one month, and then allows the development of a second UTI within six months after the first one. In particular, for month j , $j=1, \dots, 6$ after the first UTI, there is a conditional probability $m(j)$ of developing a second UTI, given no second UTI through month $j-1$. Kaplan-Meier estimates of $m(j)$, $j=1, \dots, 6$ were calculated using data from Foxman, *et al* (unpublished). A total of 263 women were followed from their first UTI for six months or until they developed a second UTI, whichever came first. Figure 3 shows the resulting Kaplan-Meier estimates. These non-parametric estimates are used in the model.

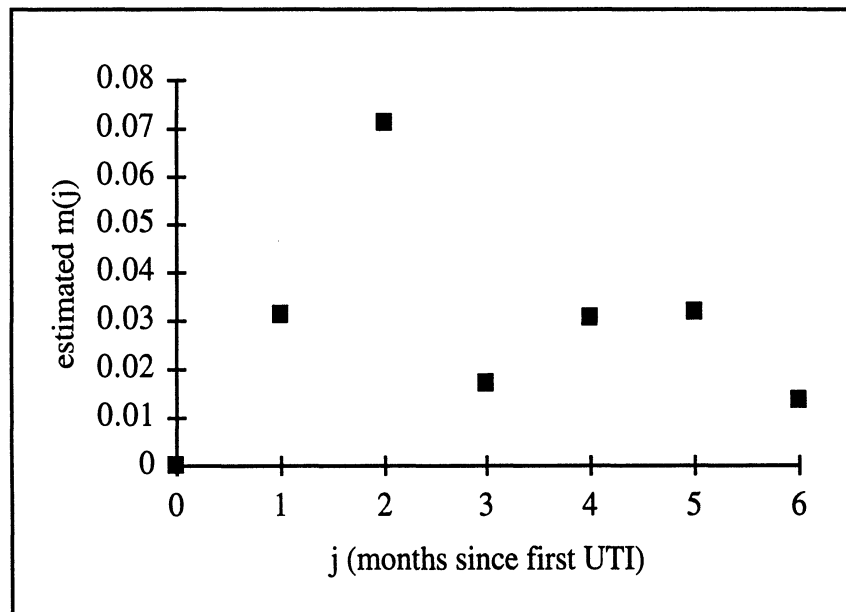


Figure 3: Conditional probability per month after first UTI of developing a second UTI

RESULTS

The model, using $\hat{g}(t)$ of equation (2) and $\hat{m}(j)$, was solved using STELLA II, a differential equation solver. The resulting cumulative probability of experiencing a first UTI as a function of time since coitarche is shown in Figure 4. The cumulative probability of having a second UTI was also found using the model, and is shown in Figure 5.

