Paper 1624-2014

Using Arrays for Epidemic Modeling in SAS®

Carl Grafe, University of Utah, Salt Lake City, UT; Aaron Wendelboe, University of Oklahoma Health Sciences Center, Oklahoma City, OK

ABSTRACT

Epidemic modeling is an increasingly important tool in the study of infectious diseases. As technology advances and more and more parameters and data are incorporated into models, it is easy for programs to get bogged down and become unacceptably slow. The use of arrays for importing real data and collecting generated model results in SAS® can help to streamline the process so results can be obtained and analyzed more efficiently. This paper describes a stochastic mathematical model for transmission of influenza among residents and healthcare workers in long-term care facilities (LTCFs) in New Mexico. The purpose of the model was to determine to what extent herd immunity among LTCF residents could be induced by varying the vaccine coverage among LTCF healthcare workers. Using arrays in SAS made it possible to efficiently incorporate real surveillance data into the model while also simplifying analyses of the results, which ultimately held important implications for LTCF policy and practice.

INTRODUCTION

Mathematical models for the spread of infectious diseases are increasingly important in the study of infectious disease policy and control (Vynnycky and White 2010). These models come in a wide variety of types and can be run using a wide variety of software programs. This paper describes the development of a stochastic mathematical model for transmission of influenza among residents and healthcare workers in long-term care facilities in New Mexico using SAS. The use of arrays in SAS made it possible to efficiently develop and run the model and analyze the model results.

EPIDEMIC MODELING

Infectious diseases spread through populations according to certain traits inherent in the diseases themselves and in traits of the individuals being infected. These traits, called parameters, include a pathogen's transmissibility, its latent and incubation periods, recovery rates for infected persons, rates of contact between persons, and so forth. These parameters can be connected together mathematically to predict how a pathogen will behave in a given population. The representation of these mathematical connections is called a model.

One way of organizing a mathematical model is to separate the different groups involved into different categories. These are called compartmental models. For example, an SEIR model (see Figure 1) is one that has been divided between those in the population who are susceptible to becoming infected (S), those who have been exposed to an infected person but have not yet become infectious (E), those who are currently infectious (I), and those who have recovered and are no longer infectious (R). The model parameters and structure dictate how long individuals remain in each category and how they move between categories.

Mathematical models used for epidemic modeling may be classified as either deterministic or stochastic. A deterministic model is one where the outcome given a specific set of parameters will always be identical for those parameters. For example, for a deterministic model, the number of infected persons after a fixed period of time will always be the same no matter how many times the model is run. For some models, especially those involving smaller populations, it may be desirable to incorporate random variation into the model. These models are called stochastic models (Vynnycky and White 2010).

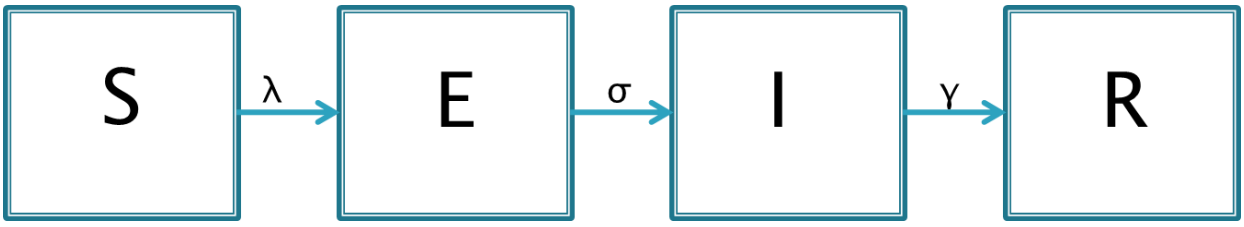

Figure 1. A simple Susceptible, Exposed, Infectious, Recovered (SEIR) compartmental model, where λ is the infection rate, σ is the incubation rate, and γ is the recovery rate.

