Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Exposure, hazard, and survival analysis of diffusion on social networks

Jiacheng Wu[a], Forrest W. Crawford[b,c,d]*, David A. Kim[e], Derek Stafford[f], Nicholas A. Christakis[d,g,h,i]

Sociologists, economists, epidemiologists and others recognize the importance of social networks in the diffusion of ideas and behaviors through human societies. To measure the flow of information on real-world networks, researchers often conduct comprehensive sociometric mapping of social links between individuals, then follow the spread of an "innovation" from reports of adoption or change in behavior over time. The innovation is introduced to a small number of individuals who may also be encouraged to spread it to their network contacts. In conjunction with the known social network, the pattern of adoptions gives researchers insight into the spread of the innovation in the population and factors associated with successful diffusion. Researchers have employed widely varying statistical tools to estimate these quantities, and there is disagreement about how to analyze diffusion on fully observed networks. Here, we describe a framework for measuring features of diffusion processes on social networks using the epidemiological concepts of exposure and competing risks. Given a realization of a diffusion process on a fully observed network, we show that classical survival regression models can be adapted to estimate the rate of diffusion, and actor/edge attributes associated with successful transmission or adoption, while accounting for the topology of the social network. We illustrate these tools by applying them to a randomized network intervention trial conducted in Honduras to estimate the rate of adoption of two health-related interventions – multivitamins and chlorine bleach for water purification – and determine factors associated with successful social transmission.

Keywords: competing risks, diffusion of innovations, social network

[a] *Department of Biostatistics, University of Washington, Seattle, WA 98105, U.S.A.*
[b] *Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, U.S.A.*
[c] *Department of Operations, Yale School of Management, New Haven, CT 06511, U.S.A.*
[d] *Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT 06511, U.S.A.*
[e] *Department of Emergency Medicine, Stanford University, Stanford, CA 94305, U.S.A.*
[f] *Department of Political Science, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[g] *Department of Sociology, Yale University, New Haven, CT 06511, U.S.A.*
[h] *Deparmtent of Medicine, Yale School of Medicine, New Haven, CT 06510, U.S.A.*
[i] *Department of Biomedical Engineering, Yale University, New Haven, CT 06511, U.S.A.*

* *Correspondence to: Forrest W. Crawford, 60 College St, LEPH 207, PO Box 208034, New Haven, CT 06510, U.S.A.*
Email: forrest.crawford@yale.edu

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

1. Introduction

Understanding the spread of new ideas, behaviors, and practices through human social networks is a major component of social science and public health research [1, 2]. Studies of the diffusion of innovations often follow adoption of a new or better product. For example, Ryan and Gross [3] tracked adoption of hybrid seed corn among farmers, Coleman *et al.* [4] followed diffusion of a medical innovation (a new antibiotic) through physician networks [see also 5–7], and Banerjee *et al.* [8] followed the adoption of a microfinance innovation in Indian villages. Many researchers have evaluated the spread of health-related interventions [9–11], especially those that seek to overturn local customs or that address sensitive topics like contraception [12–14] or household hygiene [15]. Data from online networks and exact observation of individual communication patterns have yielded studies of information diffusion through blogs, chain letters, Twitter, and other social networks [16–22].

Methodological approaches for analyzing social diffusion processes seek to uncover the reason, channel, and rate underlying the diffusion of an innovation through a human social network [2, page 10]. A major research direction is macroscopic, cascade-oriented models of diffusion in a large population [23–27], in which the adoption process is slow initially, accelerates in an intermediate stage, and finally slows as it reaches a saturation point. Another prominent framework is the threshold model, which assumes that each individual has an intrinsic exposure threshold that must be attained before he/she adopts the innovation. Exposure is usually modeled as the proportion of network alters who have previously adopted the innovation [28–30].

In addition to keeping track of the pattern of adoptions, researchers often attempt to measure the social or communication network connecting potential adopters before or during a diffusion study. Researchers have targeted two separate but related components of diffusion: individual-level factors associated with adoption, and the "spillover" or peer influence effect on adoption. Many researchers have formulated time-dependent event-history models to test the existence of a "network effect" [31–36]. These models associate the probability of adoption for an individual at a particular moment in time with the proportion of network neighbors who are prior adopters [1, 37]. Most are equivalent to logistic regression with individual adoption status as the outcome, and peer exposure to adopters as a (potentially time-varying) covariate [38]. For example, Valente [37, page 106] proposes the logistic model

$$\log\left(\frac{P_{jt}}{1-P_{jt}}\right) = \alpha_t + \boldsymbol{\beta}\mathbf{X}_{jt} + \gamma E_{jt}, \tag{1}$$

where P_{jt} is the probability that subject j adopts the innovation at time t, E_{jt} is the time-varying exposure defined as the proportion of j's network neighbors who adopted before t, α_t is a time-specific intercept, and $\mathbf{X}_{jt}$ is a vector of possibly time-dependent covariates. All subjects are assumed to be susceptible to adoption from the beginning of the study: the model assigns positive adoption probability to every subject j, even when their peer exposure is zero. A positive value of γ indicates that more network exposure to prior adopters is associated with higher probability of adoption. Extensions of these models have been proposed to incorporate spatial and temporal features of social diffusion processes [33–35, 39, 40].

Recent large-scale network intervention studies have successfully combined comprehensive sociometric data from online and real-world social networks with precisely observed adoption [8, 41, 42]. These modern diffusion studies share several key features: 1) researchers attempt to accurately and comprehensively measure the social or communication network of subjects eligible to adopt the innovation, 2) researchers have a mechanism for keeping track of the timing of adoption or behavior change, and 3) researchers observe the direction of transmission from one person to another in the social network. But application of traditional statistical modeling approaches to data from modern diffusion studies presents pitfalls for researchers. Traditional approaches sometimes treat adoption by individual subjects as conditionally independent [37], or ignore network structure by aggregating subjects into groups [43], resulting in biased estimates of contagion and lack of interpretability. Existing modeling approaches [e.g. 37, 44] often assume implicitly that adoption can occur even in the absence of peer exposure. However, this assumption may not hold in some study designs. For example,

Kim *et al.* [42] keep track of adoptions and transmission of heath-related interventions by giving subjects "tickets" carrying a unique identifier. Transmission of a ticket to another person, and redemption of the ticket in exchange for a product, constitutes adoption. Individuals whose network alters have not adopted, or have no tickets, are not eligible to adopt. A unified and rigorous approach to the statistical analysis of social network diffusion data would allow researchers to better uncover the dynamics of diffusion processes in experimental and observational studies, and could guide the design and implementation of future health-related intervention campaigns. In addition, statistical approaches for estimating diffusion dynamics on network edges may contribute to the development of approaches for rigorous causal inference in network settings [45].

Our objective here is to advance the statistical analysis of social network diffusion data, to develop methods flexible enough to accommodate the observed data from innovative new study designs [e.g. 42], and to provide tools that fit within a statistical framework familiar to sociologists, epidemiologists, and public health researchers. Our approach incorporates all available data into the analysis: the measured network, subject/link characteristics, the timing of adoptions measured continuously, and the direction of transmission/diffusion of the innovation. The key insight is that a rigorous time-dependent definition of network "exposure", borrowed from infectious disease epidemiology, permits principled estimation of the rate of diffusion and of individual characteristics associated with adoption in a traditional survival regression framework. We employ the notion of competing risks from analysis of time-to-event data to derive the likelihood of the diffusion process, while accounting for network topology and variation in vertex and edge attributes. We illustrate this new framework by estimating the rate of diffusion of two health-related interventions in a social network intervention trial in Honduras [42] and provide a network interpretation of the diffusion of the interventions.

2. Background

2.1. Terminology

We introduce generic terminology for diffusion studies on networks. Some of these assumptions have been articulated in related work on network diffusion processes in epidemiology [46]. A *seed* is a person to whom the innovation is initially introduced by the researchers. An *adopter* is someone who has adopted the innovation (in the context of the study), either because that person is a seed chosen by researchers or because the innovation has been transmitted to them via another adopter. We assume the directed graph of transmissions is observed, either using a ticket-passing design or by some other mechanism. A *susceptible* individual is one who has not yet adopted, but who is eligible or has a network contact who can transmit the innovation to them. By *transmission* we mean the social process by which the adoption of an innovation causes the susceptible neighbor to adopt. A *susceptible edge* in the network connects a prior adopter, who is able to transmit the innovation to a susceptible neighbor.

In ticket-driven studies, an adopter transmits the innovation by giving the ticket to a susceptible person who later redeems it, thereby becoming an adopter. In online studies, the "ticket" might be virtual and transmission amounts to sending an electronic message. The direction and timing of transmission may be fully observed in the sense that 1) the identity of the susceptible individual, 2) the identity of the prior adopter, and 3) the time of adoption or redemption of the ticket are all fully observed. Sometimes, tickets are exhaustible: transmission decreases the number of tickets held by the adopter by one. We also assume that a subject who adopts during the study is not eligible to adopt again, and hence is no longer susceptible.

2.2. Basic assumptions

We describe several assumptions that will guide development of a well-defined notion of network exposure. First, we assume that the social network connecting the members of the study population exists.

Assumption 1 (Network) *The population social network is a known undirected graph $G = (V, E)$ with no parallel edges or self-loops.*

Assumption 1 can be relaxed to accommodate directed graphs, but, for simplicity, we will assume here that the social network is undirected. Individuals are *vertices* in V, and their social links as *edges* in E. The network G determines who can transmit to whom.

Assumption 2 (Transmission across edges) *Transmission happens across susceptible edges in G connecting a prior adopter and a susceptible subject.*

When a subject adopts the innovation, that subject may be able to transmit the innovation to one of its network neighbors in G.

Define the directed transmission graph $G_T = (V_T, E_T)$, where V_T is the set of adopters, and E_T is the set of directed edges $(i, j) \in E_T$ indicating that i transmitted the innovation to j. Let $\mathbf{t} = (t_1, \ldots, t_n)$ be the ordered adoption times of each of the vertices in V_T. For convenience, we set $t_j = T$ for vertices who do not adopt, where T is the end of study, $j \in V$ but $j \notin V_T$. Let $\mathbf{X}$ be the collection of attributes for all vertices in V, and let $\mathbf{Z}$ be the collection of edge attributes for all edges in E.

Assumption 3 (Observed data) *We observe $(G, G_T, \mathbf{t}, \mathbf{X}, \mathbf{Z})$.*

2.3. Edge-wise hazard

The hazard of adoption is the instantaneous risk of adopting the innovation during the transmission process. Formally, let T_{ij} be the continuous waiting time for a prior adopter $i \in V$ to transmit an innovation to a susceptible network neighbor $j \in V$, with $\{i, j\} \in E$. Let t_i be the adoption time for i, and t_j be the adoption time for j if j adopts and the end of study T if j does not adopt. Obviously $t_i < t_j$. Note that the times t_i and t_j are measured relative to the beginning of the study while the edge-wise waiting time T_{ij} to adoption is measured from the moment t_i at which i adopts. We use t to denote absolute observation time relative to the beginning of the study, and τ to denote edge-wise waiting times. $T_{ij} = \infty$ if either i is not a prior adopter or j is not susceptible.

Definition 1 (Hazard) *Suppose $0 \leq t_i < t$ for $i \in V$. The* hazard *of transmission from i to $j \in V$ along the edge at absolute time t is*

$$\lambda_{ij}(t - t_i) = \lim_{\epsilon \to 0} \frac{\Pr(t_i + T_{ij} \in (t, t + \epsilon) | t_i + T_{ij} \geq t)}{\epsilon}, \tag{2}$$

for $t_i < t \leq t_j$, and $\lambda_{ij}(t - t_i)$ is non-zero only when i is connected to j, i adopts the innovation before j, and j is susceptible.

The edge-wise hazard $\lambda_{ij}(t - t_i)$ is defined to be zero if i has not yet adopted ($t < t_i$), or if j is not susceptible.

Definition 2 (Cumulative hazard) *The* cumulative hazard *is the cumulative hazard of adoption for transmission from prior adopter i to susceptible j up to time $t \leq t_j$,*

$$\Lambda_{ij}(t - t_i) = \int_{t_i}^{t} \lambda_{ij}(s - t_i)\, ds. \tag{3}$$

Let $F_{ij}(\tau) = \Pr(T_{ij} < \tau)$ be the cumulative distribution function of this waiting time, and $f_{ij}(\tau) = dF_{ij}/d\tau$ be its probability density function. Both $f_{ij}(\tau)$ and $F_{ij}(\tau)$ can be written in terms of hazard function $\lambda_{ij}(\tau)$ and cumulative hazard function $\Lambda_{ij}(\tau)$: $f_{ij}(\tau) = \lambda_{ij}(\tau) \exp[-\Lambda_{ij}(\tau)]$, and $F_{ij}(\tau) = 1 - \exp[-\Lambda_{ij}(\tau)]$.

Definition 3 (Exposure) *Let $j \in V$ be a susceptible subject. The exposure to j is*

$$E_j(t) = \sum_{i \in N_j} \lambda_{ij}(t - t_i), \tag{4}$$

where N_j is the set of network neighbors of j.

In words, exposure is the sum of the edge-wise adoption hazards from all prior adopters connected to the susceptible subject j.

Definition 4 (Cumulative exposure) *Let $j \in V$ be a susceptible subject. The cumulative exposure to j is*

$$CE_j(t) = \sum_{i \in N_j} \int_{t_i}^{t} \lambda_{ij}(s - t_i)\, ds. \tag{5}$$

In words, the cumulative exposure to j is the cumulative hazard from all prior adopters connected to the susceptible up to time t.