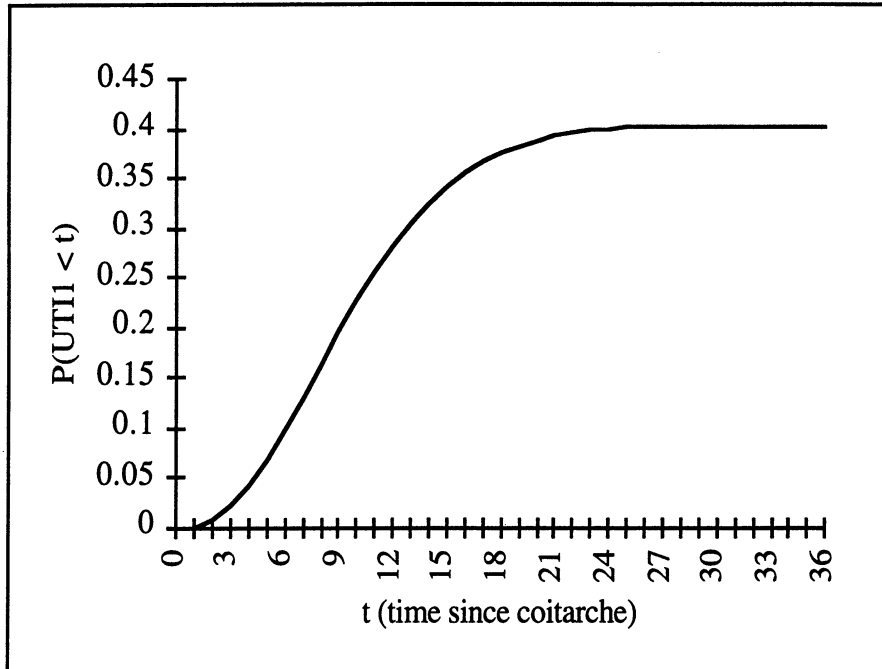


Figure 4: Cumulative probability of experiencing a first UTI

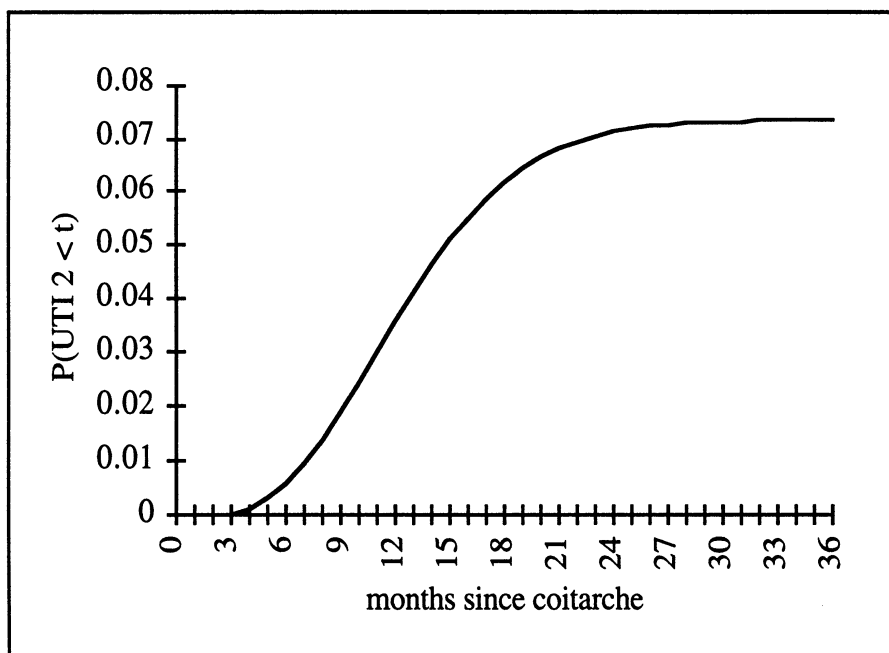


Figure 5: Cumulative probability of experiencing a second UTI

In Figure 4, we see that by 2 years after coitarche, approximately 40% of women will have experienced a first UTI. Figure 5 shows that approximately 7% of women will have had at least two UTI's within 2 years of coitarche. Clearly, women will experience a second UTI only after having had a first one, thus the curve in Figure 5 lags behind that in Figure 4 by a few months. Also, the two figures are displayed on very different scales, indicating the different magnitude of probabilities of first and second UTI's.

Potential Use of the Model

The model can readily provide information on how intervention strategies might change the probability of developing a first or second UTI. A simple example serves to illustrate how we can evaluate possible public health strategies. Suppose an intervention strategy were available that would lower all $h(t)$ by 20%; the results are shown in Figures 6 and 7. Lowering $\hat{h}(t)$ by 20% does not appear to have a significant effect on the asymptotic values of $P(\text{UTI } 1 \leq t)$ or $P(\text{UTI } 2 \leq t)$, but shifts both curves to the right somewhat, indicating that for time t , the probability of having developed a UTI before t is lower after the intervention. The shift is greater for second UTI than first UTI. Information such as this could help public health officials determine whether intervention strategies are justified. Other examples of the model's usefulness in evaluating alternatives include:

a) *Different Strains of Bacteria*

Approximately 50% of second UTI's are caused by the same strain of bacteria that caused the first UTI. The model can be modified to allow for the possibility of second infection by the same or a different strain, by adding more states. Thus instead of a single "UTI 2" state, there would be two: one for re-infection with the same strain, and one for infection with a different strain. Intervention strategies might affect the probabilities per month of developing a second infection with the same or a different strain as the first. Similarly, the model can be used to study the difference between being infected with *E. Coli* versus non-*E. Coli* bacteria for the first UTI. There may be different recurrence rates for these two types; thus the model could be used to study possible interventions based on behavior profiles or symptom profiles.