INFLUENZA IN LONG-TERM CARE FACILITIES

Influenza is a leading cause of death in general in the United States (Centers for Disease Control and Prevention 2013), and is even more prevalent among residents of long-term care facilities (LTCFs), where case-fatality rates can range from 10%-20% (Pearson, Bridges, and Harper 2006). In connection with this elevated risk, the Centers for Disease Control and Prevention (CDC) recommends that all healthcare workers receive the influenza vaccine annually (Centers for Disease Control and Prevention 2011).

The primary motivation for vaccinating HCWs in LTCFs is to induce herd immunity (Centers for Disease Control and Prevention 2013), the indirect protection that unvaccinated or otherwise vulnerable persons receive by being surrounded by vaccinated persons. For some diseases, there is also a herd immunity threshold, a level of vaccination coverage in the general population such that the number of cases of that disease drops near to zero (Vynnecky and White 2010). The question investigated in this study is whether there exists a herd immunity threshold among the HCW population in a LTCF such that the attack rate in the resident population of the LTCF goes to zero. Previous research using a mathematical model has found that no such threshold exists for influenza in LTCFs (van den Dool, Bonten, Hak, Heijne, and Wallinga 2008). However, these findings have not previously been validated using real data.

During the 2006-2007 influenza season (November-April), the New Mexico State Department of Health (NMDOH) collected data from all 75 LTCFs in the state as part of an active surveillance system for influenza. On a monthly basis, from each facility, the average numbers of residents and health care workers (HCWs), the average numbers of residents and HCWs vaccinated against influenza, and the numbers of cases of influenza among residents and HCWs were sought by NMDOH. More detailed information on the study population and on the data collected are included elsewhere (Wendelboe, Avery, Andrade, Baumbach, and Landen 2011).

ARRAYS IN SAS®

A SAS ARRAY is a tool that can be used to simplify the DATA step in a SAS program. Its usefulness comes from its provision of "an alternative method for referring to a variable rather than using the name of the variable" (SAS Institute Inc. 2014a, p. 1). It is simply a set of variables of the same type that can be referenced using the same array name (Waller 2010). So if you have a large number of variables of the same type that you want to perform the same operation on, an array can keep you from having to write out excessive lines of code that are essentially doing the same thing.

SAS ARRAYS have an important feature increasing their flexibility; specifically, the variables do not have to be "already existing" (SAS Instute Inc. 2014a, p. 2). This means that the variables can be created when the array is populated, which can be convenient when using a model to generate data. On the other hand, SAS ARRAYS also have an important limiting property in that the variables of which they are composed must all be of the same type (character or numeric).

The basic syntax of the ARRAY statement is straightforward (SAS Institute Inc. 2014b):

```
ARRAY array-name { subscript } <$><length>
<array-elements><(initial-value-list)>;
```

Detailed descriptions of how to properly use the ARRAY statement in different circumstances can be found elsewhere (for example, SAS Institute Inc. 2014b), but it is useful to point out how they can be helpful in making SAS programming more efficient. The primary benefit of using arrays is that they substantially reduce the amount of coding required. This makes a program easier to debug and maintain (SAS Institute Inc. 2014a). While arrays have been demonstrated to run slightly slower than direct coding in some circumstances (Langston 2005), it has also been shown that in some circumstances this can be avoided with thoughtful programming (Keelan 2002). In any case, depending on complexity and the number of variables involved, the reduction in time for programming and debugging attributable to using arrays should be considered against any corresponding potential loss in runtime efficiency (Langston 2005).

METHODS

A detailed description of the model design is beyond the scope of this paper, but a brief overview of the general structure of the model and the statistical analyses that were performed is included below. The use of arrays in the model is covered in greater detail.

MATHEMATICAL MODEL

In order for our results to be comparable to previous findings, we developed our model based closely on the model described by van den Dool et al. (2008), with deviations limited to parameters particular to the real data collected, such as changing the size of the community population to that of the state of New Mexico, allowing numbers and vaccination rates of HCWs and residents to vary according to the collected surveillance data by facility, and so forth. Only data from 63 of the 75 LTCFs were used in the model due to missing data in 12 facilities. We used a stochastic individual-based SEIR model for transmission of influenza in the LTCFs combined with a deterministic SIR model for transmission of influenza in the community. We used discrete time intervals of eight hours to correspond with what might be expected for a typical HCW shift. The model simulated the six month "influenza season," covering the time period from November to April. Most parameters were identical to those used in the previous model (van den Dool et al. 2008).

