

No Magic Bullet: A Theory-Based Meta-Analysis of Markov Transition Probabilities in Studies of Service Systems for Persons With Mental Disabilities

Hugh Stephen Leff, Ph.D., Clifton M. Chow, Ph.D., Stephen C. Graves, Ph.D.

Objective: A random-effects meta-analysis of studies that used Markov transition probabilities (TPs) to describe outcomes for mental health service systems of differing quality for persons with serious mental illness was implemented to improve the scientific understanding of systems performance, to use in planning simulations to project service system costs and outcomes over time, and to test a theory of how outcomes for systems varying in quality differ.

Methods: Nineteen systems described in 12 studies were coded as basic (B), maintenance (M), and recovery oriented (R) on the basis of descriptions of services provided. TPs for studies were aligned with a common functional-level framework, converted to a one-month time period, synthesized, and compared with theory-based expectations. Meta-regression was employed to explore associations between TPs and characteristics of service recipients and studies.

Results: R systems performed better than M and B systems. However, M systems did not perform better than B systems. All systems showed negative as well as positive TPs. For approximately one-third of synthesized TPs, substantial interstudy heterogeneity was noted. Associations were found between TPs and service recipient and study variables

Conclusions: Conceptualizing systems as B, M, and R has potential for improving scientific understanding and systems planning. R systems appear more effective than B and M systems, although there is no “magic bullet” system for all service recipients. Interstudy heterogeneity indicates need for common approaches to reporting service recipient states, time periods for TPs, service recipient attributes, and service system characteristics. TPs found should be used in Markov simulations to project system effectiveness and costs of over time.

Psychiatric Services 2017; 68:278–287; doi: 10.1176/appi.ps.201500523

In 1986 in a seminal article on the need for a theory of psychiatric treatment systems for persons with serious mental illness, Hargreaves (1) wrote, “To optimize treatment system effectiveness within available resources, we need some logical tools to help us calculate the implications of the knowledge gained from clinical trials. . . . This requires a theory of the way treatment systems interact with the life course of persons in each major target group. . . . Such a theory of mental health services must be sufficiently detailed and valid to forecast an array of impacts of proposed system changes. Such a theory would be a stimulus and guide to research, as well as a tool for program management.”

Hargreaves proposed that a stochastic model (that is, one based on outcomes expressed as probabilities) in the form of a discrete first-order stationary Markov process (that is, one consisting of a fixed spectrum of outcome states and probabilities of transitioning between states that remain the same in different time periods) provides a promising approach to the formulation of such a theory. This article assumes a theory similar to that of Hargreaves, operationally defining outcomes as transition probabilities (TPs) from functional

levels (FLs) (described below) prior to the receipt of services to destination FLs after receipt and estimating how service system characteristics (also described below) affect these transitions. An example of a TP would be the probability of moving from having acute symptoms to not being acutely symptomatic after receiving recovery-oriented services for one month.

Populating a model for program management and planning requires different types of data, but, as Hargreaves noted, perhaps the most difficult to estimate for a Markov model is data on treatment system outcomes in the form of TPs. Fortunately, since the publication of Hargreaves’ article, a number of studies of psychiatric treatment systems presenting TP data have appeared. This article presents the first random-effects meta-analysis of such data known to us. Random-effects meta-analyses synthesize the outcomes of related but not identical studies, taking into account both within- and between-study sources of sampling error (2).

For this meta-analysis, we identified studies of persons with diagnoses of schizophrenia, bipolar disorder, borderline disorder, and antisocial personality disorder, referring to

these collectively as “serious mental illness.” Our theoretical rationale for this focus was not because these diagnoses provide an exhaustive account of serious mental illness—they do not (3–6)—but because persons in these groups have similar needs for community systems, are typically treated in public systems, are frequently classified in ways that align with the FL system we use, and are generally included in definitions of serious mental illness. Our practical reason was that persons with these diagnoses are frequently the focus of studies of public mental health systems.

There is widespread agreement that public community mental health systems for persons with serious mental illness are in crisis, which has resulted in large numbers of incarcerations and homelessness and has strained emergency room and inpatient resources. Data from 2002 and 2005 suggest that census declines in state psychiatric hospitals are reversing (7), such that some have called for a “return to the asylum” (8). Although this crisis is undoubtedly a result of resource constraints, it is also attributable to inadequate systems planning of services and unrealistic estimates of the resources required to provide these services. In 1979, Bachrach (9) noted that planning for deinstitutionalization was inadequate and later wrote (10), “Although some planners and planning agencies continue to stress the development of model programs as solutions for the varied problems of deinstitutionalization, discrepancies between isolated successful model endeavors and widespread service system failures are becoming so apparent that the need for systems-oriented planning strategies is increasingly acknowledged.”

Markov TP-based planning models have the potential to improve mental health system planning by clarifying services and resources necessary to adequately care for persons with serious mental illness. A concrete example resulting from the application of a Markov model to settle a right-to-treatment suit in Arizona was provided by Leff and colleagues (11). Although by no means guaranteeing that services and resources required will be provided, Markov planning models can alert mental health system stakeholders to relationships between needs and services, magnitude of need, and extent and consequences of shortfalls. Markov models also allow for more nuanced theories of service recipient outcomes through subgroup analyses and system component planning. As James and colleagues (12) noted, “Health state models have several distinct advantages over traditional . . . approaches to analyzing data for complex diseases such as schizophrenia. First, they provide a convenient framework for performing longitudinal analyses. . . . Second, the partitioning of the population into health states leads to a more richly informative analysis of the differences between populations than simply examining mean differences. For example, it may be the case that one population does not dominate the other in terms of overall level of health but that extreme states are more common in one group than the other. Finally, stationary distributions can be combined with a wide variety of outcome variables, such as costs [for planning].”