b) *Condom Use*

There is evidence that condom use is protective against a second infection but may be associated with higher probabilities of first infection. These elements could be accounted for in the model by raising or lowering the appropriate probabilities of infection to see how the final probabilities of having an infection. To simulate the protective effect of condoms, $\hat{h}(t)$ or $\hat{m}(j)$ could be raised and lowered, respectively, and the results examined to determine the effect on the probability of developing an infection by a given time.

In all of these examples, the model can also provide insight on how sensitive the final distributions are to changes in parameters. For example, figures 6 and 7 show that $\hat{h}(t)$ would have to be lowered by more than 20% to have an impact on the asymptotic cumulative probabilities.

The responses of 2000 women to an extensive questionnaire are currently being analyzed. The questions cover various areas that could be related to the etiology of UTI, including number of sexual partners, type of contraceptive used, and estrogen therapy. We anticipate that these data will provide many avenues of research for which the model described here might be useful, such as an examination of the effects of estrogen therapy or sexual behaviors on the probability of developing UTI.

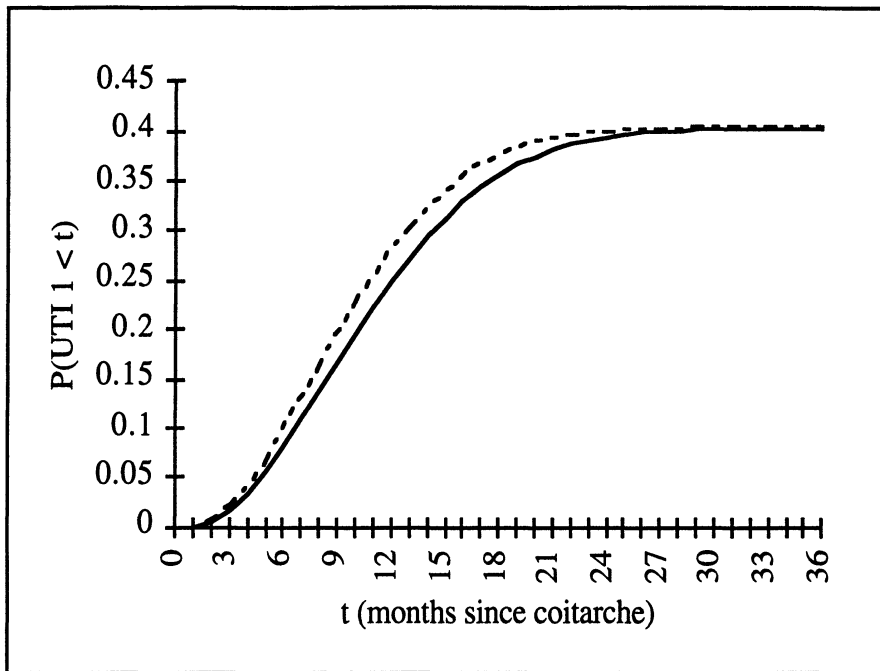


Figure 6: Cumulative probability of experiencing a first UTI
 (--- original $\hat{h}(t)$, — $\hat{h}(t)$ decreased by 20%)