Transmission

The simulated influenza season begins with a single infected individual in the community. Vaccination of residents and HCWs, contact between individuals, and movement from one compartment to another were determined by sampling from Bernoulli distributions with means set to the appropriate parameters. Thus during each shift the model determined for each individual in each facility whether they (if initially unvaccinated) had been vaccinated since the last shift, if residents had contact with visitors from the community, if HCWs made "close" contact with a resident during their shift, and so forth. Similarly, if contact between a susceptible individual and an infected individual occurred, a Bernoulli distribution was sampled from to determine whether the individual in the susceptible compartment (S) became infected and moved into the exposed compartment (E), and then for each successive shift whether they progressed to the infectious compartment (I) and then to the recovered compartment (R), depending on the corresponding parameters.

Outcome

Based on the analyses in the previous model (van den Dool et al. 2008), the outcome of interest was determined to be the seasonal attack rate (the total number of infected residents in a facility divided by the total number of residents who occupied a bed in the facility at any point during the season) in each LTCF at each level of HCW vaccination coverage from 0% to 100% (in 10% intervals). In order to generate these outcomes, the model was run 1,000 times at each level of HCW vaccination coverage, and then the resulting attack rates for the 1,000 runs were averaged for each LTCF.

Herd Immunity

The determination of whether herd immunity was attained was made by evaluating whether a curvilinear trend existed between the seasonal attack rates and the levels of HCW coverage in the 63 included facilities. If herd immunity against influenza in residents could be induced by vaccinating HCWs, the relationship between resident attack rate and HCW vaccination coverage would be expected to be concave, with greater percentages of HCWs vaccinated corresponding to lower attack rates among residents, and with greater declines in attack rates as HCW vaccination coverage approached 100%. If a herd immunity threshold existed, the trend would be expected to change to convex at some point as the attack rate approached zero. If no herd immunity existed, the trend would be expected to be linear.

Statistical Analyses

Because of the nature of the estimates generated by the epidemic model, a mixed model was used to determine whether a curvilinear trend existed between the seasonal attack rates and the levels of HCW vaccination coverage in the 63 included facilities:

```
proc mixed data=grafe.mean_ar covtest cl;
   class ut facilityID;
   model ARpred = uptake uptake*uptake / solution cl;
   repeated ut / subject=facilityid type=un r;
run;
```

In the mixed model, uptake is the HCW vaccination coverage, facilityID is the unique identifier for each LTCF, and ARpred is the mean attack rate generated by the model for each LTCF. The explanatory quadratic term uptake*uptake is included in the model statement in order to determine whether the relationship between the variables is curvilinear. The variable ut is a dummy variable for uptake used for the analysis.

Other analyses were also conducted, including a goodness-of-fit test, which compared the results of an additional 1,000 runs of the model using the actual observed HCW vaccination coverage rates for the LTCFs against the actual

observed seasonal attack rates for the LTCFs, and sensitivity analyses for the contact parameter and the community visitor parameter.

USE OF ARRAYS IN THE MODEL

Arrays were used extensively in the epidemic model, both for reading existing data into the model and for storing the data generated by the model. The arrays containing the newly generated data were incorporated into the variables used to generate the summary counts and rates used for the analyses.

Reading Existing Data into Arrays

Since the surveillance data for the 63 LTCFs were collected on a monthly basis, the simulation held parameters based on surveillance data relatively fixed during each 30 day "month." Allowed variation included such things as the number of residents, which was allowed to vary slightly as residents died or were discharged from LTCFs (the beds were assumed to be refilled at the same rate as they were emptied, keeping the mean resident count approximately constant throughout each month), and the vaccination coverage for residents and HCWs, which increased gradually as the model allowed additional individuals to become vaccinated. Four arrays were used for the imported monthly data from the 63 LTCFs (NUMRES, the average number of residents during each month; NUMEMP, the average number of HCWs during each month; COVRES, the vaccination coverage among residents each month; and COVEMP, the vaccination coverage among HCWs each month):

```
array NumRes{*} NumResNov NumResDec NumResJan NumResFeb NumResMar NumResApr;
array NumEmp{*} NumEmpNov NumEmpDec NumEmpJan NumEmpFeb NumEmpMar NumEmpApr;
array CovRes{*} CovResNov CovResDec CovResJan CovResFeb CovResMar CovResApr;
array CovEmp{*} CovEmpNov CovEmpDec CovEmpJan CovEmpFeb CovEmpMar CovEmpApr;
```

