Multi-scale models and data for infectious diseases: A systematic review

Lauren M Childs 1, Fadoua El Moustaid 2,3, Zachary Gajewski 2,3,4, Sarah Kadelka 1, Ryan Nikin-Beers 1,5, John W. Smith, Jr. 4, Melody Walker 1, Leah R. Johnson Corresp. 2,3,4,6

1 Department of Mathematics, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, Virginia, United States
2 Department of Biological Sciences, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, Virginia, United States
3 Global Change Center, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, Virginia, United States
4 Department of Statistics, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, Virginia, United States
5 Department of Mathematics, University of Florida, Gainesville, Florida, United States
6 Computational Modeling and Data Analytics, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, Virginia, United States

Corresponding Author: Leah R. Johnson
Email address: lrjohn@vt.edu

The observed dynamics of infectious diseases are driven by processes across multiple scales. First is within-host, that is how an infection progresses inside a single individual (for instance viral and immune dynamics). Second is how the infection is transmitted between multiple individuals of a host population. The dynamics of each of these may be influenced by the other, particularly across evolutionary time. Thus understanding each of these scales, and the links between them, is necessary for a wholistic understanding of the spread of infectious diseases. One approach to combining these scales is through mathematical modeling. We conducted a systematic review of the published literature on multi-scale mathematical models of disease transmission to determine the extent to which mathematical models are being used to understand across-scale transmission, and the extent to which these models are being confronted with data. Following the PRISMA guidelines for systematic reviews, we identified 19 of 139 qualifying papers across 30 years that include both linked models at the within and between host levels and that used data to parameterize/calibrate models. We find that the approach that incorporates both modeling with data is under-utilized, if increasing. This highlights the need for better communication and collaboration between modelers and empiricists to build well-calibrated models that both improve understanding and may be used for prediction.
Multi-scale models and data for infectious diseases: A systematic review

Lauren M. Childs¹, Fadoua El Moustaid²,³, Zachary Gajewski²,³,⁴, Sarah Kadelka¹, Ryan Nikin-Beers¹,⁶, John W. Smith, Jr⁴, Melody Walker¹, and Leah R. Johnson²,³,⁴,⁵,*

¹Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, USA
²Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, USA
³Global Change Center, Virginia Polytechnic Institute and State University, Blacksburg, USA
⁴Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, USA
⁵Computational Modeling and Data Analytics, Virginia Polytechnic Institute and State University, Blacksburg, USA
⁶Department of Mathematics, University of Florida, Gainesville, FL 32611

Corresponding author:
Leah R. Johnson*
Email address: lrjohn@vt.edu

ABSTRACT

The observed dynamics of infectious diseases are driven by processes across multiple scales. First is within-host, that is how an infection progresses inside a single individual (for instance viral and immune dynamics). Second is how the infection is transmitted between multiple individuals of a host population. The dynamics of each of these may be influenced by the other, particularly across evolutionary time. Thus understanding each of these scales, and the links between them, is necessary for a holistic understanding of the spread of infectious diseases. One approach to combining these scales is through mathematical modeling. We conducted a systematic review of the published literature on multi-scale mathematical models of disease transmission to determine the extent to which mathematical models are being used to understand across-scale transmission, and the extent to which these models are being confronted with data. Following the PRISMA guidelines for systematic reviews, we identified 19 of 139 qualifying papers across 30 years that include both linked models at the within and between host levels and that used data to parameterize/calibrate models. We find that the approach that incorporates both modeling with data is under-utilized, if increasing. This highlights the need for better communication and collaboration between modelers and empiricists to build well-calibrated models that both improve understanding and may be used for prediction.

INTRODUCTION

In the study of biological systems, phenomena are often observed at either the between-host scale or the within-host scale. At the between-host scale, which may include how a disease spreads among organisms in a population, or the within-host scale, which may include intracellular or inter-cellular interactions with an invading pathogen. Since biological systems do not exhibit a clear separation of temporal or spatial scales, there has been increased interest in recent years in how interactions at one scale can affect interactions at the other.