Consider a susceptible subject $j \in V$ at time t before j's adoption. For a prior adopter $i \in N_j$, let T_{ij} be the hypothetical waiting time for i to transmit the innovation to j. Note that $T_{ij} = \infty$ if either i has not adopted ($t < t_i$) or j is not susceptible ($t > t_j$). Adoption of the innovation by j occurs at time

$$t_j = \min_{i \in N_j} (t_i + T_{ij}). \tag{6}$$

The set of prior adopters connected to j, $A_j = \{i : i \in N_j, t_i < t_j\}$, represent sources of competing risks for transmission to j. All prior adopters in A_j can transmit the innovation to the susceptible subject j, but only the minimum of their corresponding edge-wise waiting times to transmission is observed. We borrow the terminology of competing risk from survival analysis that patients can die from multiple diseases, and, analogously, all prior adopters in A_j are competing to transmit the innovation to j.

Finally, we state an additional assumption that is common to most statistical models of network diffusion, but rarely made explicit, which makes possible rigorous statistical analysis using established tools from survival analysis.

Assumption 4 (Conditional independence) *Suppose $i, k \in V$ are prior adopters with adoption times t_i and t_k respectively. Furthermore suppose that $j \in N_i$ and $l \in N_k$ are susceptible, and either $i \neq k$ or $j \neq l$. Then the edge-wise waiting times T_{ij} and T_{kl} are conditionally independent given nodal attributes X_i, X_j, X_k, X_l and edge attributes Z_{ij}, Z_{kl}.*

In other words, when we condition on adoption status and node/edge attributes, the waiting times to adoption along *susceptible edges* are conditionally independent. It is not necessarily the case that the overall waiting times to adoption $t_i + T_{ij}$ and $t_k + T_{kl}$ are conditionally indepenent.

Proposition 1 *Let $\lambda_j(t)$ be the hazard of adoption to a susceptible subject $j \in V$ at time t. Under Assumption 4,*

$$\lambda_j(t) = \sum_{i \in N_j, t > t_i} \lambda_{ij}(t - t_i). \tag{7}$$

Proof is given in the appendix. In words, when we condition on the covariates X_j, X_i and Z_{ij} for $i \in N_j$, the hazard $\lambda_j(t)$ is the sum of the edge-wise hazards of transmission from network neighbors who are prior adopters. Note that (7) is the same as Definition 3 for exposure.

[Figure 1 about here.]

Figure 1 shows a hypothetical diffusion process on a network. Starting from an initial seed, labeled 1, diffusion occurs along the network edges. Vertices are numbered in the order of adoption. Vertices labeled by letters never adopt, but may experience exposure or hazard of adoption from their adopting alters. The first two rows show the adopters, susceptible edges, and exposed vertices just after each adoption event. The hazard/exposure for a particular susceptible individual increases over time with the addition of prior adopters connected to that individual. The last four rows show how hazard/exposure changes over time for each subject under constant edge-wise hazard of adoption. The exposure increases one step whenever the number of prior adopters connected to the subject increases. The area under the curve is the cumulative exposure experienced by each vertex over the course of the study.

3. Survival models of network diffusion

We now develop a flexible class of models for diffusion processes on networks, and show that these models can be formulated and fitted using the familiar framework of survival analysis. Let r_j denote the subject who transmits the innovation to the susceptible subject j. Let $r_j = 0$ in the situation where j is a seed or does not adopt the innovation. If i successfully transmits innovation to j before any other adopters, then $r_j = i$ and the edge-wise waiting time $T_{ij} = t_j - t_i$ is fully observed. On the other hand, two types of intervening events can cause observation of the waiting time T_{ij} to be *censored*. First, if $k \neq j$ transmits the innovation to j at time t_j before i, then we only observe $T_{ij} > t_j - t_i$ and the edge waiting time T_{ij} is censored. In this case, only the first transmission time is observed, and other longer waiting times are censored. Second, suppose t_i^* is the time that i uses its last ticket, or the end of the study, whichever comes first (if i receives no tickets, then $t_i^* = t_i$). Then we only observe the censored waiting time $T_{ij} > \min\{t_j, t_i^*\} - t_i$.

By Assumption 4, edge-wise waiting times T_{ij} are conditionally independent, given subject covariates X_i, X_j and edge covariates Z_{ij}. Let $t_{ij} = \min\{t_j, t_i^*\} - t_i$ and $S_i(t)$ be the set of susceptible individuals connected to the prior adopter i at time t. The likelihood is

$$\begin{aligned} L &= \prod_{i=1}^{n} \prod_{j \in S_i(t_i^+)} [f_{ij}(t_{ij})]^{\mathbb{1}\{r_j = i\}} \left[1 - F_{ij}(t_{ij})\right]^{\mathbb{1}\{r_j \neq i\}} \\ &= \prod_{i=1}^{n} \prod_{j \in S_i(t_i^+)} \lambda_{ij}(t_{ij})^{\mathbb{1}\{r_j = i\}} \exp\left[-\Lambda_{ij}(t_{ij})\right] \end{aligned} \tag{8}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function taking value 1 when its argument is true, and zero otherwise, t_i^+ is the time just after i's adoption, and n is the number of individuals who adopt the innovation. Below we describe several special cases corresponding to particular choices of the hazard function.

3.1. Example: constant hazard without covariates

Suppose we model $\lambda_{ij}(\tau) = \lambda$, a constant edge-wise hazard of transmission, for $\tau > 0$. Then edge-wise waiting times to transmission are exponentially distributed with rate λ. The likelihood becomes

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{|\{j:\, r_j = i\}|} \exp\left[-\lambda \sum_{j \in S_i(t_i^+)} t_{ij}\right] \tag{9}$$

and the maximum likelihood estimator of λ is

$$\hat{\lambda} = \frac{n-m}{\sum_{i=1}^{n}\sum_{j\in S_i(t_i^+)} t_{ij}}, \tag{10}$$

where m is the number of seeds. Intuitively, the estimated edge-wise rate of transmission is the number of non-seed adopters divided by the total edge-wise waiting time.

3.2. *Example: Weibull proportional hazard model*

The Weibull proportional hazard model has the multiplicative form,

$$\lambda_{ij}(\tau) = \exp(\delta + \alpha' X_i + \beta' X_j + \eta' Z_{ij}) k\tau^{k-1}, \tag{11}$$

where $k\tau^{k-1}$ is a time-varying baseline hazard common to all edge-wise waiting times. Subject-specific effects are captured by the exponential term, where δ is the intercept, and α, β, and η are coefficient vectors. The Weibull hazard is increasing in time when $k > 1$, decreasing when $k < 1$, and constant when $k = 1$. Estimation of $(\delta, \alpha, \beta, \eta)$ is performed by maximum likelihood. The likelihood is

$$L(\alpha,\beta,\eta) = \prod_{i=1}^{n-1}\prod_{j\in S_i(t_i^+)} \left[\exp(\delta + \alpha' X_i + \beta' X_j + \eta' Z_{ij}) k t_{ij}^{k-1}\right]^{\mathbb{1}\{r_j=i\}} \exp\left[-\exp(\delta + \alpha' X_i + \beta' X_j + \eta' Z_{ij}) t_{ij}^{k}\right]. \tag{12}$$

3.3. *Example: semi-parametric proportional hazards*

Cox's semi-parameteric proportional hazard model [47] is

$$\lambda_{ij}(\tau) = \lambda_0(\tau)\exp(\alpha' X_i + \beta' X_j + \eta' Z_{ij}), \tag{13}$$

where $\lambda_0(\tau)$ is a possibly time-varying baseline hazard common to all edges. The Cox model is semi-parametric because no parametric assumptions are made about the baseline hazard, but the covariate effects are assumed to multiply the baseline hazard. Treating $\lambda_0(\tau)$ as a nuisance function, estimates of the regression coefficients can be obtained by maximmizing the partial likelihood, assuming that all non-censored waiting time t_{ij} are distinct:

$$L(\alpha,\beta,\eta) = \prod_{(i,j):i\in N_j, r_j=i} \frac{\exp(\alpha' X_i + \beta' X_j + \eta' Z_{ij})}{\sum_{k=1}^{n}\sum_{l\in S_k(t_k^+)}\exp(\alpha' X_k + \beta' X_l + \eta' Z_{kl})\mathbb{1}\{t_{kl} > t_{ij}\}}. \tag{14}$$

The baseline hazard $\lambda_0(\tau)$ can be estimated by maximizing full likelihood as a function of baseline hazard [48, page 258].

4. Application: health-related interventions in rural Honduras

We now apply the survival regression methodology to a real-world diffusion study whose aim was to promote two health-related interventions – chlorine for water purification and multivitamins for micronutrient deficiencies – in rural Honduras [42]. The study was conducted in 32 isolated villages in Lempira, Honduras, providing an ideal environment for diffusion studies in distinct social networks, and comparison of the rates of diffusion in different villages. The social network of

subjects for each village was mapped by asking participants to identify spouses, siblings, and friends from a photographic census. Two villages received neither intervention.

The trial employed three targeting methods for seeds. Random targeting selected 5% of villagers as seeds, uniformly at random, in each village. Indegree targeting selected the 5% of villagers in each village with highest network degree as seeds. Nomination targeting was based on choosing a random alter nominated by each member of a 5% random sample of villagers, exploiting the "friendship paradox" whereby friends of random individuals tend to have higher network degree than the random individuals themselves [42, 49]. Initially targeted individuals (seeds) were given a product (chlorine or multivitamin), an associated educational intervention, and four tickets to distribute to network alters (first-wave) within the village who could redeem them in a local store for products. After redeption of tickets, these first-wave individuals also received four tickets for distribution to second-wave individuals. Redemption of a ticket is regarded as the adoption of the innovation in the context of the study, and ticket passing signifies the diffusion of the innovation. Each ticket was marked with a uniquely identifying number traceable back to the prior adopter, and the time of ticket redemption was recorded. One third of villages had seeds chosen by random targeting, 1/3 by indegree and 1/3 by nomination. Figure 2 illustrates the network diffusion of multivitamin adoption in Village 4.

[Figure 2 about here.]

In the analysis of the original study, Kim *et al.* [42] used the proportions of redeemed tickets over time as the primary village-level outcome to evaluate diffusion under the three targeting strategies for seeds. Kim *et al.* [42] also used a mixed-effects Cox model for adoption time (measured in days since the introduction of the intervention to the village's seeds) to estimate the effect of targeting methods on eventual adoption, treating non-adopting subjects' adoption times as censored. Since the primary outcome was the proportion of villagers who adopted the intervention, and not the dynamics of diffusion on network edges *per se*, Kim *et al.* [42] did not make use of data from the social network upon which diffusion was assumed to occur, except in the targeting of seeds.

4.1. Comparison across targeting methods

We first analyzed edge-wise diffusion times by constructing Kaplan-Meier survival curves [50] for edge-wise waiting times to adoption without adjusting for covariates. Figure 3 compares Kaplan-Meier estimates of the survival curve for three targeting methods on the adoption of multivitamin tablets and chlorine bleach. Lower Kaplan-Meier curves indicate faster edge-wise diffusion. For the multivitamin intervention, villages whose seeds were chosen by nomination targeting had the fastest edge-wise diffusion, followed by random targeting, and in-degree targeting. For the chlorine intervention, random targeting was associated with the fastest edge-wise diffusion, followed by in-degree and nomination targeting.

We also conduct log rank tests to test whether the unadjusted survival curves are significantly different. For the multivitamin intervention, log rank tests suggest that random targeting is significantly faster than indegree targeting ($p < 10^{-5}$), but adoption under nomination targeting is not significantly faster than under random targeting ($p = 0.146$). For the chlorine intervention, random targeting is not significantly faster than indegree targeting ($p = 0.155$), and nomination targeting is not significantly faster than random targeting ($p = 0.277$).

[Figure 3 about here.]

Figure 4 shows the cumulative edge-wise hazards. The first two days after exposure to prior adopters show the highest rate of adoption. The multivitamin intervention had a higher edge-wise diffusion rate than the chlorine intervention (reflecting its greater appeal in this setting).

[Figure 4 about here.]

4.2. *Baseline diffusion rate and covariate effects*

Next, we computed estimates of the baseline hazard of edge-wise transmission by fitting a Cox proportional hazards regression model for edge-wise waiting times to adoption. Table 1 shows the estimated coefficients from the Cox regression model. The first six covariates are measured at the village level, and the last four are characteristics of individual prior adopters. We estimated an edge-wise hazard ratio of 0.73 (95% CI 0.64-0.83) for multivitamin diffusion under indegree targeting compared to random targeting, adjusting for village-level characteristics and the prior adopter's characteristics. The edge-wise hazard ratio for the multivitamin intervention under nomination targeting is 1.05 (95% CI 0.92-1.19) compared to random targeting. After adjusting for covariates, we find that, across all waves of adoption and all villages, those assigned to nomination targeting exhibited faster edge-wise diffusion than random targeting for the multivitamin intervention, but the effect was not significant. In the original analyses, Kim *et al.* [42] estimated that, among the first-wave multivitamin tickets, nomination targeting had a significantly faster adoption rate than random targeting, while among second-wave multivitamin tickets, nomination targeting was faster than random targeting but was not significantly different. Our analysis provides an estimate of edge-wise diffusion rate that aggregates diffusion across two waves and provides a network-based interpretation of diffusion while adjusting for potential confounders. Our results generally agree with those described by Kim *et al.* [42] in that nomination targeting was faster than random targeting, though our estimates of effects differ in magnitude. However, the purpose of our method here is to estimate edge-wise diffusion rates, and to evaluate how interventions diffuse through specific network structures, rather than to characterize the aggregate effects of targeting methods on population-level adoption.