Markov models can be usefully contrasted with conventional growth modeling approaches. As Jung and Wickrama (13) noted, conventional growth modeling approaches assume that individuals come from a single population and that a single growth trajectory can adequately approximate an entire population. These approaches also assume that independent variables and covariates affecting growth factors influence each individual in the same way. Yet, theoretical frameworks and existing studies, such as the one reported here, often categorize individuals into distinct subpopulations differentially affected by treatments and covariates. Markov approaches more fully represent the heterogeneity of subpopulation growth trajectories within larger populations.

This study had two goals: to contribute to our theoretical understanding of psychiatric treatment system effectiveness and to generate TP inputs from multiple independent studies for more realistic Markov modeling.

SPECIFIC OBJECTIVES

In this article, we describe a methodology for meta-analyzing system outcomes in the form of Markov TPs between discrete states. We use the generic term “discrete states” (14), rather than “health states,” because some studies base states on services used rather than on functioning or symptoms. We also generated outcome estimates for analyzing the performance of systems and modeling by measuring TPs associated with different system types.

Furthermore, we tested an evidence-oriented theory of mental health systems proposing that systems consisting of more comprehensive, evidence-based, and rehabilitation-oriented services would produce better outcomes than systems that are less comprehensive. Specifically, we tested the hypothesis that TPs for service systems coded as recovery oriented (R) (services more comprehensive, evidence based, and rehabilitation oriented) would be more positive and less negative than systems coded as basic (B) (services least comprehensive, minimally evidence based, and not rehabilitation oriented) or as maintenance oriented (M) (services moderately comprehensive, treatments as usual, minimally evidence based, and rehabilitation oriented) by testing predictions that TPs would be more positive for R systems than for B and M systems and would be more positive and less negative for M systems than for B systems. Our theory that service systems could be coded as B, M, and R was based on a body of evaluation and planning studies typically comparing from two to four systems categorized as “lower cost,” “services as usual,” “more restrictive,” “lower quality,” or “minimal” with those categorized as “higher cost,” “enhanced,” “less restrictive,” “higher quality,” “community based,” or “evidence based” (15–20). Our goal was to better understand service recipient and study factors that might influence or bias TPs and identify scientific questions for further study in order to contribute to guidelines for collecting, synthesizing, and reporting TP data for scientific and planning purposes (21,22).

METHODS

Studies were eligible if they were in English and reported Markov analyses of treatments for persons with serious mental illness. Studies could be of any type. Bibliographic databases searched included Alt-HealthWatch (EBSCOhost); BIOSIS Previews (ISI Web of Knowledge); CAB Abstracts Archive; History of Science, Technology, and Medicine; PsycINFO (EBSCOhost); PubMed (MEDLINE); and Science Citation Index Expanded (ISI Web of Science). [More details of the bibliographic databases searched are provided in an online supplement to this article.]

The total number of candidate studies identified and retrieved was 61. A total of 42 studies (69%) were excluded because they focused on disorders or conditions other than serious mental illnesses (for example, depression, post-traumatic stress disorder, anorexia, substance use disorders, and suicide), and seven (11%) were excluded for one or more of the following reasons: states could not be cross-walked to the common FL framework; there was insufficient information provided on services to code systems as B, M, or R (the case for several studies of psychiatric medications); and other data, such as numbers of observations on which TPs were based—necessary to weight probabilities in the synthesis—were not provided. Twelve usable studies remained, which provided data for 19 study-level systems (20%) (1,12,23–32; personal communication, Hughes D, April 2015).

Procedure

All steps in the procedure were reviewed by the first and second authors. Coding reliability for FLs and service system type was assessed (described below). Service recipient states were cross-walked to a common FL framework employed to align different states, the Resource Associated Functional Level Scale (RAFLS) (described below). Transitions between states, whether as distributions of persons or probabilities, were represented for each system as a TP matrix of originating and destination FL states. If time periods were other than one month, we converted TPs into monthly rates by assuming that clients exit from current states at an exponential rate—a standard assumption when analyzing transitions from a Markov perspective (33–35).

All study-level systems were coded as B, M, or R. Variables were extracted for studies (for example, publication date), originating FL states (for example, number of observations), and populations (for example, percentage with schizophrenia). TPs for the same system types were synthesized. Comprehensive meta-analysis random-effects option was used (36,37). Rows of TPs for originating FLs and study-level systems were assembled to create full TP matrices for B, M, and R systems, and cells were compared with test study predictions. TP matrices were characterized in terms of average net-positive TPs (ANPTs) (defined below), and these measure were correlated, when data permitted, with service recipient and study characteristics by using meta-regression.

Variables

FLs. The common FL framework for this meta-analysis was the RAFLS, a reliable and valid measure of FL for persons with serious mental illness (11). Similar FL measures have been used frequently in mental health systems evaluation and planning (38–41). The RAFLS levels are as follows: FL 1, at risk, acutely symptomatic, unable or unwilling to participate in own care; FL 2, at risk, acutely symptomatic, able and willing to participate in own care; FL 3, symptoms not acute but lacking activities of daily life (ADL) skills; FL 4, possesses ADL skills, lacks community living skills; FL 5, possesses community living skills, vulnerable to stresses of everyday life; FL 6, requires specialty care but able to function except under unusual stress; FL 7, independent of the mental health system, can use generic health and human services. [Fuller definitions of these levels are provided in the online supplement.] The cross-walk was based on definitions of consumer behaviors before and after receipt of services. When only information on transitions to and from services was provided, FL states were coded on the basis of behaviors typically associated with the services described. Likely errors associated with the latter approach are discussed below. The first and second authors coded FLs. Interrater reliability calculated as the joint probability of agreement was .9. Where coding differed, authors discussed discrepancies. Consensus was possible in all cases.

Service system type. Service systems were coded as predominantly B, M, and R on the basis of system descriptions in the studies. If references were made to other articles or Web sites for fuller descriptions, these were consulted. Systems including only inpatient, emergency, and limited outpatient follow-up were coded as B. Systems also including a range of non-evidence-based community mental health center treatments and custodial services, such as day care, were coded as M. If reference was made to one or more evidence-based programs or to community support, psychosocial rehabilitation, or recovery, systems were coded as R. These categories are subsumptive because typically R systems offer the services of M systems and M systems offer the services of B systems. Systems might also be mixed; however, the number of studies available and level of detail about services did not support exploring this. The first and second authors also coded system type. Interrater reliability, calculated as the joint probability of agreement, was .8. If coding differed, authors discussed differences. Consensus was possible in all cases.

