PHARMACOLOGICAL REVIEWS
Copyright © 1974 by The Williams & Wilkins Co.

Vol. 26, No. 1
Printed in U.S.A.

Linearity and Superposition in Pharmacokinetics[1,2]

C. D. THRON

Department of Pharmacology and Toxicology, Dartmouth Medical School, Hanover, New Hampshire

THE analysis of the kinetics of drug absorption, distribution, metabolism, and excretion gives rise to a great variety of more or less complex mathematical models (17, 22, 23, 34, 35, 49). Many of these fall into the general class of systems known as *linear* systems. This class of systems is important for several reasons: 1) the property of linearity can be verified or excluded experimentally without reference to any particular kinetic model; 2) in formulating a kinetic model, the property of linearity can often be assumed or ruled out on general theoretical physicochemical grounds, with no commitment to a particular detailed structure for the kinetic model; and 3) linear systems all share certain properties that make it possible to predict important aspects of their behavior without detailed knowledge of their internal kinetics.

A linear system is defined as one which obeys the principle of superposition. This principle may be stated as follows:

> If ζ_1 is the system response to an input ξ_1, and ζ_2 is the response to the input ξ_2, and α and β are arbitrary coefficients, then $\alpha\zeta_1 + \beta\zeta_2$ is the response to the input $\alpha\xi_1 + \beta\xi_2$.

The great power and value of this principle comes from the fact that it is meaningful not only for inputs and responses that are simple numbers, but also for those that are

[1] This investigation was supported mostly by U. S. Public Health Service Research Grant NB 08710 from the National Institute of Neurological Diseases and Stroke. The author is grateful for the approval by the National Advisory Neurological Diseases and Stroke Council of the renewal application which would have supported the completion of this project, had funding been available.

[2] This paper is dedicated to the memory of those on both sides who suffered and died in the Christmas 1972 bombing of Hanoi and Haiphong: *That nation, that was wont to conquer others, Hath made a shameful conquest of itself.*

more complicated mathematical entities, such as functions or sets. The "response" or "output" ζ of a pharmacokinetic system, for example, is typically a function of time giving the drug concentration at some point in the system (*e.g.*, the plasma level). The sum $\alpha\zeta_1 + \beta\zeta_2$ is then likewise a function of time, formed by adding the functions ζ_1 and ζ_2, after multiplying by α and β, respectively. More generally, the output of a pharmacokinetic system might be considered to be a set of time-functions giving the drug concentrations at several points (*e.g.*, plasma, cerebrospinal fluid, myocardial tissue, *etc.*). In that case the sum $\alpha\zeta_1 + \beta\zeta_2$ is likewise a set of time-functions, formed by adding the corresponding time-functions of ζ_1 and ζ_2 after multiplying by α and β respectively. The possibilities for "input" are similar but slightly more complicated, because in order for us to state the principle of superposition in the simple form above it is necessary that every input uniquely determine the response. To do this, the "input" must comprise not only the time-functions specifying the injection rates at various points but also the initial conditions, and in some cases the past history of the system. Nevertheless, in spite of this complexity of the input ξ, we can still form the sum $\alpha\xi_1 + \beta\xi_2$ by adding the corresponding components (injection rates, initial conditions, *etc.*) of the two inputs ξ_1 and ξ_2 after multiplying by α and β respectively.

Readers long accustomed to intuitive addition in arithmetic and algebra may feel some intuitive resistance to this broader concept of "addition" of functions or sets of numbers or functions. The cure, if the author correctly remembers his own experience, is to accept the fact that an intuitive "feel" for this type of addition may not come at once, and to be willing to accept these new definitions of "addition" as purely formal rules.

Two examples will illustrate these concepts. Figure 1 shows simulated "plasma" levels for a compartmental pharmacokinetic model with three different inputs. Input *A* comprises 1) initial conditions of 2 units in the "plasma" and 0 units in the "gastrointestinal tract," together with 2) a constant intravenous infusion during the first 3 time units. Input *B* comprises 1) initial conditions of 0 units in both compartments, and 2) single doses introduced into the gastrointestinal tract at 1 and 4 time units. The "sum" of these two inputs can be defined as an input such that: a) the initial plasma level is the sum of the two initial plasma levels; b) the initial gastrointestinal tract level is the sum of the two initial gastrointestinal levels; and c) the drug injections and infusions given comprise all those given in the two inputs being added. In the example of figure 1, input *C* is the "sum" of inputs *A* and *B*, *i.e.*, it comprises 1) initial conditions of 2 units in the plasma and 0 units in the gastrointestinal tract, together with 2) a constant intravenous infusion during the first 3 time units and single oral doses at 1 and 4 time units. The com-