The edge-wise hazard ratio for chlorine tablet adoption under indegree targeting is 0.91 (95%CI 0.80-1.04) compared to random targeting. The edge-wise hazard ratio for chlorine adoption under nomination targeting is 0.99 (95% CI 0.83-1.16) compared to random targeting. This result is consistent with the result from Kim *et al.* [42] whose analysis also showed that the three targeting methods were not significantly different for the chlorine intervention. For both multivitamin and chlorine interventions, lower village socioeconomic status (SES) led to faster edge-wise diffusion, and male prior adopters were less likely to spread the innovation than female prior adopters.

[Table 1 about here.]

4.3. *Fixed effect for villages*

We included village-level fixed effects for the adoption of multivitamins and chlorine after controlling for prior adopter's attributes; the results are given in Tables 12 and 13 in the Appendix. Figure 5 shows the average village-level diffusion rate, defined as average expected number of transmissions per edge per day from Cox model with village fixed effects. The rate of diffusion differed greatly from village to village. Most villages exhibited faster edge-wise diffusion of the multivitamin intervention than chlorine, consistent with the finding in Kim *et al.* [42].

[Figure 5 about here.]

4.4. *Event-history model*

The event-history model (1) is an alternative approach to analyze diffusion studies on social networks [37, 38]. Table 2 shows the results of logistic regression from (1). Exposure is the proportion of network neighbors who are prior adopters. The odds of adoption for individuals with 100% exposure is 1.33 (95%CI 1.02-1.74) times larger than those with zero exposure in the multivitamin intervention. The odds of adoption for individuals with 100% exposure is 1.29 (95%CI 0.95-1.76) times larger than those with zero exposure in the chlorine intervention. The exposure in the alternative model corresponds to the individual hazard of adoption defined in (7) in the edge-wise diffusion model if the edge-wise hazard is a constant. Exposure in this logistic regression model can be interpreted as a special case of the sum of edge-wise hazard. Exposure to prior adopters in the multivitamin intervention is significantly different from zero, while exposure in the chlorine intervention is not.

[Table 2 about here.]

4.5. *Model comparison*

In addition to the analyses using the edge-wise hazard and event-history models, we conducted several additional analyses to compare model specifications and evaluate the assumptions of the edge-wise diffusion model. These results are given in the Appendix. We first compare the results to a logistic model [37, page 106]; by treating adoptions as realizations of Bernoulli trials, we compare the Akaike information criterion (AIC) [51] of the logistic and edge-wise diffusion models to show that the diffusion model exhibits better fit to the data from the Honduras experiment. Next, we evaluate the dependence of the adoption hazard (7) on the number (via the total hazard) of prior adopters, rather than the proportion, or average hazard. By dividing the total hazard $\lambda_j(t)$ by the degree d_j of the susceptible individual j, we introduce an offset $(-\log(d_j))$ in the edge-wise diffusion model; comparison of AICs shows that the original model exhibits better fit. We evaluate random effects/frailty terms for both prior adopters and susceptible individuals to account for possible actor-specific effects; we find that the AIC of the random effects model based on the integrated log partial likelihood is lower than that of the Cox diffusion model for the multivitamin intervention, but higher for the chlorine intervention.

We also evaluate Aalen's additive hazard model [52] to account for the possibility that some prior adopters may decrease the total hazard of adoption. While the hazard interpretation of the adoption rate $\lambda_j(t)$ for a susceptible j requires that it be positive, some of its constituent components $\lambda_{ij}(t)$, for particular prior adopters i, may be negative. We find five such edges under the multivitamin intervention and two edges for the chlorine intervention that have negative cumulative hazard up to the moment of adoption or censoring on the edge $\{i, j\}$. The additive model shows good overall fit, with slightly smaller Cox-Snell residuals.

Next, we assess a mixture cure-rate model based on the observation that some individuals never adopt the intervention, even when their network "exposure" is large. The cure model permits some edges to be "cured" so that no ticket is passed across them. The edge-wise waiting time distribution is estimated by the edge-wise diffusion model, and the cure probability model is logistic. The cure model exhibits smaller AIC than the edge-wise diffusion Cox model, suggesting that accounting for edges along which no adoption can occur improves the Cox model fit. Finally, we report estimated regression coefficients for the Honduras data under the Exponential and Weibull models of edge-wise diffusion, and village-level fixed effects.

5. Discussion

A major focus of contemporary social science and public health is the delivery of effective health and behavioral interventions in a social setting. Experimental studies in which researchers carefully control for network composition and information availability have demonstrated a significant contagious effect of health-related interventions [11, 42, 53–55]. Modern diffusion studies, in which the network is measured with as much precision as possible before experimental introduction of an intervention, hold promise for sidestepping many of the methodological challenges for traditional peer-influence analyses [1, 37]. But there is still a wide gap between what sociologists and public health researchers know about the social diffusion of behaviors, and the statistical tools at their disposal to design and analyze real-world network diffusion studies in the populations that stand to benefit the most from these interventions.

The proposed methodological framework leverages data that are often ignored in traditional approaches: the direction of information transmission, the network on which diffusion occurs, and measurements of network exposure in continuous time. The survival analysis framework provides a convenient method of "adjusting" for network topology, yielding inferences that are interpretable across network structures. The estimated parameters are readily interpreted in real-world terms: the diffusion rate per susceptible network link over the entire study period. The hazard model developed here also has an intuitive justification in terms of competing risks of transmission, which gives rise to the familiar additive

form of the individual-level hazard of adoption. The framework of survival analysis, familiar to public health researchers, epidemiologists, and many social scientists, should be straightforward to apply in future studies.

In this paper we assume that the network topology does not change during the study period. However, for some real-world networks, edges and vertices may appear or disappear during a given diffusion process. When dynamic network data are available, our proposed framework could be adapted, under particular assumptions about how the network dynamic process is related to the adoption process. For example, if edge deletion events occur independently of adoptions, then deletion of a susceptible edge before adoption occurs would result in censoring of the edge-wise adoption time. Likewise, addition of a susceptible edge could initiate an edge-wise adoption time.

Our reanalysis of the Honduras study has several limitations. First, we assumed that the redemption of tickets in exchange for a product signified the adoption of the innovation, but that may not always be true. In medical innovation studies, for example, patients may make use of a medication, but stop using it soon afterwards. Without long-term follow-up, it is impossible to determine whether adoption in the context of the study signifies long-term behavior change. Second, because the follow-up time in the Honduras study was relatively short, we assumed that adopters who had remaining tickets could pass a ticket to a susceptible alter at any time. However, the survival regression framework could easily accommodate cessation in the ability or willingness to transmit a ticket. For example, if tickets expire after a certain date, or if subjects become unwilling to pass a ticket, the waiting time to transmission and adoption by the alter would be censored before the end of the study. Third, if network information is not complete, the proposed method may be subject to bias because competing risks of transmission may not be correctly modeled [56]. Moreover, the social network may be accurately measured, but if participants pass their tickets to individuals not enumerated in the network census, this relevant network information might be missing, and estimates could be in error. Sensitivity analyses conducted by imputation of missing edges may be useful in exploring the magnitude of errors due to missing network information. Fourth, missing or incomplete information about adopters or susceptible subjects could result in bias. In this reanalysis of the data from Kim *et al.* [42], the identity of some ticket redeemers (adopters) was not recorded, or they were not present in the network census. We discarded data from a small number of adoptions by individuals not enumerated in the village network census. Fifth, the additive total hazard in the edge-wise diffusion model arises naturally from Assumption 4 (conditional independence). However, this assumption does not incorporate "synergistic" effects wherein the hazard of adoption increases super-linearly, or as a function of connections between prior adopters themselves. Likewise, our construction does not incorporate the possibility that some prior adopters may negatively influence the hazard of adoption in one of their susceptible alters (though we have explored this possibility in additional analyses in the Appendix).

In addition to descriptive inferences about the edge-wise rate of diffusion and factors associated with successful adoption, the models we develop here may help yield insights into the causal mechanisms that govern adoption of innovations in the network context. Statistical inference for causal peer effects may be complicated by treatment interference or contagion in outcomes [45, 57–61]. Existing approaches typically address treatment interference, in which the intervention applied to one unit affects the outcome of that unit and others [62, 63]. In the diffusion context, interference may also occur temporally between outcomes themselves via contagion/transmission processeses [64], or between multiple interventions diffusing simultaneously via "dueling contagions" [65]. Under particular causal assumptions, the diffusion models developed in this paper may have a causal interpretation, and could yield valid causal inferences for both the direct effect of an intervention on seed individuals, as well as the "spillover" or peer effects whereby network exposures influence adoption by individuals not directly targeted by the intervention. We are exploring these topics in ongoing research.

Acknowledgements

FWC was supported by NIH grants NICHD DP2 OD022614, NCATS KL2 TR000140, and NIMH P30 MH062294, the Yale Center for Clinical Investigation, and the Yale Center for Interdisciplinary Research on AIDS. DAK was partially supported by the Canadian Institutes of Health Research. NAC was supported by NIH grants P01 AG031093 and P30 AG034420, and the Bill & Melinda Gates Foundation. We are grateful to Liza Nicoll for help accessing and formatting the data from the Kim *et al.* [42] study.

Appendix

A. Proof of Proposition 1

Consider the competing risk of transmission for a susceptible j from all prior adopters i connected to j. Let T_{ij} be the edge-wise waiting time for i to transmit to j, let $f_{ij}(t-t_i)$ be the density function, let $F_{ij}(t-t_i)$ be the cumulative distribution function, and let $S_{ij}(t-t_i) = 1 - F_{ij}(t-t_i)$ be the survival function. For simplicity, we abbreviate the conditional distribution of T_{ij} given covariates X_i, X_j, Z_{ij}. The random adoption time of j is

$$T_j = \min_{i \in N_j} t_i + T_{ij} \tag{15}$$

The survival function of T_j is given by

$$\begin{aligned} S_j(t) &= \Pr(T_j > t) \\ &= \Pr(T_{ij} + t_i > t, \forall i \in N_j, t_i < t) \\ &= \prod_{i \in N_j, t > t_i} \Pr(T_{ij} + t_i > t) \\ &= \exp\left[- \sum_{i \in N_j, t > t_i} \Lambda_{ij}(t - t_i) \right] \end{aligned} \tag{16}$$

where the third line follows by conditional independence of the T_{ij}'s for all prior adopters i connected to j given X_i, X_j and Z_{ij}. The hazard function of T_j is given by

$$\begin{aligned} \lambda_j(t) &= \frac{f_j(t)}{S_j(t)} \\ &= -\frac{\frac{\mathrm{d}S_j(t)}{\mathrm{d}t}}{S_j(t)} \\ &= -\frac{\exp\left[- \sum_{i \in N_j, t > t_i} \Lambda_{ij}(t - t_i) \right]}{\exp\left[- \sum_{i \in N_j, t > t_i} \Lambda_{ij}(t - t_i) \right]} \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i \in N_j, t > t_i} -\Lambda_{ij}(t - t_i) \\ &= \sum_{i \in N_j, t > t_i} \lambda_{ij}(t - t_i) \end{aligned} \tag{17}$$

as claimed.

B. Logistic model

We compare the fit of the edgewise diffusion model with Valente's model [37, page 106]. Let Y_j be the indicator of adoption before the end of the study and let $P_j = \Pr(Y_j = 1)$. Valente's model has the logistic regression form

$$\log\left(\frac{P_j}{1-P_j}\right) = \alpha + \beta X_j + \gamma E_j \tag{18}$$

where $E_j = \frac{1}{d_j}\sum_{i\in N_j} Y_i$ is the proportion of network friends who are prior adopters before j's adoption or at the end of the study. After estimating $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$, we predict the adoption probabilities by

$$\hat{p}_j^{\text{logistic}} = \frac{\exp[\hat{\alpha} + \hat{\beta} X_j + \hat{\gamma} E_j]}{1 + \exp[\hat{\alpha} + \hat{\beta} X_j + \hat{\gamma} E_j]} \tag{19}$$

For the edgewise Cox model $\lambda_{ij}(\tau) = \lambda_0(\tau)\exp(\alpha' X_i + \beta' X_j + \eta' Z_{ij})$, we compute the estimated adoption probabilities as follows. The individual hazard of adoption is

$$\hat{\lambda}_j(t) = \sum_{i\in N_j, t>t_i} \hat{\lambda}_{ij}(t - t_i). \tag{20}$$

The cumulative hazard of individual hazard is the sum of edgewise cumulative hazards,

$$\hat{\Lambda}_j(t) = \sum_{i\in N_j, t>t_i} \hat{\Lambda}_{ij}(t - t_i). \tag{21}$$

We predict the individual adoption probability at the end of the study T by one minus the survival probability,

$$\hat{p}_j^{\text{edgewise}} = 1 - \exp\left[-\hat{\Lambda}_j(T)\right]. \tag{22}$$

To compare Valente's logistic model with the edgewise Cox model, we treat the adoption status before the end of the study as a Bernoulli trial with probability p_j and compute the binomial log likelihood for both models:

$$l(\mathbf{y}|\mathbf{p}) = \sum_{j=1}^{n} y_j \log(\hat{p}_j) + (1 - y_j)\log(1 - \hat{p}_j).$$

By putting these models into the same binomial family, we can compare models using AIC $= -2l + 2k$ where k is the number of parameters [66]. The AIC for logistic model is 3578.731, while the AIC for edgewise Cox model is 3398.262. We conclude that the edgewise model fits the data better.