Mathematical and computational modeling, which has a rich history of application to the dynamics of ecological systems and infectious diseases, has been used to study phenomena at both scales. At the between-host scale, classic compartmental models like the SIR model, which models the interactions between susceptible individuals $S$, infected individuals $I$, and recovered individuals $R$, have been used to...
predict the spread of infectious diseases between individuals in a population (Kermack and McKendrick, 1991, 1927; Anderson and May, 1992). At the within-host scale, models such as the $TIV$ model of viral dynamics, which models the interactions between target cells $T$, infected cells $I$, and virus $V$, were used to understand viral load within hosts (Perelson et al., 1996; Nowak and May, 2000).

To understand the outcomes produced by the interactions in and between different scales, a multi-scale model that links the scales may be constructed. For example, an $SIR$ model may be used to model the spread of a viral disease in a population. If the transmission rate between hosts is dependent on the outcome of the viral load from a $TIV$ model (since higher viral loads often are associated with higher disease transmission), the models at the between-host scale and the within-host scale depend on one another, and are thus considered linked. These models can be diverse in their structure and formulation (Garira, 2017). Thinking about the implications across scales is important but is also challenging as the relationships are often complex, nonlinear and, therefore, unintuitive. Previously, theoretical models of multi-scale phenomena have been reviewed (Mideo et al., 2008; Reiner et al., 2013; Dorratoltaj et al., 2017; Murillo et al., 2013; Severins, 2012). In 2015, Handel and Rohani highlighted the need for a better incorporation of data into multi-scale models (Handel and Rohani, 2015).

In this review, we aim to illuminate the state of the field joining experimental data with mathematical and computational models that bridge multiple scales. In doing so we expect to identify potential gaps in understanding and methodology. To do so, we examine papers that incorporate models that contain both within-host and between-host model components as well as utilize data. While we have related an example that involves the linking of two compartmental models in the context of a viral disease, we do not restrict our search to only compartmental models or those of viral disease. We find 19 papers which contain both (i) the within-host and between-host connection and (ii) data. In section 2, we explain how we searched for and chose papers. In section 3, we explain trends of the models in the papers we selected. We then conclude in section 4 with some overall thoughts on the current literature using multi-scales models with data.

**SURVEY METHODOLOGY**

To perform our systematic review we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). PRISMA is a standard protocol for conducting a systematic review or a meta-analysis. The flowchart showing our procedures are presented in Fig. 1.

We searched All Databases on Web of Science using the search terms “within-host between-host pathogen* transmi*” for papers published before November 30, 2017. We obtained 159 results, for which we screened abstracts (Fig. 1A). We initially eliminated 20 search results, which included duplicates and other results that were not papers. Further, there was one paper that could not be obtained (Verenini, 1983). This left us with 139 papers, which we initially screened by the abstract.

In the initial abstract screening phase, two randomly assigned people (i.e., two of LMC, FEM, ZG, SK, RNB, MW, or LRJ) separately categorized each of the 139 papers into three categories based on whether it included a linked model with data: ‘Yes’, ‘Maybe’, and ‘No’. A linked model was defined as a mathematical model that includes at least two scales, within-host and between-host, as well as some explicit link between the scales. The abstract was labeled as follows: ‘Yes’ if it appeared to describe both a linked model with data; ‘Maybe’ if it either (i) clearly described a linked model but was unclear on data, or (ii) clearly referred to data, but was unclear if the included model was linked; ‘No’ if it did not meet any of the above criteria, was obviously a review, or obviously out-side the scope of our review. A set of study properties (Fig. 1B, Q2.1 - Q2.7) were also collected for each of 110 papers at the abstract screening stage including the focal host species, other mentioned species, the type of infection, and the main results of the paper. Study properties were not recorded for the other 29 papers as they were either review papers (15) or deemed out of scope (14).