Study variables. Study-level variables coded (Table 1) were system type, study-level system description, first author, publication date, material type, study-level state measure, and RAFLS FLs coded.

Attributes of systems coded as B, M, and R. Table 2 lists attributes of system types: number of systems, number of service recipients (unique), number of observations for transitions or TPs, percentage of studies appearing or

TABLE 1. Characteristics of 12 studies and 19 systems included in the meta-analysis

System type	Study-level system description	Reference	Publication date	Material type	Study-level state measure ^a	RAFLS FLs ^b
B	Fee-for-service county system	Elsesser (23)	1991	Thesis	RAFLS	1–7
B	Mental health system (Australia)	Langley-Hawthorne (28)	1996	Thesis	Service type and location	1, 3, 6
B	Department of Veterans Affairs clinic	James et al. (12)	2006	Peer-reviewed article	PANSS	1–6
B	Hospital-based inpatient and outpatient services without needs monitor (Netherlands) (2 different data sets)	Drukker et al. (27) ^c	2012	Peer-reviewed article	Service type and location	1, 2, 3, 6
M	Community mental health center	Hargreaves (1)	1986	Peer-reviewed article	GAF	1–7
M	Ambulatory mental health treatment setting (service recipients with borderline personality disorder)	Perry et al. (24)	1987	Peer-reviewed article	Service type and location	2, 5, 6
M	Ambulatory mental health treatment setting (service recipients with bipolar disorder)	Perry et al. (24)	1987	Peer-reviewed article	Service type and location	2, 5, 6
M	Community mental health center	Liu et al. (29)	1992	Peer-reviewed article	Service type and location	1, 2, 4, 5
M	Community mental health center	James et al. (12)	2006	Peer-reviewed article	PANSS	1–6
M	Community mental health center (Spain)	Moreno et al. (30)	2007	Peer-reviewed article	Service type and location	2, 4, 6
M	County mental health system without jail diversion	Hughes et al. (31); Hughes ^d	2012	Peer-reviewed article	TRAG	1–7
R	Community support program, housing continuum	Drachman (26)	1981	Peer-reviewed article	Service type and location	1–6
R	County pilot managed care system	Elsesser (23)	1991	Thesis	RAFLS	1–7
R	Comprehensive psychosocial rehabilitation program	Miller et al. (32)	2010	Peer-reviewed article	MORS	1, 3, 4, 5, 6, 7
R	Hospital-based inpatient and outpatient services with needs monitor (Netherlands) (2 different data sets)	Drukker et al. (27) ^c	2012	Peer-reviewed article	Service type and location	1, 2, 3, 6
R	County mental health system with jail diversion	Hughes et al. (31); Hughes ^d	2012	Peer-reviewed article	TRAG	1–7
R	Comprehensive county community mental health programs with enhancements	Yoon et al. (25)	2013	Peer-reviewed article	Service type and location	2, 3, 4, 6

^a RAFLS, Resource Associated Functional Level Scale; PANSS, Positive and Negative Syndrome Scale; GAF, Global Assessment of Functioning; TRAG, Texas Recommended Assessment Guidelines; MORS, Milestones of Recovery Scale

^b FL, functional level. FL 1, at risk, acutely symptomatic, unable or unwilling to participate in own care; FL 2, at risk, acutely symptomatic, able and willing to participate in own care; FL 3, symptoms not acute but lacking activities of daily life (ADL) skills; FL 4, possesses ADL skills, lacks community living skills; FL 5, possesses community living skills, vulnerable to stresses of everyday life; FL 6, requires specialty care but able to function except under unusual stress; FL 7, independent of the mental health system, can use generic health and human services

^c Data for two systems were coded from this study.

^d Personal communication, Hughes D, April 2015

completed in 2000 or later, designs, study-level state measure, and data sources for state information.

Study-level Markov property variables. The accuracy of predictions based on matrices of Markov TPs is a function of the degree to which the TPs are shown to have “Markov properties” (1,33). Study findings were coded with respect to the three most commonly implemented tests for Markov properties: tests of “stationarity” or the stability (for example, reliability) of TPs over time (12,23), tests comparing whether current state alone or current state in combination with other variables best predicted subsequent states (termed first- versus second-order properties) (24), and tests of the predictive validity of Markov TPs based on the ability

of a set of TPs for one sample to predict transitions for different or hold-out samples.

Service recipient variables coded. Table 3 lists sociodemographic and clinical variables of service recipients by system type: percentages of persons in studies who received a diagnosis of schizophrenia or a related diagnosis, bipolar disorder, depression, and comorbid substance abuse; average age; percentage male; and percentage white.

ANPTs. Exploring the relationship between TPs and service recipient and study variables through meta-regression required calculating a standardized summary measure of how well persons were being served for each B, M, and

TABLE 2. Attributes of 19 systems included in 12 studies in the meta-analysis, by system type

Typical description	Basic (B): services or clinic care as usual		Maintenance (M): community mental health centers or clinics		Recovery-oriented (R): community support enhanced or specialized programs		All	
	N	%	N	%	N	%	N	%
N of study-level systems	5	26	7	37	7	37	19	100
N of service recipients (unique)	3,688	18	5,951	29	13,675	68	20,222	100
N of observations for transitions or transition probabilities	8,713	3	59,528	18	268,168	80	336,409	100
% of studies appearing or completed in 2000 or after	3	60	3	43	5	71	11	58
Designs								
Descriptive	2	40	5	71	3	43	10	53
Pre-post comparison	1	20			1	14	2	11
Quasi-experiment	1	20	2	29	2	29	5	26
Randomized controlled trial	1	20			1	14	2	11
Study-level state measure								
Functional level ratings	1	20	2	29	3	43	6	32
Symptom or pattern	1	20	1	14			2	11
Treatment or service type or location	3	60	4	57	4	57	11	58
Data source for state information								
Administrative data (for example, claims)	0	—	1	14	2	29	3	16
Patient registry	3	60	1	14	2	29	6	32
Rating for research or evaluation studies	2	40	5	71	3	43	10	53

R matrix. We considered an outcome positive if there was a transition to FL 5 or 6 (including static TPs), and we considered an outcome negative when there was a transition to FL 1, 2, 3, or 4 (including static TPs). For each origin FL and each service type, we next computed the net-positive TP, equal to the probability of a positive outcome minus the probability of a negative outcome. To obtain an ANPTP measure for each matrix type, we then averaged the net-positive TPs over the origin FLs.