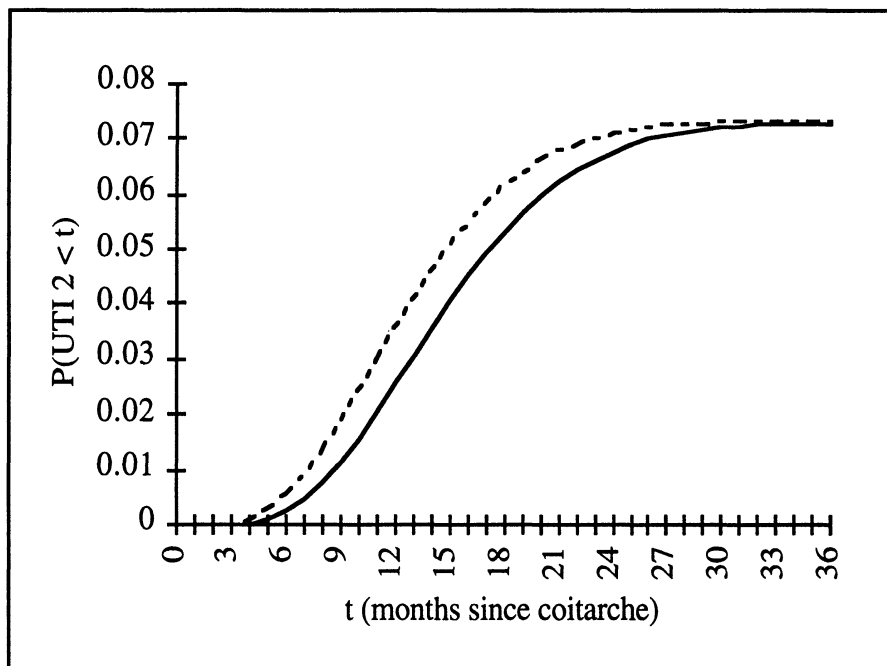


Figure 7: Cumulative probability of experiencing a second UTI
 (--- original $\hat{h}(t)$, — $\hat{h}(t)$ decreased by 20%)

APPENDIX A Kaplan-Meier Estimation

In order to estimate $h(t) = P(T = t | T > t - 1)$, we first need to estimate the conditional hazard, $g(t) = P(T = t | t - 1 < T \leq 14)$, from the available data. The condition $T > t - 1$ is implicitly assumed to be present in the work that follows.

Let n_t be the number of women who have not yet had a UTI by the start of year t , and d_t be the number of women who develop a UTI in year t . The hazard is defined as:

$$g(t) = \frac{f(t)}{S(t)} \quad (1)$$

where $f(t)$ is the probability mass function of T (the time of first UTI), and $S(t)$, the “survival” function, is $P(T \geq t)$. Both $f(t)$ and $S(t)$ can be estimated from the data. The number of women who do not develop a UTI during year t is given by $(n_t - d_t)$. An estimate of $f(t)$ is thus obtained by dividing the number of women developing a UTI during year t by the total number of women at risk. Thus for each year t , we have:

$$\hat{f}(t) = \frac{n_t}{210} \quad (2)$$

where 210 is the total number of women in the data set.

The estimation of $S(t)$ is slightly more involved. Let $p_t = P(\text{no UTI in year } t | \text{no UTI up to } t - 1)$. Then $S(t) = p_1 p_2 \dots p_t$. An estimate of p_t is:

$$\hat{p}_t = \frac{(n_t - d_t)}{n_t} = 1 - \frac{d_t}{n_t} \quad (3)$$

We can now use (3) to provide the Kaplan-Meier estimates $S(t)$:

$$\hat{S}(t) = \prod_i \left(1 - \frac{d_i}{n_i} \right) = \hat{S}(t-1) \left(1 - \frac{d_t}{n_t} \right) \quad (4)$$

Finally, using (2) and (3) in equation (1) produces an estimate of the conditional hazard, $\hat{g}(t)$. The results are shown in Table A.1 and Figure 2.

The same procedure was used to estimate $m(j) = P(J = j | J > j - 1)$. The results are shown in Table A.2 and Figure 3.

APPENDIX A, cont'd

Year	d_t	n_t	$\hat{S}(t)$	$\hat{f}(t)$	$\hat{h}(t)$
0	0	210	1	0	0
1	24	210	0.886	0.114	0.129
2	33	186	0.729	0.157	0.216
3	30	153	0.586	0.143	0.244
4	38	123	0.405	0.181	0.447
5	25	85	0.286	0.119	0.417
6	20	60	0.190	0.095	0.500
7	18	40	0.105	0.086	0.818
8	9	22	0.062	0.043	0.692
9	5	13	0.038	0.024	0.625
10	2	8	0.029	0.010	0.333
11	2	6	0.019	0.010	0.500
12	1	4	0.014	0.005	0.333
13	1	3	0.010	0.005	0.500
14 *	1	2	0.005	0.005	1.000

* Note that only one person developed a UTI during year 14, while two were at risk. The estimation method gives an estimate of zero for the hazard function for all subsequent years until the last person developed a UTI, at which point the estimate of the hazard is undefined. Thus estimates were carried out only until year 14.