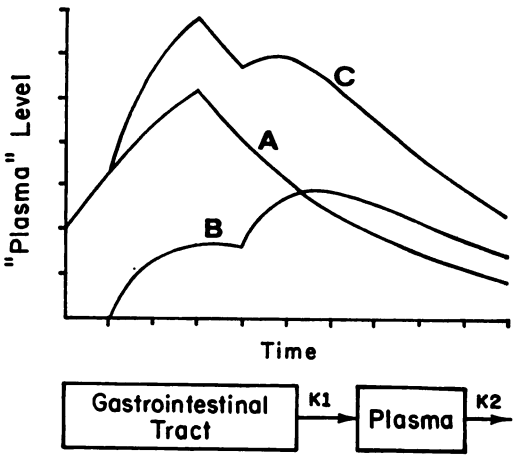

FIG. 1. Additivity of inputs and responses for a linear model. Curves *A*, *B*, and *C* are computed "plasma" levels for the compartmental model shown. First-order rate constants *K*1 and *K*2 are 0.7 and 0.25, respectively. The input for curve *C* was the sum of the inputs for curves *A* and *B*, and curve *C* itself is the sum of curves *A* and *B*. Curves were computed by numerical integration by Euler's method on the Dartmouth Time-Sharing System. The figure has been redrawn from the original computer-plotted curves. See text for further explanation.

puted responses to these inputs (*i.e.*, the computed "plasma" levels) are shown as curves *A*, *B*, and *C*, respectively, in figure 1. The principle of superposition is reflected in the fact that at every point in time the plasma level defined by curve *C* is exactly equal to the sum of those defined by curves *A* and *B*.

This definition of "addition" is expressed literally in the computer program used to draw the curves of figure 1 (fig. 2). In that program the symbol *A* represents a pair of numbers, namely the initial level *A*(1) in the gastrointestinal tract and the initial plasma level *A*(2) for the first run (curve *A*). Similarly the letters *B* and *C* represent pairs of initial values for the second and third runs, respectively. Line 7 of the program (fig. 2) defines the initial values for the third run as the sums of the corresponding initial values of the first two runs. The functions *FNA*, *FNB*, and *FNC* specify the various injections and infusions into the two compartments; and in line 8 *FNC* is defined as the sum of *FNA* and *FNB*.

A second example is illustrated in figure 3. The figure shows concentration profiles within a plane sheet of homogeneous tissue, resulting from diffusion of a substance applied to one or both surfaces. Curve *A* shows the concentration profile resulting from the

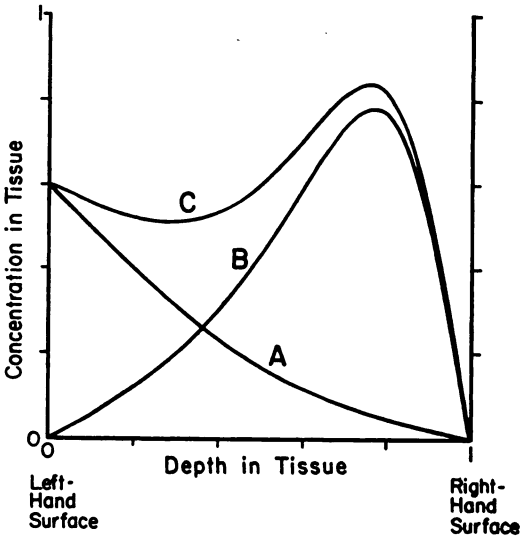

FIG. 3. Additivity of inputs and outputs in a diffusion system. Curves *A*, *B*, and *C* are computed concentration profiles within a plane sheet of homogeneous tissue with a diffusing substance applied to one or both surfaces. The diffusion coefficient is 1. The input for curve *C* was the sum of the inputs for curves *A* and *B*, and curve *C* itself is equal to the sum of curves *A* and *B*. See text for further details.

```
1 LIBRARY "PLOTLIB***:TDI", "FIG1SUB"
2 LET NO = 100
3 DIM Z(150),A(2),B(2),C(2)
4 LET A(2) = 2
5 DEF FNA(J,T) = (1-SGN(ABS(J-2)))*SGN(1-SGN(T-3))*2/NO
6 DEF FNB(J,T) = 3*(1-SGN(J-1))*SGN(2-SGN(ABS(T-1))-SGN(ABS(T-4)))
7 MAT C = A+B
8 DEF FNC(J,T) = FNA(J,T)+FNB(J,T)
9 CALL "AXES":Z()'    DRAW AXES AND MARK SCALES
10 CALL "RUN":Z(),A(),FNA,NO'  RUN WITH INITIAL COND. A() & INPUT FNA
11 CALL "RUN":Z(),B(),FNB,NO'  RUN WITH INITIAL COND. B() & INPUT FNB
12 CALL "RUN":Z(),C(),FNC,NO'  RUN WITH IN. COND. A()+B() & INPUT FNA+FNB
13 END
```