Data Analysis

Comparison of TP matrices. Our procedures yielded a matrix of 54 cells for each system type (six origin FLs and nine destination states). Our theory yielded predictions for comparing each cell. For each of the 54 comparisons, the values in one matrix can be greater than, smaller than, or tied with the values in the other. Comparisons can be made by row and by matrix. These differences can be consistent with our predictions (+) or inconsistent (−), or in the case of ties, they can be nondiscriminating (=). The sign test is a nonparametric statistical test fitting data of this type (42), calculating probabilities for numbers of +s and −s, with ties being excluded.

Meta-regression. Using linear Pearson product-moment and point-biserial regression, we correlated the ANPTP measure with service recipient and study variables.

RESULTS

Table 1 shows that study dates ranged from 1981 through 2013. Ten appeared as peer-reviewed articles. Two were

theses. The 12 studies yielded 19 study-level systems, five of which were coded as B, seven as M, and seven as R. In eight instances, study-level states were based on FL or symptoms. In the 11 others, study-level state was based on service types or locations. Coding FL7 was possible for at least one example of B, M, and R systems.

System Type Variables

Table 2 shows that B study-level systems were typically termed “services or clinical care as usual”; M systems were typically termed “community mental health centers or clinics”; and R systems were typically termed “community support, enhanced, or specialized programs.” Study-level systems coded R had the largest number of unique service recipients (13,675) and the largest number of transition observations (268,168). Systems coded M had the next largest numbers (5,951 and 59,528, respectively), and B systems had the lowest (3,688 and 8,713, respectively). Because our analyses of matrix differences and correlations with service recipient and study-level variables were reflective of thousands of individuals and tens of thousands of transition observations, we discuss moderate-to-high effect sizes despite the fact that they may have been associated with moderate p values. As Cohen (43) noted, “[T]he primary product of a research inquiry [should be] one or more measures of effect size, not p values.” Moreover, estimates of p values based on Ns for FLs and TP matrix cells ranging from six to 54 almost certainly would have been lower if we had had access to individual-level data.

To focus on the most notable differences among intervention types, systems coded as B were smaller and most

TABLE 3. Sociodemographic and clinical characteristics of service recipients reported in 12 studies included in the meta-analysis, by system type^a

Characteristic	Basic (B)			Maintenance (M)			Recovery-oriented (R)			Total		
	% or M	Systems with data N %		% or M	Systems with data N %		% or M	Systems with data N %		% or M	Systems with data N %	
N of treatment arms	na	5	26	na	7	37	na	7	37	na	19	100
Schizophrenia or related diagnosis (%)	88	4	80	53	6	86	58	4	57	65	14	74
Age (M)	43	4	80	35	5	71	41	4	57	39	13	68
Male (%)	71	2	40	58	5	71	57	1	14	68	8	42
White (%)	64	1	20	58	2	29	69	2	29	64	5	26
Bipolar disorder (%)	nr	0	—	28	3	43	38	2	29	32	4	21
Depression (%)	nr	0	—	19	1	14	22	1	14	20	2	11
Comorbid substance use disorder (%)	nr	0	—	41	1	14	54	2	29	49	3	16

^a na, not applicable; nr, not reported

likely to have included studies in which service recipient states were based on treatment or service types or locations (60%) and to have extracted data from patient registries (60%). M systems were intermediate in size, least likely to have appeared in studies completed in 2000 or after (43%), most likely to include descriptive studies (71%), and most likely to have used ratings data from research or evaluations (71%). R systems were the largest in size and most likely to have appeared in studies appearing or completed in 2000 or after (71%) and to have based states on FL ratings (43%).

With respect to Markov properties, stationarity was empirically disconfirmed for one (20%) B study-level system (23), for three (43%) M systems (1,24), and for one (14%) R system (23) [see online supplement]. In all studies in which stationarity was not confirmed, the cause was a subgroup of service recipients who tended to transition less than others (1,23,24) thereby increasing the proportions of persons in “static” TP cells. No study provided actual analyses of how changers differed from nonchangers. However, study authors speculated that nonchangers might be persons with certain diagnoses, older persons, or persons adhering to patterns of previous service use.

Testing whether TPs had first-order properties was done as follows. Testing for one of the B systems (20%) showed that the fit between expected and observed transitions was greater if persons were grouped into those who transitioned more and less frequently (23). For two of the M systems (29%), testing indicated that a second-order model based on prior service utilization fit the observed cell value data better than a first-order model (24). For three of the R systems (43%), goodness-of-fit tests for two systems (25,26) supported first-order properties while one system (23) was consistent with second-order properties. Despite the fact that some studies suggested higher-order models, only one (24) reported second-order Markov TPs. Predictive validity was found for all systems in which this was tested (23,26,27).

For no clinical or sociodemographic variable were data presented for all study-level systems (Table 3). Compared with R and M systems, B systems had higher percentages of service recipients with diagnoses of schizophrenia or related

disorders, service recipients were slightly older, and the percentage of males was higher. Data on the remaining variables were too sparse for comment.