C. Number or proportion of adopting neighbors?

To study the dependence of adoption times on the absolute number, or proportion of adopting neighbors, we define an alternative model,

$$\lambda_j^*(t) = \frac{1}{d_j}\sum_{i\in N_j, t>t_i} \lambda_{ij}(t - t_i) \tag{23}$$

where d_j is the network degree of j. Fitting this model amounts to adding an offset of $-\log(d_j)$ in the edgewise hazard regression model. We compare the log likelihood of exponential hazard regression, Weibull hazard regression and Cox proportional hazard model with and without dividing the hazard by the susceptible's network degree. These models have the same degree of freedom. Denote the log likelihood of the original model as l and the log likelihood of the model divided by the susceptible's network degree as l^*. Table 3 shows the $l - l^*$ for three models and two interventions, and the original model fits the data better than the model divided by the network degree.

[Table 3 about here.]

We plot the Cox-Snell residuals versus the estimated cumulative hazard of the residuals for Exponential, Weibull, Cox proportional hazard regression without and with dividing by network degrees in Figure 6 and 7. These results suggest that the Cox proportional hazards model fits better than Exponential and Weibull models, and the model that does not divide total hazard by network degree fits the data better than the model that divides by network degree.

[Figure 6 about here.]

[Figure 7 about here.]

D. Random effects/frailty terms

We can incorporate frailty terms to represent shared dependence of edge-wise waiting times on the prior adopter i,

$$\lambda_{ij}(\tau) = \lambda_0(\tau)\exp(\alpha' X_i + \beta' X_j + \eta' Z_{ij} + \theta_i),$$

where θ_i is an adopter-specific random effect/frailty term. The distribution of θ_i is assumed to be Gaussian. Likelihood ratio tests based on intergrated and penalized likelihoods both reject the null hypothesis that random effects are zero. The AIC of random effects model based on the intergrated log partial likelihood is 25470.87 while AIC of Cox's model is 25725.28 for multivitamin. AIC of random effects model is 21857.99 while AIC of Cox's model is 21846.95 for chlorine. The adopter-specific random effects does not improve the model fit than Cox's model. We show the distribution of the estimated random effects in Figure 8 with standard deviation 0.4647 for multivitamin and 0.5275 for chlorine. The estimated adopter-specific random effects are approximately normally distributed. Table 4 shows fixed effect coefficients.

[Figure 8 about here.]

[Table 4 about here.]

Another type of random effects/frailty terms are for susceptible individuals

$$\lambda_{ij}(\tau) = \lambda_0(\tau)\exp(\alpha' X_i + \beta' X_j + \eta' Z_{ij} + \theta_j),$$

where θ_j is a susceptible-specific random effects/frailty term. This model permits a susceptible individual j with a large negative value for θ_j to be very unlikely to adopt, regardless of their exposure. Likelihood ratio tests based on intergrated and penalized likelihood both reject the null that random effects are zero. AIC of random effects model based on the intergrated log partial likelihood is 25728.53 while AIC of Cox's model is 25725.28 for multivitamin. AIC of random effects model is 21876.31 while AIC of Cox's model is 21846.95 for chlorine. The susceptible-specific random effects does not improve the model fit than Cox's model.

The standard deviation of random effects are 0.8144 for multivitamin and 0.7983 for chlorine. Figure 9 shows the distribution of random effect. The estimated susceptible-specific random effects have two modes and do not look similar to normal distribution. Table 5 shows the fixed effect coefficients.

[Figure 9 about here.]

[Table 5 about here.]

E. Additive hazard model

Aalen's additive hazard model [52] allows estimated components of the hazard to be negative,

$$\lambda_{ij}(\tau) = \lambda_0(\tau) + \alpha(\tau)' X_i + \beta(\tau)' X_j + \eta(\tau)' Z_{ij}$$

where $\lambda(\tau)$ is the baseline hazard and the coefficients $\alpha(\tau)$, $\beta(\tau)$ and $\eta(\tau)$ are time-varying. Figure 10 and 11 show estimates of the cumulative coefficients and their 95% pointwise confidence intervals for the multivitamin and chlorine interventions. Figure 12 shows the Cox-Snell residuals for Aalen's additive hazard model. We find five edges $\{i, j\}$ with have negative cumulative hazard (at the moment of adoption or censoring) for the multivitamin intervention, and two such edges for the chlorine intervention. Table 6 and 7 show the village- and adopter-level covariates for these edges. Comparison of these residuals with those of the Cox model in Figure 6 and 7 shows slightly smaller residuals in the additive model.

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Table 6 about here.]

[Table 7 about here.]

F. Semi-parametric proportional hazards mixture cure model

We fit a semi-parametric proportional hazards mixture cure model [67] in the edgewise diffusion framework. Let $1 - \pi(Z)$ be the probability of an edge being "cured" (no ticket being passed along that edge) and let $S(t|X)$ be the survival probability of "uncured" edges, and X and Z are covariates that may affect survival and cure probability. The mixture cure model can be expressed as

$$S_{mix}(t|X, Z) = \pi(Z)S(t|X) + 1 - \pi(Z),$$

where $S(t|X)$ is estimated by survival regression such as the Cox proportional hazard model, and $\pi(Z)$ can be estimated by logistic regression. Table 8 shows logistic regression coefficients for the cure probability model, and Table 9 shows the Cox regression coefficients for the edge-wise diffusion model. We predict the individual probability of adoption at the end based the cumulative cure probability,

$$\hat{p}_j^{\text{cure}} = 1 - \prod_{i \in N_j, T > t_i} \hat{S}_{ij}(T - t_i)$$

where $\hat{S}_{ij}$ is the predicted edgewise survival from semi-parametric cure model. We calculate the binomial log likelihood and the AIC (based on the binomial log likelihood) of this model as 3177.428, smaller than that of the logisitc model and the edgewise Cox model, suggesting that the cure model fits the data better.

[Table 8 about here.]

[Table 9 about here.]

G. Exponential and Weibull model

Tables 10 and 11 show regression coefficients, hazard ratio and 95% confidence interval from the Exponential and Weibull hazard models. The coefficients of Exponential and Weibull regression have the same sign as Cox regression, and their p-values have the same significance level as Cox's regression despite some slight differences, suggesting that Cox's model agrees with the parametric models.

[Table 10 about here.]

[Table 11 about here.]

H. Village-level fixed effects

Tables 12 and 13 show the village-level fixed effects for the adoption of multivitamin and chlorine respectively after controlling for prior adopter's attributes. Village 1 was treated as base group.

[Table 12 about here.]

[Table 13 about here.]

References

1. Valente TW. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory* 1996; **2**(2):163–164.

2. Rogers EM. *Diffusion of innovations*. Simon & Schuster, 2010.

3. Ryan B, Gross NC. The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* 1943; **8**(1):15–24.

4. Coleman JS, Katz E, Menzel H, *et al.*. *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Company Indianapolis, 1966.

5. Burt RS. Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology* 1987; :1287–1335.

6. Van den Bulte C, Lilien GL. Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology* 2001; **106**(5):1409–1435.

7. Friedkin NE. A multilevel event history model of social diffusion: Medical innovation revisited. *Journal of Mathematical Sociology* 2010; **34**(2):146–155.

8. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO. The diffusion of microfinance. *Science* 2013; **341**(6144).

9. Valente TW. Network interventions. *Science* 2012; **337**(6090):49–53.

10. Valente TW, Ritt-Olson A, Stacy A, Unger JB, Okamoto J, Sussman S. Peer acceleration: effects of a social network tailored substance abuse prevention program among high-risk adolescents. *Addiction* 2007; **102**(11):1804–1815.

11. Centola D. The spread of behavior in an online social network experiment. *Science* 2010; **329**(5996):1194–1197.

12. Park J, Chung K, Han D, Lee S. Mothers clubs and family planning in Korea. *Seoul Korea Seoul National University School of Public Health May 1974. 312* 1974; .

13. Rogers EM, Kincaid DL. *Communication networks: Toward a new paradigm for research.* Free Press, 1981.

14. Valente TW, Watkins SC, Jato MN, Van Der Straten A, Tsitsol LPM. Social network associations with contraceptive use among Cameroonian women in voluntary associations. *Social Science & Medicine* 1997; **45**(5):677–687.

15. Wellin E. Water boiling in a Peruvian town. *Health, Culture and Community* 1955; :71–103.

16. Gruhl D, Guha R, Liben-Nowell D, Tomkins A. Information diffusion through blogspace. *Proceedings of the 13th international conference on World Wide Web*, ACM, 2004; 491–501.

17. Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences* 2008; **105**(12):4633–4638.

18. Cha M, Mislove A, Gummadi KP. A measurement-driven analysis of information propagation in the Flickr social network. *Proceedings of the 18th international conference on World wide web*, ACM, 2009; 721–730.

19. Cha M, Haddadi H, Benevenuto F, Gummadi PK. Measuring user influence in Twitter: The million follower fallacy. *ICWSM* 2010; **10**:10–17.

20. Bakshy E, Hofman JM, Mason WA, Watts DJ. Everyone's an influencer: quantifying influence on Twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011; 65–74.

21. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. The dynamics of protest recruitment through an online network. *Scientific reports* 2011; **1**.

22. Lerman K, Ghosh R, Surachawala T. Social contagion: An empirical study of information spread on Digg and Twitter follower graphs. *arXiv preprint arXiv:1202.3162* 2012; .

23. Bass FM. A new product growth for model consumer durables. *Management Science* 1969; **15**(5):215227.

24. Borge-Holthoefer J, Baños RA, González-Bailón S, Moreno Y. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks* 2013; **1**(1):3–24.

25. Valente TW. Social network thresholds in the diffusion of innovations. *Social Networks* 1996; **18**(1):69–89.

26. Guardiola X, Diaz-Guilera A, Perez CJ, Arenas A, Llas M. Modeling diffusion of innovations in a social network. *Physical Review E* 2002; **66**(2):026 121.

27. Wang P, González MC, Hidalgo CA, Barabási AL. Understanding the spreading patterns of mobile phone viruses. *Science* 2009; **324**(5930):1071–1076.

28. Granovetter M. Threshold models of collective behavior. *American Journal of Sociology* 1978; :1420–1443.

29. Granovetter M, Soong R. Threshold models of diffusion and collective behavior. *Journal of Mathematical Sociology* 1983; **9**(3):165–179.

30. Granovetter M, Soong R. Threshold models of interpersonal effects in consumer demand. *Journal of Economic Behavior & organization* 1986; **7**(1):83–99.

31. Marsden PV, Podolny J. Dynamic analysis of network diffusion processes. *Social Networks Through Time*, Weesie J, Flap H (eds.). Utrecht: ISOR/Rijksuniversiteit Utrecht, 1990; 197–214.

32. Strang D. From dependency to sovereignty: An event history analysis of decolonization 1870-1987. *American Sociological Review* 1990; :846–860.

33. Strang D. Adding social structure to diffusion models an event history framework. *Sociological Methods & Research* 1991; **19**(3):324–353.

34. Strang D, Tuma NB. Spatial and temporal heterogeneity in diffusion. *American Journal of Sociology* 1993; :614–639.

35. Greve HR, Strang D, Tuma NB. Specification and estimation of heterogeneous diffusion models. *Sociological Methodology* 1995; **25**:377–420.

36. Valente TW, Dyal SR, Chu KH, Wipfli H, Fujimoto K. Diffusion of innovations theory applied to global tobacco control treaty ratification. *Social Science & Medicine* 2015; **145**:89–97.

37. Valente TW. Network models and methods for studying the diffusion of innovations. *Models and Methods in Social Network Analysis*, Carrington PJ, Scott J, Wasserman S (eds.). Cambridge University Press New York, NY, 2005; 98–116.

38. Allison PD. Discrete-time methods for the analysis of event histories. *Sociological Methodology* 1982; **13**(1):61–98.

39. Myers DJ. The diffusion of collective violence: Infectiousness, susceptibility, and mass media networks. *American Journal of Sociology* 2000; **106**(1):173–208.

40. Greenan CC. Diffusion of innovations in dynamic networks. *Journal of the Royal Statistical Society: Series A* 2015; **178**(1):147–166.

41. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH. A 61-million-person experiment in social influence and political mobilization. *Nature* 2012; **489**(7415):295–298.

42. Kim DA, Hwong AR, Stafford D, Hughes DA, O'Malley AJ, Fowler JH, Christakis NA. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* 2015; **386**(9989):145–153.

43. Coviello L, Sohn Y, Kramer AD, Marlow C, Franceschetti M, Christakis NA, Fowler JH. Detecting emotional contagion in massive social networks. *PloS One* 2014; **9**(3):e90 315.

44. Hill AL, Rand DG, Nowak MA, Christakis NA. Emotions as infectious diseases in a large social network: the sisa model. *Proceedings of the Royal Society of London B: Biological Sciences* 2010; **277**(1701):3827–3835.

45. VanderWeele TJ, An W. Social networks and causal inference. *Handbook of Causal Analysis for Social Research*. Springer, 2013; 353–374.

46. Crawford FW. The graphical structure of respondent-driven sampling. *Sociological Methodology* 2016; **46**(1):187–211.

47. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* 1972; **34**:187–220.

48. Klein JP, Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

49. Feld SL. Why your friends have more friends than you do. *American Journal of Sociology* 1991; :1464–1477.

50. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**(282):457–481.