If an abstract was labeled with two ‘Yes’ or with one ‘Yes’ and one ‘Maybe’, we retained the paper for full paper screening; if an abstract was labeled with two ‘No’ we excluded the paper from screening. If an abstract was labeled with one ‘Yes’ and one ‘No’, we reviewed the abstract collectively to relabel it to either two ‘Yes’, two ‘No’, or one ‘Maybe’. We obtained 40 abstracts labeled two ‘Yes’ or one ‘Yes’ and one ‘Maybe’, 36 abstracts labeled two ‘No’, 34 abstracts labeled one ‘Maybe’ and one ‘No’, and 29 abstracts labeled two ‘Maybe’. If an abstract was labeled with one ‘Maybe’ and one ‘No’, the person who labeled ‘Maybe’ was assigned to skim the paper to decide if the paper should be kept or eliminated. If an abstract was labeled with two ‘Maybe’, a third randomly assigned person was assigned to skim the paper.
**Figure 1. Schematic of survey methodology.** (A) PRISMA flowchart showing the inclusion of papers. “Non-papers” refers to database entries that were figures or codes. (B) Schematic of the screening and evaluation questions used. Dashed lines indicate links between questions that were conditional, i.e., answering the second question/box depended on the answers to the earlier question. For example, details on the study properties (Q2.1 - Q2.7) and questions from the final screening stage (Q3.1 - Q3.3) were only collected for the 139 papers that were retained following the abstract screening stage. Questions in boxes 4 through 8 were completed for all 19 papers that remained following the final screening stage. Questions are found in Text S1; Responses are found in Tables S1-S8; References to all included papers are found in Text S2; References to all excluded papers are found in Text S3; All recorded data can be found in our Supplemental Data Sets.

to decide whether it should be kept or eliminated. A paper was kept if it appeared to have a linked model and/or data, but still was unclear if it had both; the paper was excluded otherwise. Once this process was completed, we kept 46 papers for further screening, and excluded 93 papers based on the abstracts. The reason for exclusion (lacking data, lacking a model, lacking a within-host component, lacking a between-host component, review, or another reason, which needed to be described) was recorded for all 93 papers excluded at this stage (Fig. 1B, Q3.2 - Q3.3).

We then conducted a final screening of the remaining 46 papers by having two individuals (randomly assigned from the full author list) read through the full text of each paper. During this step, a final determination was made for each paper whether to keep it for further analysis or to exclude it. A paper was kept if it contained a linked model with data; a paper was otherwise excluded. The reason for exclusion (lacking data, lacking a model, lacking a within-host component, lacking a between-host component, review, or another reason, which needed to be described) was recorded for all 27 papers excluded at this stage (Fig. 1B, Q3.2 - Q3.3). In all, we included 19 papers in the full analyses (Fig. 1A).

For the papers that were included, we answered a detailed set of questions, which described important aspects of the model (such as the transmission route), how the models were linked, how the data was used in the model, etc. (Supplemental Text S1, Fig. 1B, Q4 - Q8). We further characterized the journal in which each paper appeared as a general audience journal, a specialized biological journal, a primarily mathematical/computational journal, or a biology sub-discipline journal (Fig. S1).

**RESULTS**

**Traits of included compared to excluded papers**

Our initial search yielded 139 papers published over the span of than 30 years. While the earliest included paper was from 1990 (Dwyer et al., 1990), the next papers that met our requirements were published 15 years later (Cooper and Heinemann, 2005; McKenzie and Bossert, 2005). In the interim, a few more papers were published, but interest in this general area grew quickly starting in 2005. Both the number of papers loosely related to the topic (i.e. those excluded) and papers meeting our criteria to include both models and data (i.e. those included) increased in that time frame (Fig. 2A).
Figure 2. Summary of papers considered. Both included and excluded papers by (A) year of publication, (B) host species, and (C) the reason for exclusion (only for excluded papers). Papers were classified as included (gray), out of scope (orange) or excluded (blue) for (A) and (B). ‘Out of scope’ designated papers that literally included the search terms but were not topically related.