The model rotated through the data in these arrays over the six month simulation as follows (where t is an index variable that counts from -60 to 120 in increments of 1/3):

```
if -60 le t lt -30 then month = 1;
else if -30 le t lt 0 then month = 2;
else if 0 le t lt 30 then month = 3;
else if 30 le t lt 60 then month = 4;
else if 60 le t lt 90 then month = 5;
else month = 6;
```

Thus the numbers of residents and HCWs and the vaccination coverage for the residents were reset at the beginning of each simulated month. The simulated HCW vaccination coverage was allowed to vary from 0% to 100% during the run of the model, so the array for the actual HCW coverage at each LTCF obtained from the surveillance data was included only for model validation. This array was used to generate the seasonal attack rates predicted by the model at the actual HCW vaccination rates so that these predictions could be compared with the actual observed attack rates at each facility to determine how well the model matched reality.

Populating Arrays with Newly Generated Data

Additional arrays were created for each resident's SEIR status (0 = "Vacant," 1 = "Susceptible," 2 = "Exposed," 3 = "Infectious," 4 = "Recovered"), each HCW's SEIR status (1 = "Susceptible," 2 = "Exposed," 3 = "Infectious," 4 = "Recovered"), and each HCW's work status (0 = "At work, 1 = "Not at work"). The HCW work schedule was designed to fairly distribute shifts among the HCWs in a facility while maintaining adequate staffing levels. The work schedule also incorporated the expected variation in staffing levels during different times of the day (day shift, evening shift, and night shift). If a HCW was "Not at work" during a given iteration, the probabilities for contact and exposure for the HCW were based on the community population part of the model (as were the probabilities for LTCF visitors), allowing the amount of infectiousness in the community to directly influence the likelihood of infection in the LTCF. As the HCW work status array played an important role in what parameters applied to HCWs during a given shift, the array was incorporated into calculations pertaining to the HCW SEIR status array, which likewise incorporated the resident's SEIR status array (which was also likewise incorporated into the HCW SEIR status array), enabling the various interactions between residents and HCWs and visitors and other community members to be predicted and recorded.

Preparation of Array Data for Analysis

As the model iterated through the influenza season and through each simulated individual in the various arrays, other variables were used to capture the data needed in order to run the statistical analyses. Counter variables for the total number of infected residents and for the total number of residents who occupied a bed at any time during the season in a LTCF were incorporated into the model. During each iteration of the model, if a bed in the resident SEIR array

was found to contain a newly infected resident (SEIR status = 3 where it had been 2 or 0 during the previous iteration), the counter variable for the total number of infections during the season was incremented by one. Similarly, if a bed in the resident SEIR array was found to contain a new resident (SEIR status = 1, 2, 3, or 4 where it had been 0 during the previous iteration), the counter variable for the total number of residents in the facility during the season was incremented by one. As the discharge/mortality rate was relatively low at 1/425 per day (from van den Dool et al. 2008, citing [Anonymous] 2006), the possibility of double counting readmitted infected persons was assumed to be negligible.

After using the arrays to increment the counter variables throughout the season, attack rates for each LTCF were calculated by dividing the counter variable for the total number of infected residents during the season by the counter variable for the total number of residents who occupied a bed at any time during the season. These seasonal attack rates were used in the statistical analyses.

RESULTS

The model predicted a slight concave curvilinear trend between the seasonal resident attack rates and the HCW vaccine coverage rates in the LTCFs (unpublished observations). However, the mixed model for the predicted mean attack rates never came close to zero, indicating that while the model predicts a small amount of herd immunity may be induced by vaccinating HCWs, there is no apparent threshold at which further increases in HCW vaccine uptake would cease to significantly affect the resident attack rate. While this finding of induced herd immunity differs from the previous model (van den Dool et al. 2008), it corroborates their finding that no herd immunity threshold can be induced strictly by vaccinating HCWs. Results from other statistical analyses were also obtained (unpublished observations).