51. Akaike H. Likelihood of a model and information criteria. *Journal of econometrics* 1981; **16**(1):3–14.

52. Aalen OO. A linear regression model for the analysis of life times. *Statistics in medicine* 1989; **8**(8):907–925.

53. Centola D. An experimental study of homophily in the adoption of health behavior. *Science* 2011; **334**(6060):1269–1272.

54. Aral S, Walker D. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* 2011; **57**(9):1623–1639.

55. Rand DG, Arbesman S, Christakis NA. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences* 2011; **108**(48):19 193–19 198.

56. Onnela JP, Christakis NA. Spreading paths in partially observed social networks. *Physical Review E* 2012; **85**(3):036 106.

57. Hudgens MG, Halloran ME. Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association* 2006; **101**(473):51–64.

58. Rosenbaum PR. Interference between units in randomized experiments. *Journal of the American Statistical Association* 2012; .

59. Hudgens MG, Halloran ME. Toward causal inference with interference. *Journal of the American Statistical Association* 2012; .

60. Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine* 2013; **32**(4):556–577.

61. Staples PC, Ogburn EL, Onnela JP. Incorporating contact network structure in cluster randomized trials. *Scientific Reports* 2015; **5**.

62. Toulis P, Kao EK. Estimation of causal peer influence effects. *ICML (3)*, 2013; 1489–1497.

63. Aronow PM, Samii C. Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156* 2013; .

64. Ogburn EL, VanderWeele TJ, *et al.*. Causal diagrams for interference. *Statistical Science* 2014; **29**(4):559–578.

65. Fu F, Fowler JH, Christakis NA. Overlying biological and social contagions 2017; :To Appear.

66. Akaike H. Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike*. Springer, 1998; 199–213.

67. Cai C, Zou Y, Peng Y, Zhang J. smcure: An r-package for estimating semiparametric mixture cure models. *Computer methods and programs in biomedicine* 2012; **108**(3):1255–1260.

Statistics
in Medicine

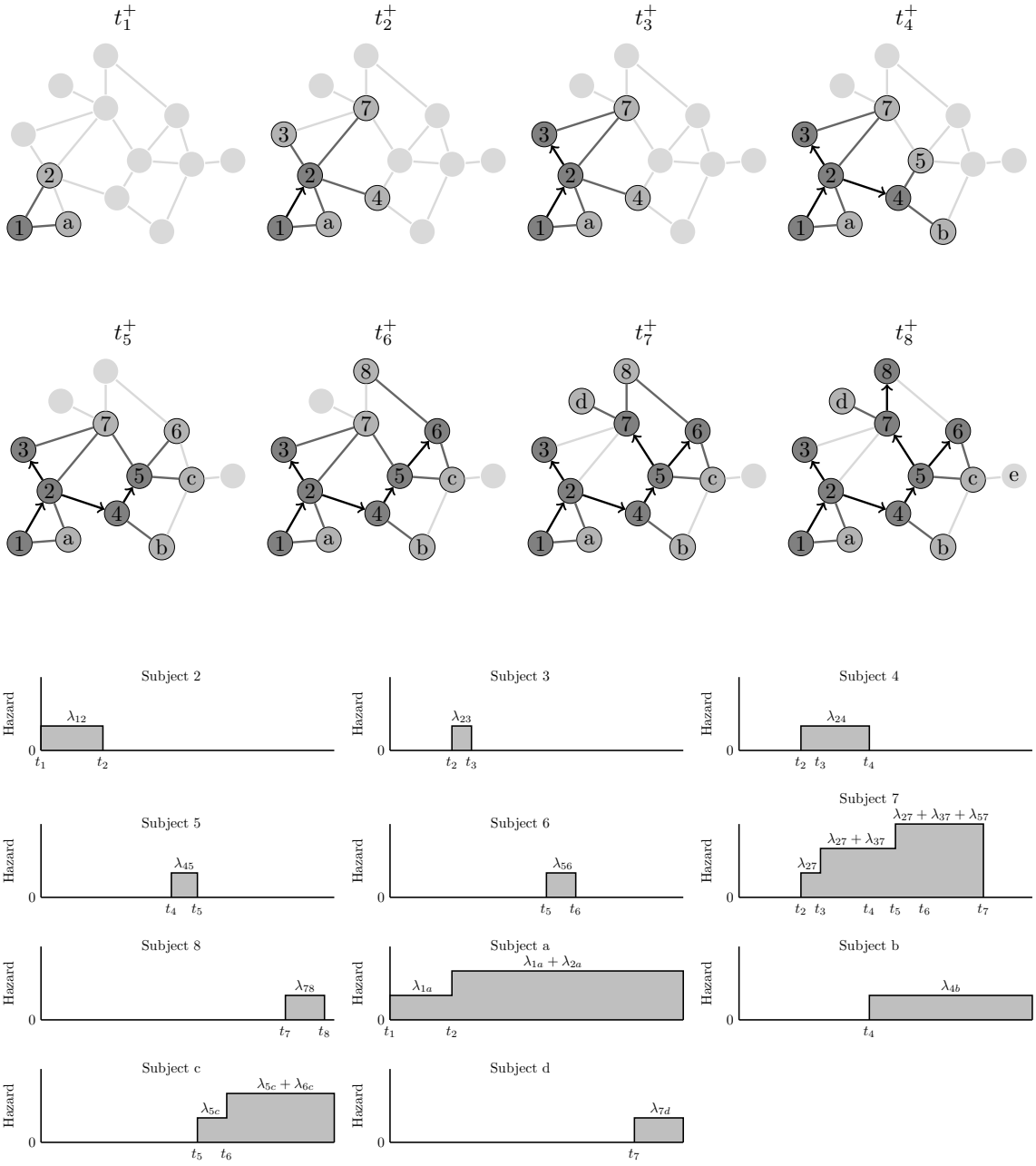

Figure 1. How network exposure works in a diffusion process. The first two rows show the evolution of an adoption process on an example network, starting with a seed labeled 1. The numbered circles denote the order of adoption and arrows represent transmission of the innovation. The time just after ith adoption is denoted as t_i^+. Light gray lines and circles are susceptible edges and individuals at the moment of each adoption. The last four lines show how the total hazard/exposure of adoption felt by susceptible individuals changes over time, assuming constant edge-wise hazards. The exposure increases one step whenever the number of prior adopters connected to the individual increases. The shaded area under each subject's curve is the cumulative exposure experienced by that subject.

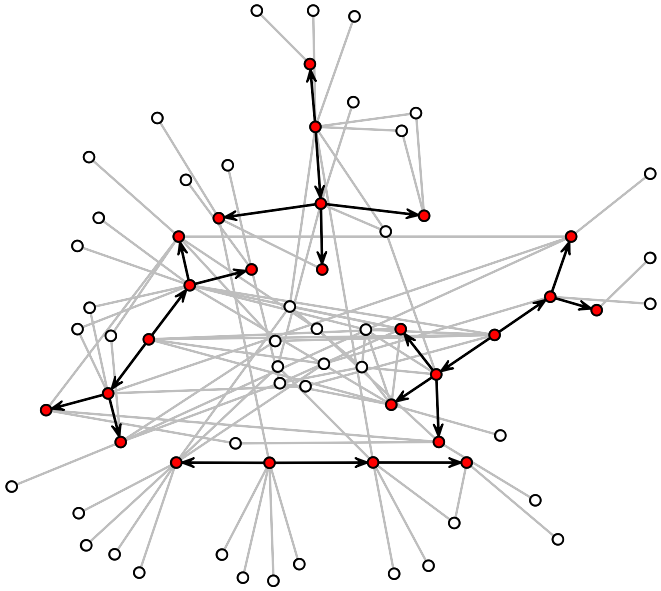

Figure 2. Diffusion of multivitamin adoption in the social network of Village 4. Social network edges, measured before the diffusion study began, are shown in gray. Red circles represent multivitamin adopters, and white circles are susceptible subjects who did not adopt. Arrows represent transmission (and redemption) of multivitamin tickets.

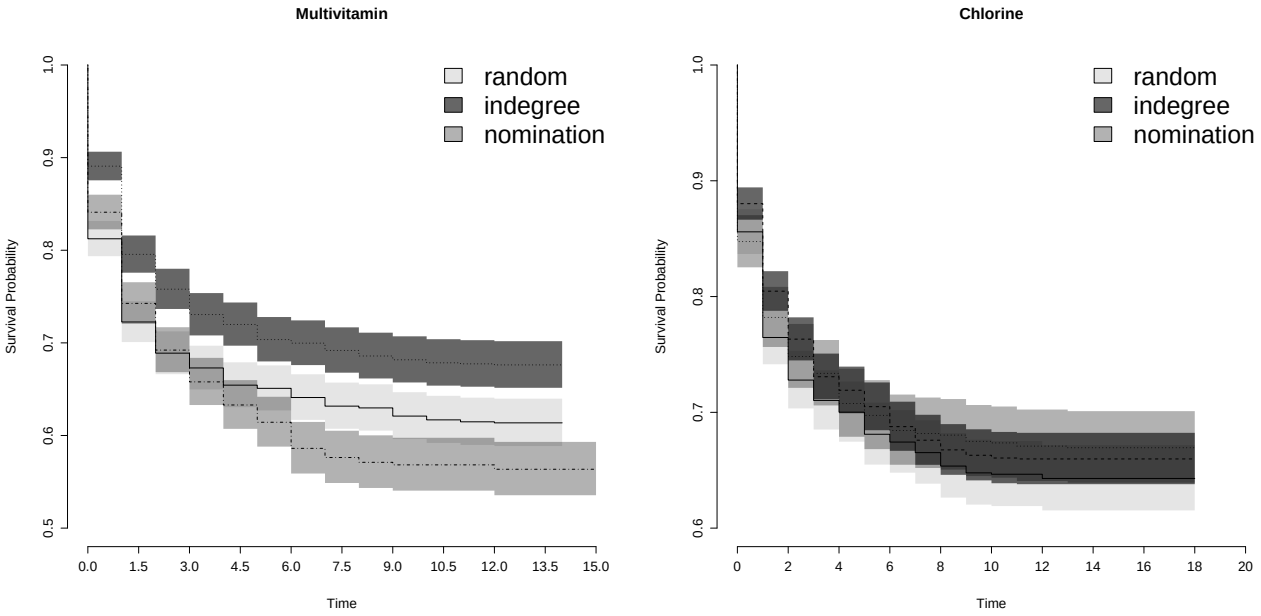

Figure 3. Comparison of Kaplan-Meier curves for edge-wise diffusion among three targeting methods for the diffusion of multivitamin and chlorine interventions. Lower curves indicate faster adoption across network edges. Semi-transparent areas are 95% pointwise confidence intervals for each unadjusted curve.

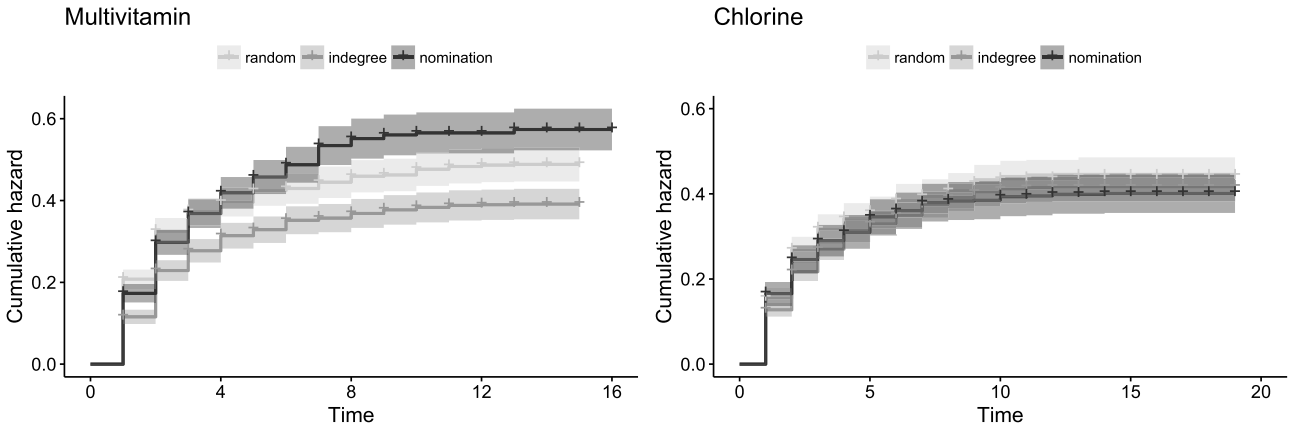

Figure 4. Cumulative edge-wise hazards for adoption of multivitamins and chlorine, across all villages. The first two days after exposure to prior adopters saw the highest rates of adoption, followed by much slower rates of adoption thereafter. The multivitamin intervention had a higher overall diffusion rate than the chlorine intervention. Shaded areas indicate 95% pointwise confidence intervals.