Synthesized TP Matrices

For each system type and TP cell, Table 4 summarizes the numbers of observations on which TPs were based along with the numbers of systems represented in the cell. ND indicates no data found for TPs to death. Also shown with asterisks are cells for which the probability of Q, a measure of interstudy heterogeneity, was less than the equivalent of .05 adjusted for the large number of comparisons. A total of 53 cells (35%) with data were found to have adjusted Qs with p values equivalent to <.05. These Q values probably reflect a mixture of true differences in services and service recipients between systems correctly coded as similar and a coding error.

Looking across system types, TPs have common features, many found in earlier studies. First, the most common one-month TP was to the same FL. Without new arrivals a static TP of .938, the highest in the table, would leave only 46% of original persons in that FL after 12 months. Next, for TPs from FL 3 and above (non-“acute” FLs), the next-highest TPs grouped by whether they were forward or backward were to immediately adjacent FLs (1). A high proportion of large changes should not be expected in short time periods. TPs from FLs 1 and 2, more “acute” states, had more variable destinations, suggesting that positive symptoms respond to medication more quickly, returning persons to a variety of baseline FLs, whereas negative symptoms and behaviors that cause people to be categorized at FL 3 and above require remediation by slower-acting psychosocial interventions.

For all but one FL group—persons originating from FL 6 in R systems—disappearance and death are the only ways persons exit mental health systems. Without disappearances, the number of persons in systems increases continuously, straining system capacities. Because of this Levin (44) has suggested that disappearance rates may be “the solution, not the problem” in providing care to meet expressed demand. Except for persons at FL 6 in R systems, there was no

TABLE 4. Synthesized one-month transition probabilities and numbers of observations and study-level systems, by functional level (FL) and system type

Origin FL	System type and N of observations and studies	Disappearance rate	Death	Destination FL ^a						
				FL 1	FL 2	FL 3	FL 4	FL 5	FL 6	FL 7
FL 1	Basic ^b	.019	.001	.895*	.034*	.043*	.002	.000	.006	.000
	Observations	630	604	630	423	630	26	26	423	26
	Studies	5	3	5	4	5	2	2	4	2
FL 1	Maintenance ^b	.058*	nr	.819*	.038*	.062	.008*	.011	.005	.000
	Observations	4,647	0	4,647	4,647	1,713	4,647	4,647	1,713	29
	Studies	4	0	4	4	3	4	4	3	2
FL 1	Recovery ^b	.003	.001	.825*	.036*	.038*	.060	.031	.007	.000
	Observations	1,350	1,222	1,396	1,371	1,396	174	174	1,396	153
	Studies	4	2	6	5	6	4	4	6	3
FL 2	Basic ^b	.006	nr	.014	.809*	.023*	.111	.009	.027	.000
	Observations	461	0	461	461	461	461	461	461	461
	Studies	4	0	4	4	4	4	4	4	4
FL 2	Maintenance ^b	.036	nr	.023*	.755*	.016*	.010*	.015*	.002	.000
	Observations	5,428	0	5,381	5,463	1,837	5,381	5,428	1,884	370
	Studies	6	0	4	6	3	4	6	5	2
FL 2	Recovery ^b	.007	.001	.023*	.815*	.046*	.055	.021*	.033	.000
	Observations	12,278	734	1,452	12,338	12,338	11,604	718	12,338	658
	Studies	5	2	5	6	6	4	3	6	2
FL 3	Basic ^b	.014	.003	.007*	.017	.888*	.038	.030	.003	.000
	Observations	1,775	1,230	1,775	839	1,775	545	1,481	839	545
	Studies	5	3	5	4	5	2	3	4	2
FL 3	Maintenance ^b	.055	nr	.005*	.034*	.781*	.108	.016	.001	.000
	Observations	3,764	0	3,764	3,764	3,764	3,764	3,764	3,764	2,033
	Studies	3	0	3	3	3	3	3	3	2
FL 3	Recovery ^b	.012*	.001	.006	.008*	.827*	.082*	.058*	.007*	.000
	Observations	57,302	1,248	3,119	57,395	57,601	56,353	1,871	57,601	1,778
	Studies	5	2	6	6	7	5	4	7	3
FL 4	Basic ^b	.009	nr	.001	.016	.032	.902	.033	.006	.000
	Observations	1,200	0	1,200	1,200	1,200	1,200	1,200	1,200	1,200
	Studies	2	0	2	2	2	2	2	2	2
FL 4	Maintenance ^b	.032	nr	.006*	.007*	.068	.837*	.038*	.011*	.000
	Observations	15,792	0	4,937	15,792	4,308	15,792	4,937	15,163	2,086
	Studies	5	0	4	5	3	5	4	4	2
FL 4	Recovery ^b	.023	nr	.009	.005*	.046*	.821*	.076	.019*	.001
	Observations	59,973	0	1,617	60,090	60,151	60,151	1,617	60,151	1,500
	Studies	3	0	4	4	5	5	4	5	3
FL 5	Basic ^b	.034*	.003	.005	.012	.025*	.033	.864*	.025	.000
	Observations	3,025	1,593	3,025	1,432	3,025	1,432	3,025	1,432	1,432
	Studies	3	1	3	2	3	2	3	2	2
FL 5	Maintenance ^b	.025	nr	.007	.014	.010	.017*	.900*	.027	.000
	Observations	6,373	0	6,242	4,225	2,017	6,242	6,373	2,148	186
	Studies	5	0	3	5	2	3	5	4	1
FL 5	Recovery ^b	.010*	nr	.004	.005	.015	.080*	.817*	.068	.002
	Observations	6,373	0	6,242	6,373	2,017	6,242	6,373	2,148	186
	Studies	2	0	4	3	4	4	4	4	3
FL 6	Basic ^b	.009	.001	.001	.010	.004	.012	.025	.938*	.000
	Observations	1,695	839	950	1,695	1,695	856	1,695	1,695	856
	Studies	4	2	3	4	4	2	2	2	2
FL 6	Maintenance ^b	.046*	nr	.007	.009	.007	.012*	.042	.877*	.000
	Observations	16,734	0	2,671	16,734	2,671	16,616	2,789	16,734	464
	Studies	6	0	3	6	3	4	5	6	1
FL 6	Recovery ^b	.010*	.001	.003	.004*	.004*	.006*	.103*	.858*	.011
	Observations	133,337	1,644	3,245	133,487	133,608	131,964	1,601	133,608	1,451
	Studies	5	2	6	6	7	5	4	7	3