Table A.1: Kaplan-Meier Estimation of Yearly Hazard to First UTI

Month	d_t	n_t	$\hat{S}(j)$	$\hat{f}(j)$	$\hat{m}(j)$
0	0	263	1	0	0
1	8	263	0.970	0.030	0.031
2	17	255	0.905	0.065	0.071
3	4	238	0.890	0.015	0.017
4	7	234	0.863	0.027	0.031
5	7	227	0.837	0.027	0.032
6	3	220	0.825	0.011	0.014

Table A.2: Kaplan-Meier Estimation of Monthly Hazard to Second UTI

APPENDIX B Maximum Likelihood Estimation

As an alternative to using the Kaplan-Meier (non-parametric) estimates of $g(t)$, a parametric model can be used. The following functional form is assumed for the hazard:

$$g(t) = \begin{cases} at & t \leq b \\ ab & t \geq b \end{cases} \quad (1)$$

For convenience, this formulation assumes that t is continuous. From equation (1), the probability density function for T (the time of first UTI) is:

$$f(t) = g(t) e^{-\int_0^t g(\xi) d\xi} = \begin{cases} ate^{-at^2/2} & t \leq b \\ abe^{-ab(t-b)} & t \geq b \end{cases} \quad (2)$$

Given the data t_i = time of the i^{th} UTI developed and letting $m(b)$ indicate the last UTI developed before time b , the likelihood function $L(a, b)$ is:

$$\begin{aligned} L(a, b) &= \prod_i^n f(t_i) \\ &= \prod_{\{i:t_i \leq b\}} at_i e^{-at_i^2/2} \prod_{\{i:t_i > b\}} abe^{-ab(t_i-b)} \\ &= a^{m(b)} \left(\prod_{\{i:t_i \leq b\}} t_i \right) e^{-\frac{a}{2} \sum_{\{i:t_i \leq b\}} t_i^2} a^{n-m(b)} b^{n-m(b)} e^{-ab \sum_{\{i:t_i > b\}} (t_i-b)} \end{aligned} \quad (3)$$

The maximum likelihood estimates \hat{a} and \hat{b} are the values that maximize $L(a, b)$. To maximize $L(a, b)$, we take the logarithm of each side to get the log-likelihood function, $l(a, b)$.

$$l(a, b) = m(b) \ln a + \ln \left(\prod_{\{i:t_i \leq b\}} t_i \right) - \frac{a}{2} \sum_{\{i:t_i \leq b\}} t_i^2 + [n - m(b)] \ln a + [n - m(b)] \ln b - ab \sum_{\{i:t_i > b\}} (t_i - b) \quad (4)$$

Values of a and b that maximize $L(a, b)$ will also maximize $l(a, b)$. Each side of equation (4) can be differentiated with respect to a , and set equal to zero (a necessary condition for maximization, since $l(a, b)$ is a concave function of a):

$$\frac{\partial l}{\partial a} = \frac{m(b)}{a} - \frac{\sum_{\{i:t_i \leq b\}} t_i^2}{2} + \frac{n - m(b)}{a} - b \sum_{\{i:t_i > b\}} (t_i - b) = 0 \quad (5)$$

Equation (5) can be solved to get \hat{a} in terms of \hat{b} :

$$\hat{a} = \frac{n}{\left(\sum_{\{i:t_i \leq b\}} t_i^2 / 2 \right) + \hat{b} \sum_{\{i:t_i > b\}} (t_i - b)} \quad (6)$$

A crude search technique was used, in which candidate integer values of \hat{b} were selected, corresponding \hat{a} values found from equation (6), then $l(\hat{a}, \hat{b})$ evaluated to find the maximum. Final estimates were $\hat{a} = 0.18$ and $\hat{b} = 4$.