Prepared using simauth.cls

Statistics
in Medicine

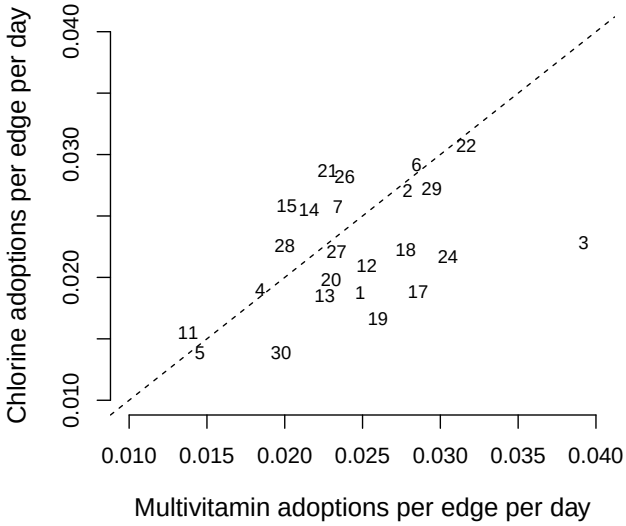

Figure 5. Comparison of edge-wise diffusion rates for the multivitamin and chlorine interventions in 24 villages. The diffusion rate is defined as the expected number of transmissions per edge per day, and the average expected number of transmissions from Cox model with village fixed effects were calculated and plotted. Note that the time is adjusted to the same scale for the two interventions so that they are comparable. The horizontal axis shows the diffusion rate of the multivitamin intervention while the vertical axis corresponds to the chlorine intervention. Diffusion rates were heterogenous among villages. Villages shown above the diagonal exhibit faster chlorine diffusion than multivitamin diffusion, while villages shown below the diagonal have faster multivitamin diffusion than chlorine.

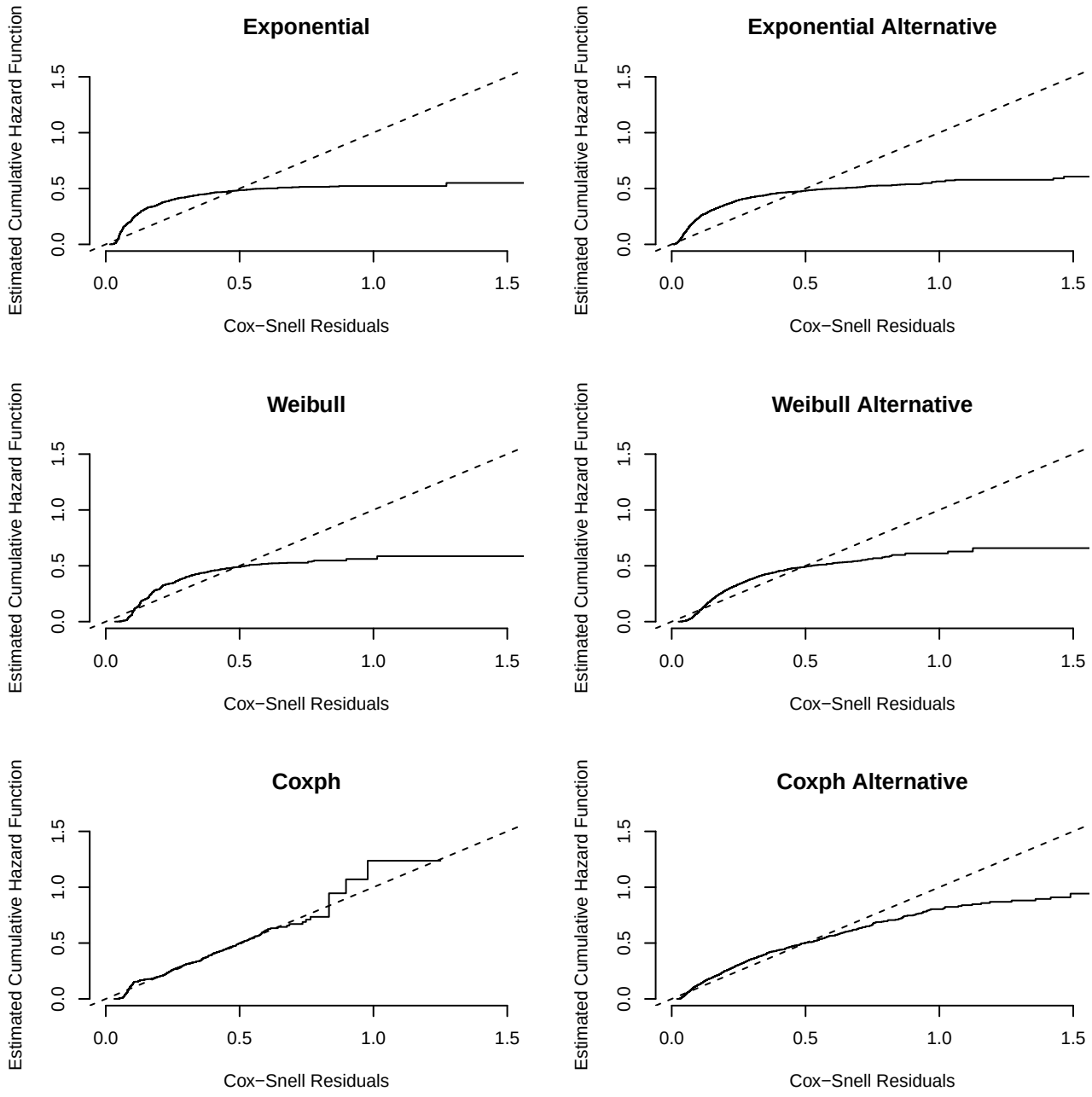

Figure 6. Cox-Snell residuals and estimated cumulative hazard of residuals for the multivitamin intervention. The dashed line is the expected relationship under correct specification of the edge-wise hazard model. The left panel shows the edgewise diffusion models, and the right panel shows the alternative models (Equation 22) that divide the edgewise hazard by the susceptible individuals' network degree.

Prepared using simauth.cls

Statistics
in Medicine

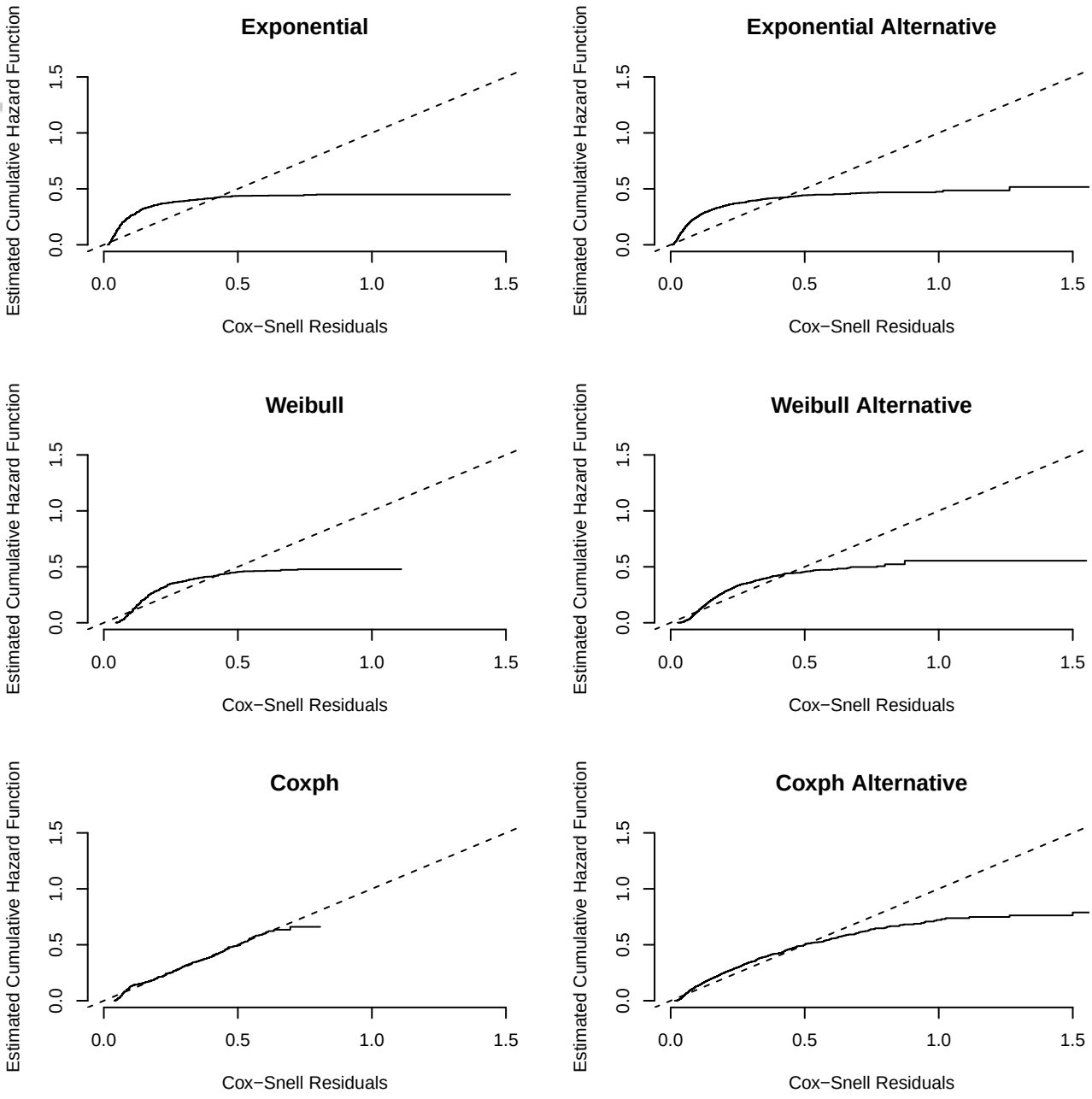

Figure 7. Cox-Snell residuals and estimated cumulative hazard of residuals for the chlorine intervention. The dashed line represents the expected relationship under correct specification of the edge-wise hazard model. The left panel shows the edgewise diffusion models, and the right panel shows the alternative models (Equation 22) that divide the edgewise hazard by the susceptible individuals network degree.