Papers spanned a variety of host species systems (Fig. S2). Infections of humans were, not surprisingly, the most common in both the excluded (28/120) and included categories (8/19). Although non-human mammals (20 overall) and invertebrates (14 overall) were the next largest individual categories of hosts overall, no papers focusing on invertebrates met the criteria for inclusion. Humans and non-human mammals also comprised a larger proportion of papers included than the other categories of host species (8/19 and 5/19, respectively). The most common reason for exclusion was a lack of data being used with the model (34%) followed by no model (20%) (note, that only one reason was recorded for each paper). That is, many papers explore within- to between-host transmission either from a modeling or empirical perspective, but many fewer link the models robustly to data. Recently, there have been a number of review papers on multi-scales models with data, another common reason for exclusion (12.5%).

Traits of included papers

We considered whether the aim of each paper was primarily strategic (trying to understand underlying dynamics) or primarily tactical (trying to make predictions) (Nisbet and Gurney, 1982). Of the papers examined, most were classified as primarily strategic and very few papers as primarily tactical. Only one paper was classified as both strategic and tactical (Vrancken et al., 2014) (Table S4.2). included papers were rarely found in highly specialized non-mathematical journals (2/19), but were relatively equally spread between mathematically focused journals, biology focused, and for a general audience (Fig. S1).

Figure 3. Focal host species and infection types for included papers. (A) Type of infection across host species for included papers as bacterial (gray), fungal (orange), macroparasite (blue), multiple (green), protozoa (yellow), viral (dark blue), or other (red). (B) Modeled transmission route across infection types for included papers as direct contact (gray), indirect contact (orange) or multiple routes (blue).
**Infection, host, and transmission categorization**

We found that the majority of the papers and models focused on a single infection. Most infection types were viral (10/19), with protozoa being the second most common (4/19). The host species were predominately mammals (13/19), of which eight were human hosts (Fig. 3A). Most papers modeled transmission as direct contact across infection types. Half of those with protozoa infection type were modeled by indirect contact. The only paper included that modeled a bacterial infection assumed indirect transmission (Chen et al., 2013). Of the papers which modeled viral infections, there was one which was indirect (Handel et al., 2014) and one with multiple modes of transmission (Handel et al., 2013) (Fig. 3B).

**Figure 4. Types of modeling framework used in included papers.** The x-axis shows the model types used in the within-host part of the model while the y-axis shows the model types used in the between-host model. The dots’ diameter represents how many papers used a particular framework.

**Model characteristics**

The multi-scale models reviewed are composed of three parts: the within-host model, the between-host model, and the linking mechanism. Approximately two thirds (12/19) of papers primarily investigated how the within-host dynamics affect the between-host dynamics; only one paper focused on the impact of the between-host dynamics on the within-host dynamics (Handel et al., 2014). The remaining third of the papers either examined both of the above directions of impact (i.e. how the within-host and between-host dynamics effect each other) or the influence of within- or between-host dynamics on another factor in their model (Table S4.3).

In a multi-scale model, the within-host component and between-host component are both modeled explicitly. We characterized each of the within-host and between-host models used as either a deterministic model, an individual-based model (IBM), a statistical model, or a stochastic model. Figure 4 shows the types of within-host and between-host models used in the included papers. Most studies used the deterministic model type at least once, either for within- or between-host models and sometimes for both. In the included papers, within-host models were most commonly deterministic (9), followed by statistical (7), individual-based (2), and stochastic (1). In contrast, for the between-host models, the vast majority were deterministic (11), with a lower and more evenly distributed representation of statistical (3), individual based (3), and stochastic (3). One study used an IBM model type for both the within-host model and the between-host model (van Dorp et al., 2014). In general, studies did not typically use the same modeling approach for both the within- and between-host components. As for host type, there was no evident correlation between model types and the focal host species used in the model (Fig. ??).

Within- and between-host models can be linked in three different ways: within- to between-, between-to within-, or bi-directionally. Among the included papers, ten of the studies linked the within-host model to the between-host model while nine linked bidirectionally - both within-host to between-host and between-host to within-host (Table S5.5). None of the included papers only linked the between-to within-host model.