^a FL 1, at risk, acutely symptomatic, unable or unwilling to participate in own care; FL 2, at risk, acutely symptomatic, able and willing to participate in own care; FL 3, symptoms not acute but lacking activities of daily life (ADL) skills; FL 4, possesses ADL skills, lacks community living skills; FL 5, possesses community living skills, vulnerable to stresses of everyday life; FL 6, requires specialty care but able to function except under unusual stress; FL 7, independent of the mental health system, can use generic health and human services. nr, not reported

^b Values in these rows are synthesized one-month transition probabilities, reflecting the estimated probability that cohorts at the origin FLs will remain at origins or transition to destination FLs, disappearance, or death after a month.

*Probability of Q (measure of heterogeneity), $p < .0003$ ($p < .05$ Bonferroni-adjusted for 152 comparisons)

evidence of movement to independence from the system. Although it is possible that some persons disappearing from systems had become system independent without its being recognized, the evidence suggests that even with our most effective services, “graduating people” to system independence is a rare event that needs to be better understood. Backward movement was present for all FLs in all types of systems. Not all services work for all recipients all the time. Systems must make provisions for recipients for whom first-line services do not work or have adverse consequences.

Table 5 shows that predicted ANPTPs for systems coded as R were greater than those for B ($p=.01$) and M ($p=.02$) systems. As predicted, average probabilities of transitioning to disappearance and for remaining the same were lower for R systems compared with B and M systems, although these differences were generally small, and sign tests indicated that these differences could have occurred by chance. Findings for backward movement, shown in Table 5, contrary to predictions, show that the means of TPs indicating backward movement were higher for R systems compared with B and M systems in many comparisons. Once again, sign tests showed that these differences could have occurred by chance. Nevertheless, these post hoc findings are interesting and may indicate a negative effect of “high expectation” programs on some service recipients, a finding in previous research on the effect of expectations on outcomes (45–48).

Predictions for M systems compared with B systems were in expected directions only for static rates. Differences were small, and sign tests indicated that the differences observed probably occurred by chance.

Meta-Regressions for Service Recipient and Study Variables

Service recipient variables. The high number of cells with values of Q unlikely by chance suggests that TPs were influenced by factors in addition to originating FL and study-level system type. Although these Q values may partially reflect coding errors, several studies have suggested that differences between interventions might also be attributable to service recipient variables (1,23,24). ANPTPs were regressed on service recipient variables to explore this possibility.

Systems consisting solely of persons with a diagnosis of schizophrenia or related disorders had lower ANPTPs than diagnostically mixed subgroups ($t=-1.99$, $df=12$, one tailed $p=.03$) [see online supplement]. Systems with more persons classified as white had higher ANPTPs, although the number of systems with data were very small ($N=5$). Subgroups with more males had higher ANPTPs, but p was above .20. Subgroups formed on the basis of age did not differ. Multicollinearity among variables is possible. Findings also

TABLE 5. Comparisons of average net-positive transition probabilities (ANPTPs), disappearance rates, static TPs, and backward TPs for B, M, and R system types^a

Variable	System		Systems compared ^b	Total comparisons minus ties ^c	Mean		
	1	2			R	M	B
ANPTP	R	B	32	47**	.098	-.021	.053
	R	M	30	45*			
	M	B	21	46			
Disappearance rate	R	B	3	6	.011	.042	.015
	R	M	6	6*			
	M	B	3	6			
Static TPs	R	B	3	6	.827	.829	.884
	R	M	3	6			
	M	B	4	6			
Backward movement	R	B	8	14	.020	.018	.014
	R	M	9	15			
	M	B	5	14			

^a Systems were coded as basic (B), maintenance (M), and recovery oriented (R).

^b Number of comparisons between system types in which values for system 1 type was as predicted when compared with system 2 values.

^c Total number of comparisons in which values were not tied and therefore could be included in sign tests. For example, ANPTP was higher for R systems in 32 of 47 comparisons with B systems.

* $p=.02$, ** $p=.01$, one-tailed sign test

raise the possibility that ANPTP differences between service types may have been moderated by the percentage of service recipients with diagnoses of schizophrenia and by methods with which functional level states were measured. Unfortunately, lack of data prevented further analyses of these possibilities.

Study variables. Except for the variable study-level state, all sign tests for study variables were two-tailed because our only hypothesis for these was that TPs based on functioning and symptoms would be higher than TPs based on services type or location, which could be constrained by service availability. The ANPTP for functioning- or symptom-based TPs was almost twice the size of the one for service-based TPs (.059 versus .030), suggesting that TPs based on service use should be considered low-side estimates of movement, although the value of p was .30. Correlations between ANPTPs and stationarity and predictive validity testing and publication date had p values above .30, giving no reason to believe that these variables influenced ANPTPs.

DISCUSSION AND CONCLUSIONS

Discrete states from diverse studies can be aligned with a common FL framework, making synthesis of TPs possible. Mental health systems described in diverse studies can be usefully characterized as B, M, and R. Although findings for interstudy heterogeneity suggest that systems may be further subdivided, current data are insufficient for this. First-order Markov TPs are highly informative ways to represent system outcomes, although in some cases it may be desirable

to characterize persons more complexly, by calculating TPs for subgroups or estimating higher-order TPs.