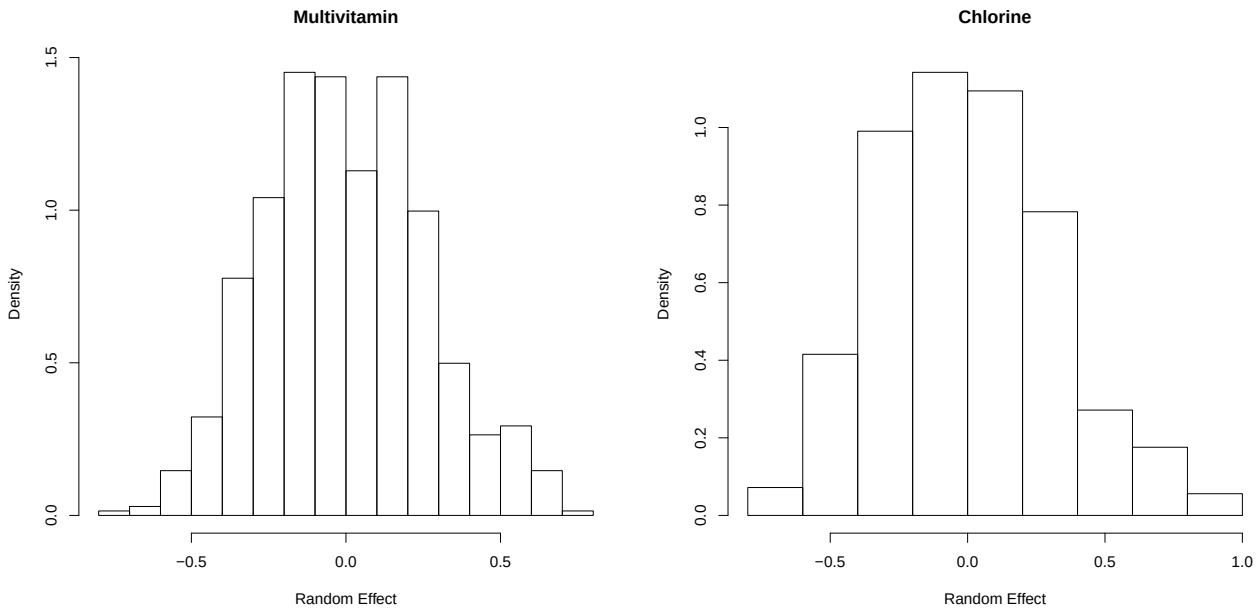

Figure 8. Distribution of adopter-specific random effects

Statistics
in Medicine

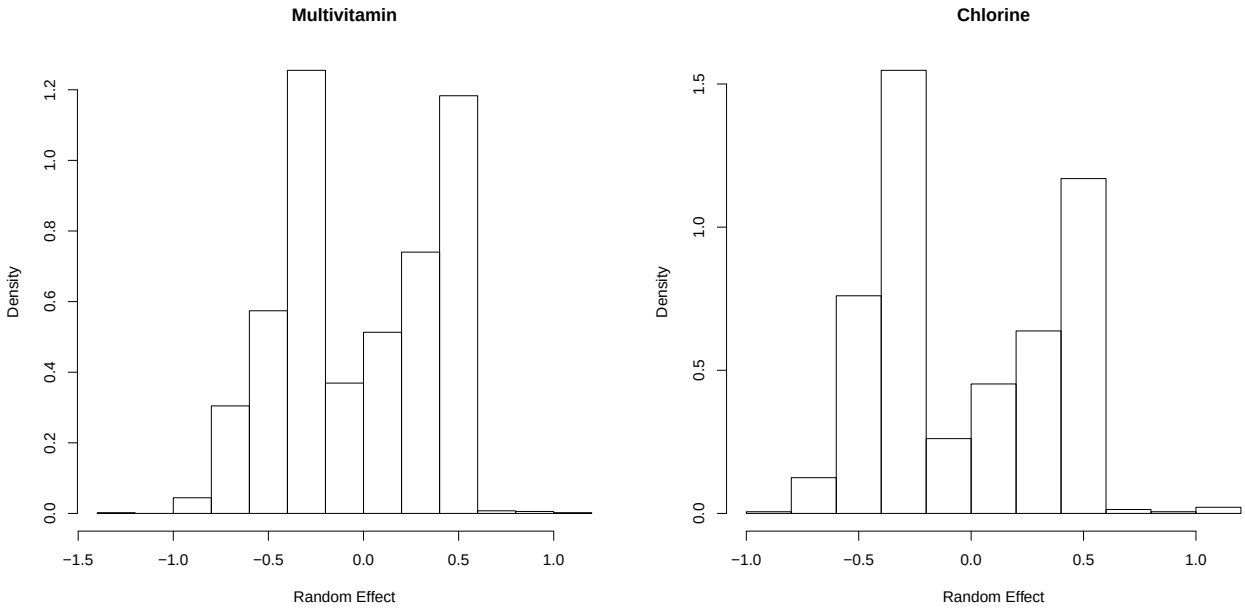

Figure 9. Distribution of susceptible-specific random effects

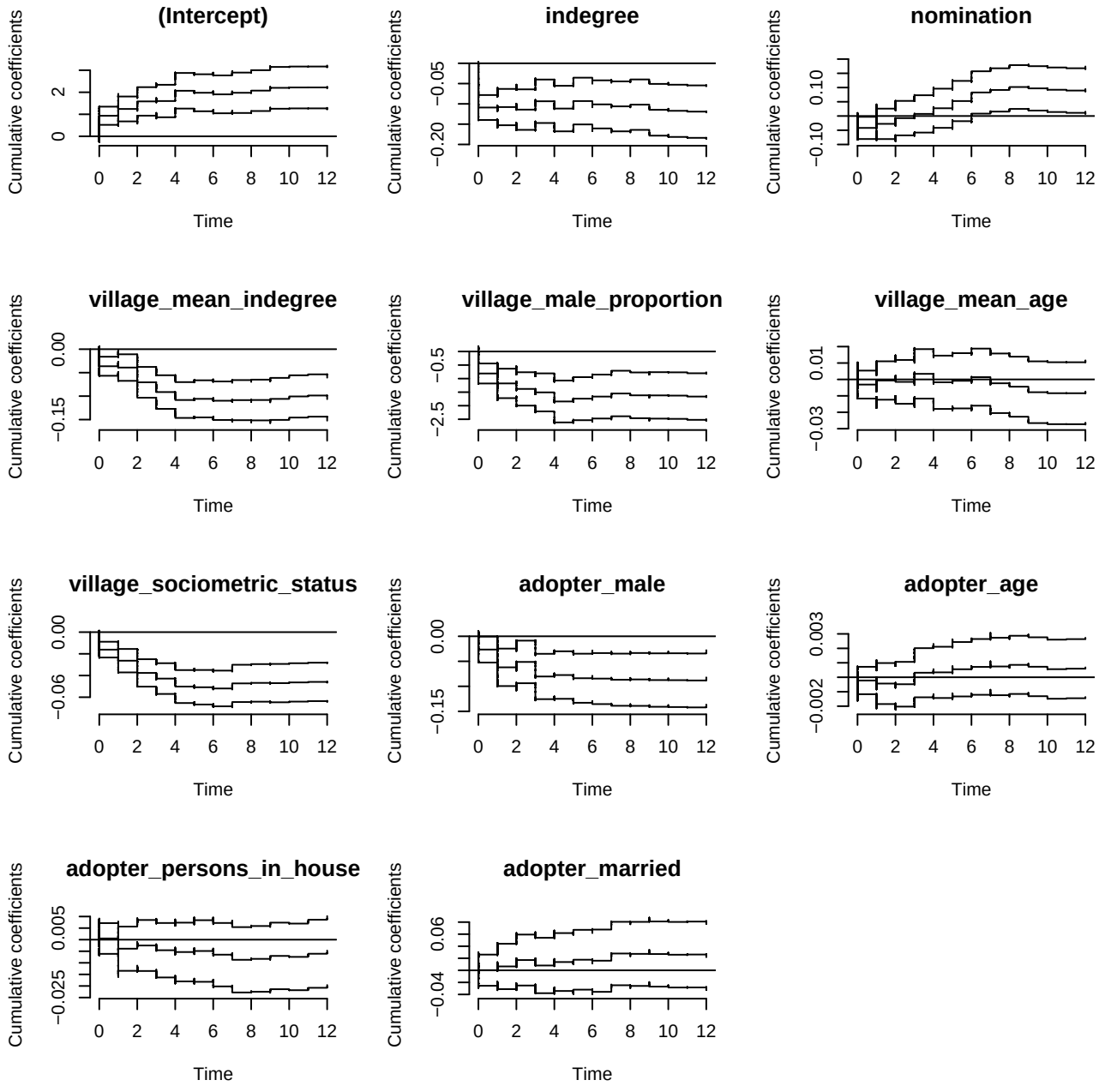

Figure 10. Aalen's additive hazard regression for diffusion of multivitamin adoption

Prepared using simauth.cls

Statistics
in Medicine

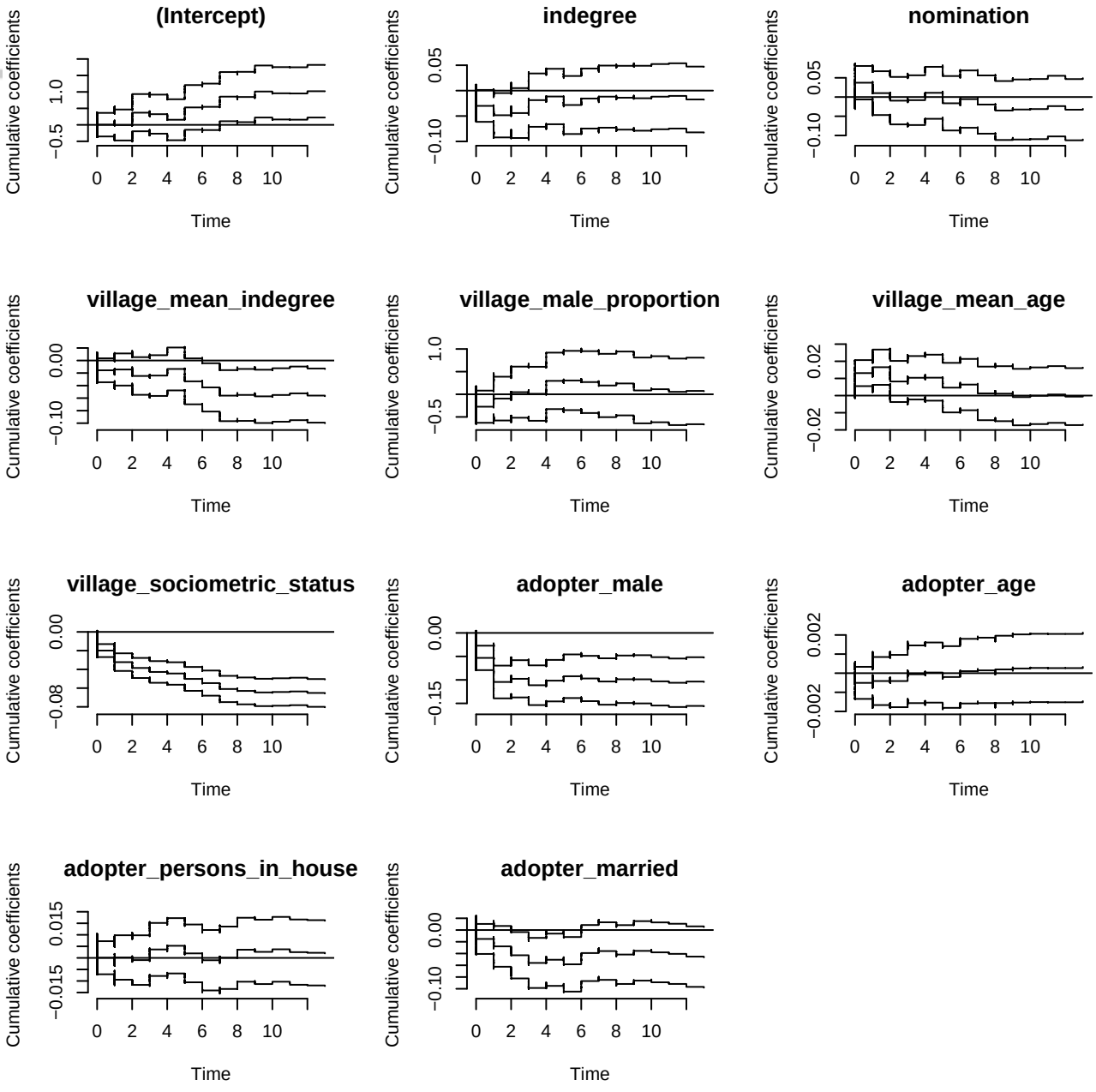

Figure 11. Aalen's additive hazard regression for diffusion of chlorine adoption

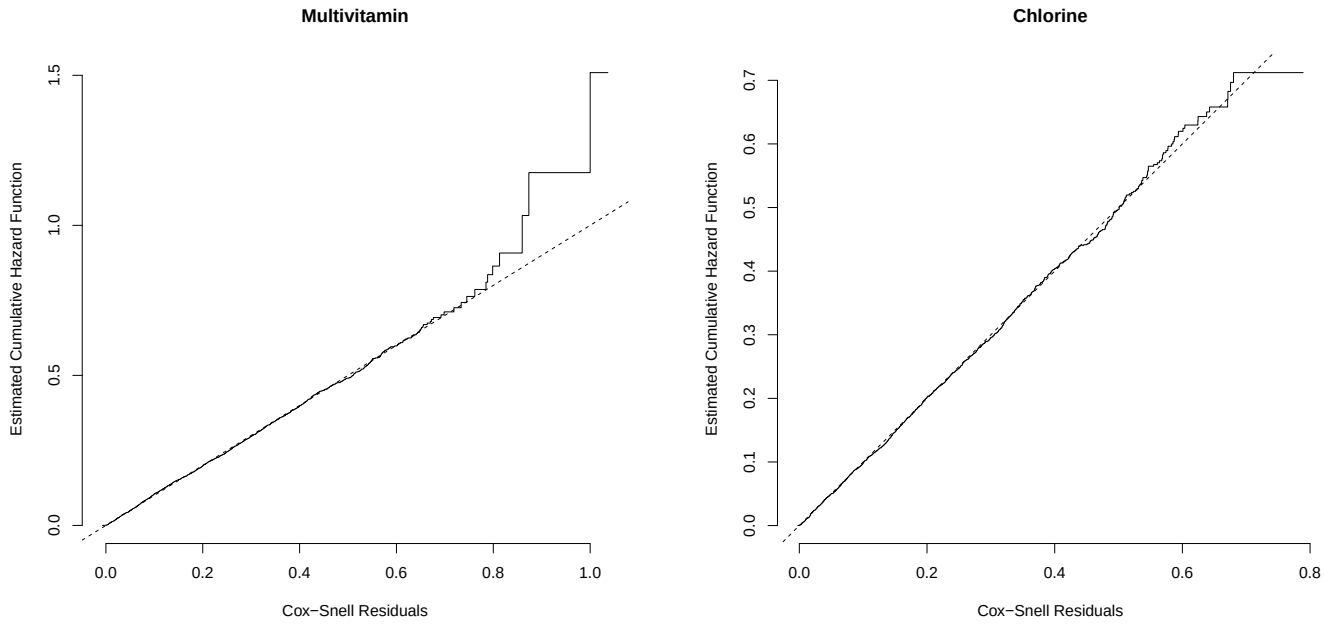

Figure 12. Aalen's additive hazard model fit

	Multivitamin				Chlorine			
	Coef	HR	95%CI(HR)	p	Coef	HR	95%CI(HR)	p
Indegree targeting	-0.310	0.733	(0.644, 0.835)	< 0.01	-0.093	0.911	(0.795, 1.044)	0.18
Nomination targeting	0.045	1.046	(0.916, 1.194)	0.50	-0.015	0.985	(0.834, 1.164)	0.86
Village mean indegree	-0.191	0.826	(0.763, 0.895)	< 0.01	-0.102	0.903	(0.826, 0.988)	0.03
Village male proportion	-3.300	0.037	(0.009, 0.156)	< 0.01	-0.231	0.794	(0.176, 3.588)	0.76
Village mean age	-0.008	0.992	(0.959, 1.026)	0.65	0.022	1.022	(0.988, 1.058)	0.21
Village SES	-0.083	0.921	(0.894, 0.948)	< 0.01	-0.125	0.883	(0.858, 0.909)	< 0.01
Adopter male	-0.198	0.820	(0.736, 0.915)	< 0.01	-0.265	0.767	(0.681, 0.864)	< 0.01
Adopter age	0.001	1.001	(0.997, 1.004)	0.59	-0.000	0.999	(0.996, 1.004)	0.90
Adopter persons in house	-0.012	0.988	(0.961, 1.015)	0.37	0.002	1.002	(0.971, 1.035)	0.88
Adopter married	0.021	1.021	(0.922, 1.130)	0.69	-0.100	0.904	(0.810, 1.010)	0.07

Table 1. Cox semi-parametric regression coefficients for the adoption of multivitamins and chlorine. The first six covariates are village-level characteristics and the last four covariates are characteristics of prior adopters.