The use of SAS ARRAYS in the model was determined to be effective based on the additional coding that would have been required to run the model without them. Considering, for example, the arrays used in reading in the surveillance data, a reference to the array for the number of residents in a facility was made by using a DO loop and a simple call to whichever variable corresponded to the month the model happened to be on:

```
bed_ct = NumRes(month);
```

In this example, bed_ct is the number of occupied beds in an LTCF, and it is being initialized using the NUMRES array, which gives the number of residents in the LTCF for the current month from the surveillance data (where month is based on t from -60 to 120 as described in the Methods section above). Direct coding, on the other hand, would have required the number of residents for each month to be called separately, perhaps using IF-THEN statements:

```
if Month = 1 then bed_ct = NumResNov;
if Month = 2 then bed_ct = NumResDec;
if Month = 3 then bed_ct = NumResJan;
if Month = 4 then bed_ct = NumResFeb;
if Month = 5 then bed_ct = NumResMar;
else bed_ct = NumResApr;
```

Changing from one line of code to six isn't too unwieldy, but consider instead the variables making up the SEIR status arrays for residents and HCWs. Some LTCFs in the surveillance data had hundreds of residents and HCWs. Trying to code a variable for each resident and each HCW in each LTCF directly would have resulted in potentially thousands of lines of additional code in order to accomplish what was accomplished with arrays using only a few lines of code. Similarly, incorporating the counter variables into statements utilizing the arrays likely prevented the need for thousands of additional lines of code while enabling identical statistical analyses. With the interplay between the arrays in the model being complicated enough as it is, trying to duplicate that interplay for each of the variables individually without the benefit of arrays would clearly have increased the difficulty of the project by substantial margins.

DISCUSSION

The primary implication of the model results is the verification of the CDC's guidelines for annual vaccination of all HCWs (Centers for Disease Control and Prevention 2011). Using real data from LTCFs in New Mexico, the model predicted that even 100% vaccination coverage among HCWs would not reduce the risk enough to be considered a threshold. This suggests that every additional HCW vaccinated against influenza contributes meaningfully to the protection of the resident population. This supports the CDC's recommendation that 100% of HCWs get vaccinated. However, the presence of a curvilinear trend suggests that the greater the percentage of HCWs that are vaccinated, the greater the protection for the residents at each level of HCW vaccination coverage, implying that some degree of herd immunity could be induced, rather than the strictly linear protection that would be provided by simply reducing the number of potentially infected contacts. This leaves some room for the possibility that a herd immunity threshold

could potentially be attained under more favorable circumstances. What those circumstances may be is beyond the scope of this study.

Regardless of the specific practical and policy implications of the study results, it is important to note that they might not have been obtained without the SAS ARRAYS that made the model design feasible. From reading in preexisting data to saving newly generated data, the simplification of the code made possible by the arrays made the model easier to design and debug, likely saving a substantial amount of coding and troubleshooting time. The improvement in organization provided by the arrays also simplified statistical analyses by reducing the difficulty in incrementing the relevant analysis variables appropriately.

This study has a number of limitations. As with any mathematical model, the epidemic model described in this paper is subject to the underlying assumptions made in its development. The parameters used in the model, while backed up with empirical data wherever possible, may not sufficiently describe the conditions that were experienced in New Mexico during the influenza season of 2006-2007, nor may they accurately describe what would transpire were the hypothetical circumstances explored in the study to actually occur. Furthermore, the restriction in the surveillance data used in the model to a single influenza season may limit the generalizability of the results. Different influenza strains and different influenza vaccines in different seasons may respond differently in similar circumstances. Also, another limitation is that herd immunity in small populations is not well-defined (van den Dool et al. 2008). The presence of a curvilinear trend may not be the best indicator of herd immunity in a LTCF.