As hypothesized, R systems generated better outcomes than B or M systems, except that R systems also produced more backward movement, suggesting negative effects of high-expectation systems on some persons. The ubiquity of backward movements suggests that all types of systems should include services for persons who do not respond to or who are negatively affected by first-line services. Contrary to our hypotheses, M service packages did not outperform B ones in expected ways. Further research into this finding is needed, especially because M systems are common. Consistent with a theory that one size does not fit all, all systems produced diverse and complex outcomes for all FLs, with some probabilities of forward and backward movements, stasis, and disappearances. We did not find and should not expect to find “magic bullet” systems for persons with serious mental illness that produce only positive outcomes for all persons all the time. Again, all types of systems should include services for service recipients who do not respond to or are negatively affected by first-line services.

Most studies lacked TP information on death, an important omission given concerns about premature mortality (49,50). The ways in which these new TPs will affect overall system outcomes are not immediately obvious because of the complex backward and forward nature of the TPs observed and the role that disappearance rates play. Lower disappearance rates for R systems, especially compared with M systems, could substantially increase service use and costs. It will be important to use these TPs in simulation models to explore how interactions between these variables affect system outcomes, service utilization, and costs over time (11,51).

Simulation modeling holds promise for increasing the scientific understanding of mental health systems for persons with serious mental illness and for making system planning and implementation more realistic. The TP estimates provided here should be used in simulations to project how outcomes, service utilization, and costs can vary over time with different system configurations. We expect B, M, and R systems will be shown to differ in costs, both in simulation and in empirical studies (a review of empirical costs studies was beyond the scope of this study). Given that R systems are more effective than M and B systems, stakeholders will prefer such systems. However, cost estimates will provide information on the extent to which R systems are affordable and will inform discussions about what system configurations are possible given resource constraints.

If mental health system evaluation and planning through simulation is to progress, researchers need to reduce inter-study heterogeneity by agreeing on a common set of methods and standards for conceptualizing, estimating, and reporting the inputs required by models for implementing and reporting clinical trials (52). TPs should be based on FL assessments, not on service utilization. TPs to death should be estimated. Disappearance should be studied to clarify its

meaning and implications. There should be a common time period for TP assessments: one month seems most reasonable, although “semi-Markov models” with varying time periods are possible and should be explored (53). Studies should collect and report agreed-upon clinical and socio-demographic data to explore what works (and does not work) for whom and certain methodological features, such as testing for stationarity, should be routinely implemented. In addition, other system attributes thought to be related to performance should be provided and included in analyses—for example, information about how mental health services are financed. The development of such data, methods, and standards will enrich the data available and improve the quality of syntheses for estimating model inputs. This is required for the good science and more realistic and detailed planning that the current crisis in treating persons with serious mental illness demands, especially given the advent of integrated care.

AUTHOR AND ARTICLE INFORMATION

Dr. Leff and Dr. Chow are with the Department of Psychiatry, Cambridge Health Alliance—Harvard Medical School, Cambridge, Massachusetts. Dr. Leff is also with the Human Services Research Institute, Cambridge, Massachusetts. Dr. Graves is with the Department of Management Science, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts. Send correspondence to Dr. Chow (e-mail: c.chow@snhu.edu).

Dr. Graves reports receipt of research funding from Ferrovia Services, Ford Motor Company, and Samsung Corporation. The other authors report no financial relationships with commercial interests.

Received December 7, 2015; revision received June 9, 2016; accepted August 3, 2016; published online October 17, 2016.

REFERENCES

1. Hargreaves WA: Theory of psychiatric treatment systems: an approach. *Archives of General Psychiatry* 43:701–705, 1986
2. Borenstein M, Hedges LV, Higgins JPT, et al: A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 1:97–111, 2010
3. Schinnar AP, Rothbard AB, Kanter R, et al: An empirical literature review of definitions of severe and persistent mental illness. *American Journal of Psychiatry* 147:1602–1608, 1990
4. Hargreaves WA, LeGoullon M, Gaynor J, et al: Defining the severely mentally disabled. *Evaluation and Program Planning* 7: 219–227, 1984
5. Goldman HH, Regier DA, Taube CA, et al: Community mental health centers and the treatment of severe mental disorder. *American Journal of Psychiatry* 137:83–86, 1980
6. Goldman HH, Gattozzi AA, Taube CA: Defining and counting the chronically mentally ill. *Hospital and Community Psychiatry* 32: 21–27, 1981
7. Manderscheid RW, Atay JE, Crider RA: Changing trends in state psychiatric hospital use from 2002 to 2005. *Psychiatric Services* 60:29–34, 2009
8. Sisti DA, Segal AG, Emanuel EJ: Improving long-term psychiatric care: bring back the asylum. *JAMA* 313:243–244, 2015
9. Bachrach LL: Planning mental health services for chronic patients. *Hospital and Community Psychiatry* 30:387–393, 1979
10. Bachrach LL: New directions in deinstitutionalization planning. *New Directions for Mental Health Services* 17:93–106, 1983
11. Leff HS, Hughes D, Chow C, et al: Mental health allocation and planning simulation model; in *Handbook of Healthcare Delivery Systems*. New York, CRC Press, 2010