	Multivitamin				Chlorine			
	Coef	OR	95%CI(OR)	p	Coef	OR	95%CI(OR)	p
Indegree targeting	-0.22	0.80	(0.70, 0.92)	< 0.01	-0.17	0.85	(0.73, 0.99)	0.03
Nomination targeting	0.12	1.13	(0.98, 1.30)	0.09	0.26	1.30	(1.10, 1.54)	< 0.01
Village mean indegree	-0.07	0.93	(0.86, 1.01)	0.09	0.07	1.07	(0.97, 1.19)	0.18
Village male proportion	-4.31	0.01	(0.00, 0.06)	< 0.01	-0.78	0.46	(0.08, 2.52)	0.37
Village mean age	-0.01	0.99	(0.95, 1.02)	0.51	-0.04	0.96	(0.92, 0.10)	0.05
Village SES	-0.09	0.91	(0.88, 0.94)	< 0.01	-0.13	0.88	(0.85, 0.91)	< 0.01
Exposure	0.29	1.33	(1.02, 1.74)	0.04	0.26	1.29	(0.95, 1.76)	0.11

Table 2. Logistic regression coefficients for adoption of multivitamins and chlorine. Exposure is defined as the proportion of network neighbors who are prior adopters.

Statistics
in Medicine

	Multivitamin	Chlorine
Exponential	156.49	94.98
Weibull	84.46	37.84
Cox	184.84	201.77

Table 3. Difference in log likelihood between baseline models and alternative models dividing the edgewise hazard by the susceptible individuals network degree. The baseline models have higher log likelihood.

Prepared using simauth.cls

	Multivitamin			Chlorine		
	Coef	Std Error	p	Coef	Std Error	p
Adopter male	-0.269	0.068	8e-05	-0.185	0.078	0.017
Adopter age	0.002	0.002	0.35	0.002	0.003	0.48
Adopter persons in house	-0.020	0.017	0.24	-0.001	0.020	0.62
Adopter married	-0.008	0.065	0.9	-0.121	0.073	0.099

Table 4. Regression coefficients of adopter-specific random effects

Statistics
in Medicine

	Multivitamin			Chlorine		
	Coef	Std Error	p	Coef	Std Error	p
Adopter male	-0.257	0.060	2.2e-5	-0.216	0.065	9e-4
Adopter age	0.002	0.002	0.33	0.002	0.002	0.35
Adopter persons in house	-0.018	0.015	0.22	-0.008	0.017	0.63
Adopter married	0.013	0.057	0.82	-0.135	0.061	0.029

Table 5. Regression coefficients of susceptible-specific random effects.

	Village				Adopter			
Intervention	prop mean indegree	prop male	mean age	SES	male	age	person in house	married
indegree	3.19	0.53	36.4	3.3	yes	36	5.3	no
indegree	4.23	0.51	36.1	8.0	yes	39	8.0	yes
indegree	4.23	0.51	36.1	8.0	yes	35	7.0	yes
indegree	4.23	0.51	36.1	8.0	yes	35	7.0	yes
indegree	4.23	0.51	36.1	8.0	yes	35	7.0	yes

Table 6. Edges that have negative cumulative hazard for multivitamin diffusion. Each row corresponds to an edge $\{i, j\}$ linking a prior adopter i to a susceptible subject j.

Statistics
in Medicine

Village					Adopter			
Intervention	prop mean indegree	prop male	mean age	SES	male	age	person in house	married
nomination	2.60	0.49	30.8	7.4	yes	46	4	yes
indegree	2.26	0.44	36.1	8.4	yes	43	6	yes

Table 7. Edges that have negative cumulative hazard for chlorine diffusion. Each row corresponds to an edge $\{i, j\}$ linking a prior adopter i to a susceptible subject j.

	Multivitamin			Chlorine		
	Coef	Std Error	p	Coef	Std Error	p
Intercept	3.708	1.200	0.002	1.046	0.903	0.25
Indegree targeting	-0.331	0.078	2.19e-5	-0.064	0.0878	0.47
Nomination targeting	0.205	0.090	0.023	-0.075	0.111	0.50
Village mean indegree	-0.274	0.058	2.46e-6	-0.163	0.057	0.004
Village male proportion	-4.305	1.002	1.73e-5	-0.437	0.944	0.64
Village mean age	-0.013	0.026	0.62	0.005	0.020	0.81
Village SES	-0.117	0.023	6.19e-7	-0.176	0.021	$<$ 1e-10
Adopter male	-0.237	0.074	0.001	-0.293	0.074	7.09e-5
Adopter age	0.002	0.003	0.49	0.0008	0.002	0.74
Adopter persons in house	-0.008	0.020	0.70	0.005	0.020	0.81
Adopter married	0.035	0.075	0.64	-0.138	0.076	0.07

Table 8. Cure probability model coefficients.

Statistics
in Medicine

	Multivitamin			Chlorine		
	Coef	Std Error	p	Coef	Std Error	p
Indegree targeting	-0.159	0.057	0.005	-0.125	0.0523	0.018
Nomination targeting	-0.227	0.059	1e-4	0.074	0.069	0.279
Village mean indegree	-0.028	0.036	0.437	0.030	0.038	0.432
Village male proportion	-1.083	0.625	0.083	0.562	0.707	0.426
Village mean age	0.016	0.015	0.265	0.047	0.015	0.001
Village SES	-0.019	0.013	0.134	0.013	0.014	0.336
Adopter male	-0.054	0.049	0.271	-0.101	0.050	0.043
Adopter age	-0.001	0.001	0.440	-0.002	0.002	0.160
Adopter persons in house	-0.013	0.014	0.379	-0.003	0.013	0.850
Adopter married	-0.002	0.045	0.959	0.011	0.043	0.799

Table 9. Failure time model coefficients from the cure mixture model

	Multivitamin				Chlorine			
	Coef	HR	95%CI(HR)	p	Coef	HR	95%CI(HR)	p
Indegree targeting	-0.31	0.73	(0.64, 0.83)	< 0.01	-0.00	0.99	(0.87, 1.15)	0.99
Nomination targeting	0.19	1.21	(1.06, 1.38)	< 0.01	-0.05	0.95	(0.80, 1.12)	0.52
Village mean indegree	-0.23	0.80	(0.74, 0.86)	< 0.01	-0.09	0.91	(0.83, 0.99)	0.04
Village male proportion	-4.5	0.01	(0.00, 0.04)	< 0.01	-0.58	0.56	(0.12, 2.55)	0.45
Village mean age	-0.01	0.99	(0.96, 1.02)	0.54	-0.00	0.99	(0.96, 1.03)	0.80
Village SES	-0.12	0.88	(0.86, 0.91)	< 0.01	-0.19	0.83	(0.80, 0.85)	< 0.01
Adopter male	-0.23	0.79	(0.71, 0.88)	< 0.01	-0.33	0.72	(0.64, 0.81)	< 0.01
Adopter age	0.00	1.00	(0.99, 1.00)	0.60	0.00	1.00	(0.99, 1.00)	0.99
Adopter persons in house	-0.01	0.99	(0.97, 1.02)	0.71	0.00	1.00	(0.97, 1.04)	0.86
Adopter married	0.04	1.04	(0.94, 1.15)	0.41	-0.13	0.87	(0.78, 0.98)	0.02

Table 10. Results from Exponential waiting time distribution.

Prepared using simauth.cls

	Multivitamin				Chlorine			
	Coef	HR	95%CI(HR)	p	Coef	HR	95%CI(HR)	p
Indegree targeting	-0.30	0.74	(0.65, 0.85)	< 0.01	-0.04	0.96	(0.84, 1.10)	0.56
Nomination targeting	0.14	1.15	(1.00, 1.31)	0.04	-0.04	0.96	(0.81, 1.13)	0.63
Village mean indegree	-0.21	0.81	(0.75, 0.88)	< 0.01	-0.10	0.90	(0.83, 0.99)	0.03
Village male proportion	-4.0	0.02	(0.00, 0.08)	< 0.01	-0.46	0.63	(0.14, 2.87)	0.55
Village mean age	-0.01	0.99	(0.96, 1.03)	0.61	0.00	1.01	(0.97, 1.04)	0.73
Village SES	-0.10	0.90	(0.88, 0.93)	< 0.01	-0.16	0.86	(0.83, 0.88)	< 0.01
Adopter male	-0.22	0.80	(0.72, 0.90)	< 0.01	-0.29	0.75	(0.67, 0.84)	< 0.01
Adopter age	0.00	1.00	(0.99, 1.00)	0.61	-0.00	1.00	(0.99, 1.00)	0.98
Adopter persons in house	-0.01	0.99	(0.97, 1.02)	0.60	0.00	1.00	(0.97, 1.03)	0.89
Adopter married	0.03	1.03	(0.93, 1.14)	0.52	-0.16	0.89	(0.80, 0.99)	0.04

Table 11. Results from Weibull waiting time distribution.

	Coef	HR	95%CI(HR)	p-value
Village 2	0.27	1.31	(0.78, 2.22)	0.31
Village 3	0.53	1.69	(0.86, 3.31)	0.12
Village 4	-0.51	0.60	(0.36, 0.98)	0.04
Village 5	-0.70	0.50	(0.34, 0.72)	< 0.01
Village 6	0.31	1.36	(0.98, 1.90)	0.07
Village 7	0.04	1.04	(0.75, 1.45)	0.8
Village 11	-0.71	0.49	(0.36, 0.67)	< 0.01
Village 12	0.08	1.08	(0.76, 1.55)	0.67
Village 13	-0.09	0.91	(0.67, 1.24)	0.56
Village 14	-0.18	0.83	(0.55, 1.27)	0.40
Village 15	-0.34	0.71	(0.46, 1.09)	0.11
Village 17	0.13	1.13	(0.61, 2.10)	0.69
Village 18	0.30	1.35	(0.69, 2.63)	0.39
Village 19	0.04	1.05	(0.65, 1.69)	0.86
Village 20	-0.16	0.85	(0.52, 1.42)	0.54
Village 21	-0.04	0.96	(0.62, 1.51)	0.87
Village 22	0.52	1.67	(1.22, 2.31)	< 0.01
Village 23	-0.27	0.76	(0.56, 1.04)	0.09
Village 24	0.18	1.20	(0.89, 1.62)	0.22
Village 25	0.54	1.71	(1.21, 2.42)	< 0.01
Village 26	0.07	1.07	(0.78, 1.47)	0.66
Village 27	-0.03	0.97	(0.70, 1.35)	0.87
Village 28	-0.44	0.65	(0.38, 1.09)	0.10
Village 29	0.42	1.53	(1.02, 2.29)	0.04
Village 30	-0.26	0.77	(0.57, 1.05)	0.10
Village 32	0.07	1.07	(0.70, 1.64)	0.75
Adopter male	-0.16	0.86	(0.77, 0.96)	< 0.01
Adopter age	0.00	1.00	(0.99, 1.00)	0.40
Adopter persons in house	-0.02	0.98	(0.95, 1.00)	0.11
Adopter married	0.02	1.02	(0.92, 1.13)	0.73

Table 12. Village fixed effects for the adoption of multivitamins. Each village had a dummy variable in Cox regression, and village 1 was treated as the base group. Villages 22 and 25 had the highest diffusion rate while villages 5 and 11 had the lowest diffusion rate.

Statistics
in Medicine

	Coef	HR	95%CI(HR)	p-value
Village 2	0.34	1.40	(0.76, 2.60)	0.28
Village 3	0.15	1.17	(0.62, 2.21)	0.63
Village 4	-0.10	0.90	(0.51, 1.59)	0.72
Village 5	-0.44	0.65	(0.42, 0.99)	0.05
Village 6	0.76	2.14	(1.52, 3.02)	< 0.01
Village 7	0.46	1.58	(1.09, 2.28)	0.02
Village 9	0.19	1.21	(0.82, 1.78)	0.34
Village 10	-0.29	0.75	(0.50, 1.13)	0.17
Village 11	-0.24	0.78	(0.55, 1.12)	0.18
Village 12	0.20	1.22	(0.82, 1.83)	0.32
Village 13	-0.01	0.99	(0.70, 1.39)	0.95
Village 14	0.43	1.54	(0.93, 2.55)	0.09
Village 15	0.32	1.38	(0.90, 2.11)	0.14
Village 16	-0.11	0.90	(0.59 1.35)	0.60
Village 17	-0.08	0.93	(0.50, 1.71)	0.81
Village 18	0.28	1.33	(0.62, 2.82)	0.46
Village 19	-0.20	0.82	(0.45, 1.49)	0.52
Village 20	0.02	1.02	(0.53, 1.97)	0.95
Village 21	0.60	1.82	(1.15, 2.87)	0.01
Village 22	0.86	2.36	(1.66, 3.35)	< 0.01
Village 24	0.09	1.09	(0.77, 1.56)	0.63
Village 26	0.62	1.85	(1.31, 2.63)	< 0.01
Village 27	0.32	1.37	(0.97, 1.95)	0.08
Village 28	0.19	1.21	(0.73, 2.00)	0.45
Village 29	0.65	1.92	(1.21, 3.07)	0.01
Village 30	-0.34	0.71	(0.50, 1.01)	0.06
Adopter male	-0.22	0.80	(0.71, 0.90)	< 0.01
Adopter age	0.00	1.00	(0.99, 1.00)	0.77
Adopter persons in house	0.00	1.00	(0.97, 1.04)	0.78
Adopter married	-0.08	0.92	(0.82, 1.03)	0.17

Table 13. Village fixed effects for the adoption of chlorine. Each village had a dummy variable in the Cox regression, and village 1 was treated as the base group. Village 22 had the highest diffusion rate while villages 5 and 30 had the lowest diffusion rate.