CONCLUSION

To conclude, using arrays in the design of an epidemic model for herd immunity against influenza in LTCFs contributed to the simplicity of the model and, ultimately, to the ability of the model to answer the research question of interest. Arrays were used successfully to read in real surveillance data from LTCFs in New Mexico collected during the 2006-2007 influenza season, and other arrays were used successfully for generating and storing model predictions. Counter variables to be used in statistical analyses were successfully combined with array references and used in a mixed model and other analyses. The mixed model revealed the existence of a curvilinear trend, suggesting herd immunity was induced, but not sufficiently to indicate the presence of a herd immunity threshold, corroborating previous findings (van den Dool et al. 2008) and supporting recommendations for vaccination of HCWs made by the CDC (2011). SAS ARRAYS were important in the attainment of these findings and can serve as a powerful tool in epidemic modeling.

REFERENCES

- [Anonymous]. 2006. "Arcares jaarverslag 2005." Utrecht: Dutch National Association for Nursing and Care.
- Centers for Disease Control and Prevention. "Interim Guidance for Influenza Outbreak Management in Long-Term Care Facilities." Seasonal Influenza (Flu). December 19, 2011. Available at http://www.cdc.gov/flu/professionals/infectioncontrol/ltc-facility-guidance.htm.
- Centers for Disease Control and Prevention. "Leading causes of death." *FastStats*. January 11, 2013. Available at http://www.cdc.gov/nchs/FASTATS/lcod.htm.
- Keelan, Stephen. 2002. "Off and Running with Arrays in SAS®." *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*. Cary, NC: SAS Institute Inc. Available at http://www2.sas.com/proceedings/sugi27/p066-27.pdf.
- Langston, Rick. 2005. "Efficiency Considerations Using the SAS® System." *Proceedings of the Thirtieth Annual SAS® Users Group International Conference*. Cary, NC: SAS Institute Inc. Available at http://www2.sas.com/proceedings/sugi30/002-30.pdf.
- Pearson, Michele, Bridges, Carolyn, and Harper, Scott. February 24, 2006. "Influenza Vaccination of Health-Care Personnel." *MMWR Recommendations and Reports*. 1-16. Centers for Disease Control and Prevention.
- SAS Institute Inc. "Using Arrays in SAS® Programming." *support.sas.com*. Retrieved February 7, 2014. Available at http://support.sas.com/resources/papers/97529_Using_Arrays_in_SAS_Programming.pdf.
- SAS Institute Inc. "ARRAY statement." *support.sas.com*. Retrieved February 7, 2014. Available at http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000201956.htm.
- Van den Dool, Carline, Bonten, Marc, Hak, Eelko, Heijne, Janneke, Wallinga, Jacco. October 28, 2008. "The Effects of Influenza Vaccination of Health Care Workers in Nursing Homes: Insights from a Mathematical Model." *PLoS Medicine*.

e200. Available at http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0050200.

- Vynnycky, Emilia and White, Richard. July 15, 2010. *An Introduction to Infectious Disease Modelling*. Oxford, NY: Oxford University Press.

- Waller, Jennifer. 2010. "How to Use ARRAYS and DO Loops: Do I DO OVER or Do I DO i?" *Proceedings of SAS® Global Forum 2010 Conference.* Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings10/158-2010.pdf.

- Wendelboe, Aaron, Avery, Catherine, Andrade, Bernardo, Baumbach, Joan, and Landen, Michael. August 17, 2011. "Importance of Employee Vaccination against Influenza in Preventing Cases in Long-Term Care Facilities." *Infection Control and Hospital Epidemiology*. 990-997. Chicago, IL: The University of Chicago Press.

ACKNOWELDGEMENTS

The New Mexico Department of Health provided assistance and support during the collection of the LTCF surveillance data used in this project. No external sources of funding were used to conduct the surveillance.

RECOMMENDED READING

An Introduction to Infectious Disease Modeling (see Vynnycky and White 2010 in References)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Carl Grafe
Organization: University of Utah
Address: 421 S Wakara Way Ste 140
City, State ZIP: Salt Lake City, UT 84108-3514
Email: Carl.Grafe@utah.edu
Web: http://medicine.utah.edu/bmi/