12. James GM, Sugar CA, Desai R, et al: A comparison of outcomes among patients with schizophrenia in two mental health systems: a health state approach. *Schizophrenia Research* 86:309–320, 2006
13. Jung T, Wickrama KAS: An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass* 2:302–317, 2008
14. Sugar CA, James GM, Lenert LA, et al: Discrete state analysis for interpretation of data from clinical trials. *Medical Care* 42:183–196, 2004
15. Dickey B, Fisher W, Siegel C, et al: The cost and outcomes of community-based care for the seriously mentally ill. *Health Services Research* 32:599–614, 1997
16. Grusky O, Tierney K: Evaluating the effectiveness of countywide mental health care systems. *Community Mental Health Journal* 25: 3–20, 1989
17. Grusky O: Models of local mental health delivery systems. *American Behavioral Scientist* 28:685, 1985
18. Frisman L, McGuire T: The economics of long-term care for the mentally ill. *Journal of Social Issues* 45:119, 1989
19. Regan J, Daleiden E, Chorpita B: Integrity in mental health systems: an expanded framework for managing uncertainty in clinical care. *Clinical Psychology: Science and Practice* 20:78–98, 2013
20. Miles DG: The Georgia experience: unifying state and local services around the balanced service system model. *New Directions for Mental Health Services* 1983:53–65, 1983
21. Kuntz K, Sainfort F, Butler M, et al: Decision and Simulation Modeling in Systematic Reviews. *Methods Research Report*. Rockville, Md, Agency for Healthcare Research and Quality, 2013
22. Caro JJ, Briggs AH, Siebert U, et al: Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Medical Decision Making* 32:667–677, 2012
23. Elssesser K: The Validation of a Simulation Model for the Allocation of Mental Health Services. Cambridge, Massachusetts Institute of Technology, 1991
24. Perry JC, Lavori PW, Hoke L: A Markov model for predicting levels of psychiatric service use in borderline and antisocial personality disorders and bipolar type II affective disorder. *Journal of Psychiatric Research* 21:215–232, 1987
25. Yoon J, Bruckner TA, Brown TT: The association between client characteristics and recovery in California's comprehensive community mental health programs. *American Journal of Public Health* 103:e89–e95, 2013
26. Drachman D: A residential continuum for the chronically mentally ill: a Markov probability model. *Evaluation and the Health Professions* 4:93–104, 1981
27. Drukker M, Joore M, van Os J, et al: The use of a Cumulative Needs for Care Monitor for individual treatment v care as usual for patients diagnosed with severe mental illness, a cost-effectiveness analysis from the health care perspective. *Epidemiology and Psychiatric Sciences* 21:381–392, 2012
28. Langley-Hawthorne C: Modeling the Lifetime Costs of Schizophrenia in an Australian Treating Environment. Bandoora, Victoria, Australia, La Trobe, 1996
29. Liu C-Y, Hu T-W, Jerrell J: A Markov analysis of the service system for severe mental illness. *Biometrical Journal* 34:443–457, 1992
30. Moreno B, Cervilla J, Luna JD, et al: Pattern of care for schizophrenia patients in Granada (Spain): a case register study. *International Journal of Social Psychiatry* 53:5–11, 2007
31. Hughes D, Steadman HJ, Case B, et al: A simulation modeling approach for planning and costing jail diversion programs for persons with mental illness. *Criminal Justice and Behavior* 39: 434–446, 2012
32. Miller L, Brown T, Pilon D, et al: Patterns of recovery from severe mental illness: a pilot study of outcomes. *Community Mental Health Journal* 46:177–187, 2010
33. Sonnenberg FA, Beck JR: Markov models in medical decision making: a practical guide. *Medical Decision Making* 13:322–338, 1993
34. Dias S, Welton NJ, Sutton AJ, et al: Evidence synthesis for decision making: 5. the baseline natural history model. *Medical Decision Making* 33:657–670, 2013
35. Pilla Reddy V, Kozielska M, Johnson M, et al: Modeling and simulation of the Positive and Negative Syndrome Scale (PANSS) time course and dropout hazard in placebo arms of schizophrenia clinical trials. *Clinical Pharmacokinetics* 51:261–275, 2012
36. Pettiti DB: Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. New York, Oxford University Press, 1994
37. Basu A, Meltzer HY, Dukic V: Estimating transitions between symptom severity states over time in schizophrenia: a Bayesian meta-analytic approach. *Statistics in Medicine* 25:2886–2910, 2006
38. Green RS, Gracely EJ: Selecting a rating scale for evaluating services to the chronically mentally ill. *Community Mental Health Journal* 23:91–102, 1987
39. Newman FL, Hunter RH, Irving D: Simple measures of progress and outcome in the evaluation of mental health services. *Evaluation and Program Planning* 10:209–218, 1987
40. Evans S, Greenhalgh J, Connelly J: Selecting a mental health needs assessment scale: guidance on the critical appraisal of standardized measures. *Journal of Evaluation in Clinical Practice* 6:379–393, 2000
41. Carter DE, Newman FL: A Client-Oriented System of Mental Health Service Delivery and Program Management: A Workbook and Guide. Washington, DC, US Department of Health, Education, and Welfare, Public Health Service, 1976
42. Siegel S, Castellan NJ: Nonparametric Statistics for the Behavioral Sciences, 2nd ed. New York, McGraw-Hill, 1988
43. Cohen J: Things I have learned (so far). *American Psychologist* 45: 1304–1312, 1990
44. Levin G: Point of view: poor quality is the solution, not the problem. *Health Care Management Review* 2:69–72, 1977
45. Bell M, Lysaker P: Levels of expectation for work activity in schizophrenia: clinical and rehabilitation outcomes. *Psychiatric Rehabilitation Journal* 19:71–76, 1996
46. O'Connell MJ, Stein CH: The relationship between case manager expectations and outcomes of persons diagnosed with schizophrenia. *Community Mental Health Journal* 47:424–435, 2011
47. Lamb HR, Goertzel V: High expectations of long-term ex-state hospital patients. *American Journal of Psychiatry* 129:471–475, 1972
48. Buhrmaster D, Hartman J, Menefee P, et al: Clients' reasons for dropping out of rehabilitation centers. *Psychological Reports* 51: 1307–1316, 1982
49. Leff HS, McPartland JC, Banks S, et al: Service quality as measured by service fit and mortality among public mental health system service recipients. *Mental Health Services Research* 6: 93–107, 2004
50. De Hert M, van Winkel R, Silic A, et al: Physical health management in psychiatric settings. *European Psychiatry* 25(suppl 2): S22–S28, 2010
51. Leff HS, Graves SC, Natkins J, et al: A system for allocating mental health resources. *Administration in Mental Health* 13:43–68, 1985
52. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials. Geneva, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998. http://www.fda.gov/ohrms/dockets/ac/02/briefing/3837b1_03_ICH%20e9.pdf
53. Cao Q, Buskens E, Feenstra T, et al: Continuous-time semi-Markov models in health economic decision making: an illustrative example in heart failure disease management. *Medical Decision Making* 36:59–71, 2016