To link the within-host and between-host models, a linking mechanism was needed, which we categorized either as a state or a trait. Linking via a state meant that an outcome of the model was
Figure 5. Mechanisms used to link between and within-host models together. The number of included papers that used the each of the (A) within-host linking mechanisms and (B) between-host linking mechanisms to connect the models together.

used; for example, the pathogen load at the within-host level or the number of infected individuals at the between-host level. In contrast, a trait was a parameter of the model; for example, the pathogen growth rate at the within-host level or the transmission rate at the between-host level. The model framework was categorized in one of three ways: linked only by states, only by traits, or by both traits and states. Furthermore, models could also have multiple linking mechanisms. In the included papers, nine studies used state variables, three used trait variables, and seven used both (Table S5.6).

Within-host models (Fig. 5A) are linked to the between-host models mostly via the pathogen load, with more than half the papers using this linking mechanism (14/19). Pathogen growth rate was the second most used trait to link the within- to between-host model (4/19 papers). All other within-host linking mechanisms were used in two or fewer papers. Between-host models were also linked into the within-host models (Fig. 5B) based on primarily a single trait, the transmission rate (13/19). All other between-host linking mechanisms were used in at most three papers.

Figure 6. Role of data in multi-scale modeling efforts. (A) Scale (within-host, linking, or between-host) at which data was incorporated (orange) in the multi-scale models. Some models used data at more than one level. (B) How the data was incorporated into the models: bottom-up, i.e. fitting traits (orange); top-down, i.e. fitting states (green); both (gray) or other (blue).

Role and method of data incorporation
All papers that passed the screening criteria utilized data in at least one component: the within-host component, the linking mechanism, or the between-host component. Even among the relatively small sample of papers that included data at all in multi-scale models, most did not use it for more than one level of their model (Fig. 6A). While most of the included papers (17/19) used data at the within-host level, only six papers used data at both the within-host and between-host levels, of which only three also
used data for the linking mechanism. Papers that included data for both linking mechanisms and the
between-host level also included data for the within-host level.

Across all model scales, bottom-up, i.e., fitting of traits, was utilized more than top-down, i.e. fitting of
states, or other methods (Fig. 6B). None of the papers used a mixture of bottom-up and top-down data
fitting at different levels, although one paper did not specify explicitly how the data was incorporated
(Cooper and Heinemann, 2005). For data fitting that was bottom-up, the majority of papers (6/10 within-
host, 3/6 linking mechanism) used maximum likelihood or least-squares (3/10 within-host, 3/6 linking
mechanism). Bayesian inference, although a popular statistical method, was only used twice in the papers
considered (Fig. S4). Only a single paper (Volz et al., 2017) recorded using multiple fitting methods at
the same scale, and most papers used the same fitting method across all scales.

**DISCUSSION AND CONCLUSION**

Our objective in this review was to determine how multi-scale infectious disease models are used when
they directly incorporate data. We focused on which species are modeled, which pathogens are modeled,
which types of models are used, how the within-host and between-host dynamics are linked, and at what
level data has been used. We found that it was most common for these models to describe a human
population, to model a viral disease, to use a deterministic model at either level, to link the pathogen
load at the within-host level, to link the transmission rate at the between-host level, and to use data at the
within-host level. It was least common for these models to describe a plant, fish, reptile, or amphibian
population, to model a bacterial, macroparasite, or fungal infection, to use a stochastic model at either
level, to link host symptoms at the within-host level, to link the host recovery rate at the between-host
level, and to use data at the between-host level.

We speculate on the reasons for these outcomes. As human disease has tangible consequences directly
impacting the wider population, it is unsurprising that the primary host species to examine these multi-
scale interactions was found to be humans. However, the importance of other species both economically
and ecologically leaves the door open for further study of these interactions. The dominance of viral
disease as the focal pathogen likely results from the rich history of mathematical modeling in viral
disease as well as their prominence in the human community. In choosing which type of model to use,
deterministic models do not include the mathematical and computational complication of randomness
as stochastic models do, making them often easier to simulate and analyze. Linking the within-host and
between-host scales is challenging. Many studies, thus, defaulted to the standard assumption that a higher
pathogen load often correlates with a higher chance of disease transmission, making pathogen load the
simplest way to link the within-host and between-host scales. Other linking mechanisms are often difficult
to model because there may not be an obvious relationship in how two elements at different scales affect
one another. The incorporation of data was primarily at the within-host level, perhaps stemming from
the fact that some of these relationships can be obtained through laboratory based research. In contrast,
between-host data may often require large-scale resources and monitoring.

We were quite surprised that our search yielded only nineteen papers that included both the across-
scale modeling and substantial use of data. It is possible that our particular search terms may have been
overly restrictive. For instance, the search term “pathogen” may be less likely to be used to describe
infectious macro-parasites (e.g., worms). None the less, our relatively small included set indicates that
there is considerable scope for further work to be done in the area of data-driven multi-scale modeling
of infectious diseases. Given the specific results of our review, we propose that future research could
productively focus on i) exploring alternative linking mechanisms and ii) incorporating more and varied
data at all scales.

Most studies we reviewed appeared to use the simplest assumption to link the within-host and between-
host scales, namely, linking the pathogen load at the within-host level to the transmission rate at the
between-host level. While this assumption may be appropriate for some diseases, there are other potential
mechanisms that could be used to provide links between scales, perhaps along with pathogen load. For
example, these could include how host immunity affects the transmission rate or how pathogen load
affects the pathogen virulence among the population. Within-host data of antibodies, when available,
could be used as a measure of host immunity. Accounting for these interactions could produce models
that make complementary or potentially divergent predictions of transmission outcomes, and in turn be
used to elucidate the effects different treatments have on disease spread.

The lack of data was the major reason that our search only uncovered nineteen papers (Fig. 2C). Thus,
a major gap in bridging within-host infection dynamics and between-host transmission is the existence and incorporation of data. This appears not to have improved significantly since a similar observation in 2015 (Handel and Rohani, 2015). Although there are cases where appropriate data for a model does not currently exist and must be collected in a new experiment, a greater effort should be put forth to work with and incorporate existing data sets. This is especially true at the population level, where data are particularly difficult and expensive to collect. In addition, far more of the papers we examined included data at the within-host level (Fig. 6A), likely due to the accessibility and scale of data that can be collected in a lab setting. Along with more data overall, the incorporation of more varied data at a variety of scales will enhance the utility of multi-scale disease modeling.

In summary, important results about disease spread can be gleaned from modeling the interactions at both the within-host and between-host scales. While current research has mainly focused on simplistic assumptions, we believe that including additional complexities in future models may help to better explain observations from the field. Multi-scale modeling provides a great opportunity for empiricists and theorists to work together, and to contribute to the understanding of the drivers, treatments, and control of infectious disease.
**SUPPLEMENTAL FIGURES**

**Figure S1. Journals where included papers appeared.** (A) Counts of included papers from different journals. (B) Counts of different journal types. General audience journals included *Philosophical Transactions of the Royal Society B: Biological Sciences*, *PLOS One*, and *Proceedings of the Royal Society of London B: Biological Sciences*. Primarily mathematical and computational journals included *American Naturalist*, *PLOS Computational Biology*, and *Journal of Theoretical Biology*. Specialized journals included *Molecular Biology and Evolution* and *Preventive Veterinary Medicine*. Sub-discipline journals included *Ecology*, *Ecological Monographs*, *Evolution*, and *Journal of Virology*. All journals of the 19 included papers are included.

**Figure S2. Reason for exclusion of papers by host species.** The reason for exclusion was categorized as no between-host component (gray), no data (orange), no model (blue), or no within-host component (green). All other reasons were included under other (yellow).
Figure S3. Types of focal host species and the modeling type. Types of focal host species used in the within-host and between-host models and the modeling type used to represent the model components. Model types were classified as deterministic (gray), individual-based (orange), statistical (blue) or stochastic (green).
Figure S4. Method used in data fitting at each scale. Three fitting methods were considered: Bayesian inference (gray), least squares (orange), maximum likelihood (blue). All other fitting methods were included under other (green). Different fitting methods could be used in the same papers for different scales.

REFERENCES