FIG. 2. Computer program for drawing figure 1. The programming language is BASIC. Lines 1 to 3 identify subprogram libraries and set the integration step size and matrix dimensions. Model compartments are numbered 1 (gastrointestinal tract) and 2 (plasma). Line 4 sets the initial "plasma" level at 2 units for input *A*; all other initial levels are automatically set at 0. Line 5 defines input function *A* as a constant infusion into the "plasma" compartment ($J = 2$) during the first three time units. Line 6 defines input function *B* as two pulses (single doses) given into the "gastrointestinal tract" ($J = 1$) at 1 and 4 time units, respectively. Lines 7 and 8 define the initial conditions and the input function for run *C* as the sums of the initial conditions and input functions, respectively, of runs *A* and *B*. Lines 9 to 12 call subprograms to draw and scale the axes and to compute and plot the "plasma" levels for each input in turn.

application of a fixed concentration of 0.6 units to the left-hand surface for 0.11 time units, with the right-hand surface exposed to a solution containing none of the diffusing substance. Curve *B* resulted from the application of concentrations of 1.5 units to the right-hand surface and 0 units to the left-hand surface for 0.1 time units, followed by a partial washing-out by exposure for 0.01 time units to solutions of 0 units' concentration on both surfaces. Curve *C* resulted from the application of both these inputs, *i.e.*, 0.6 units on the left and 1.5 units on the right for 0.1 time units, followed by 0.6 units on the left and 0 units on the right for 0.01 time units. The principle of superposition is reflected in the fact that at every depth within the tissue curve *C* is exactly equal to the sum of curves *A* and *B*. We note that although these particular curves represent only a single point in time (*i.e.*, 0.11 time units after first applying the diffusing substance to the surfaces), the same relationships will hold at any fixed point in time. In short, the output *C* equals the sum of the outputs *A* and *B* at every depth within the tissue and at every time.

As indicated in the above statement of the principle of superposition, the inputs in either of these examples could have been multiplied by arbitrary numbers before being combined. In addition, we could have considered diffusion in three-dimensional space, rather than the one-dimensional diffusion illustrated in figure 3; and in that case we could show that the additivity of output concentrations holds at every point in three-dimensional space and time. However, the simpler examples of figures 1 to 3 suffice to illustrate the basic concepts of addition of complicated and dissimilar inputs, addition of outputs which are functions of time, and addition of outputs which are functions of time and space.

Many authors (11, 23, 26, 49) have implicitly or explicitly recognized the distinction between linear and non-linear systems. What are often referred to nowadays as "dose-dependent" systems are in fact non-linear systems. Concepts of linearity and superposition have from time to time been invoked in pharmacokinetic discussions (25, 32, 50, 53). However, in most of the pharmacokinetic literature the concept of linearity has been related to only one specific class of pharmacokinetic models, namely multicompartment models with constant transfer rate coefficients, uniform concentrations within compartments and instantaneous transfers between compartments. Furthermore, there has been little exploitation of the fact that the principle of superposition applies to an immense variety of more or less complicated inputs (*i.e.*, schedules of dosage and routes of administration).

In this review, we shall try to emphasize the great generality of the concepts of linearity and superposition. We shall stress the fact that these concepts apply not only to conventional multicompartment systems, but also to multicompartment systems with time-dependent rate coefficients, intercompartmental diffusion, incomplete mixing in compartments, or time delays (single-valued or statistically-distributed) in absorption or transport, and also to so-called "discriminating systems" (34), where molecules leave a compartment at a rate dependent on the length of time they have resided in it. Because of this generality, these concepts can often be applied with confidence to actual pharmacokinetic systems about which little of a specific nature is known. We shall also stress the fact that these concepts apply to inputs of great variety and complexity, and that therefore these concepts have potentially wide practical application, not only for predicting responses to multiples of a given dose, but also for analyzing the results of combining different schedules of dosage and multiple routes of administration.

Theory

Pharmacokinetics deals with the time-course of drug concentrations at various points in the body. A pharmacokinetic system therefore consists of a set of anatomical points, together with the drug concentra-

tions at those points. If within a certain anatomical region the drug concentration is everywhere the same, then the points of that region can be lumped together as a "compartment." In that case we can speak of the quantity of drug in the compartment, as well as the concentration. Except for its containing a finite quantity of drug, a compartment can be treated in pharmacokinetic analysis as equivalent to a single anatomical point. Compartments which behave the same pharmacokinetically (*i.e.*, their concentrations are always the same) can be lumped together as a single compartment or "point" for purposes of pharmacokinetic analysis, even if in reality they are anatomically separate. Either the concentration or the quantity of drug in a compartment can be used to characterize it pharmacokinetically; and both are single-valued functions of time. Drug can be injected into a compartment from outside the system, and the rate of injection is likewise a single-valued function of time.

If the drug concentration within a given anatomical region is non-uniform, and if the region cannot be subdivided into compartments with internally uniform concentrations, then the region may be thought of as comprising infinitely many anatomical points, *i.e.*, the points of physical space.[3] We can speak of the quantity of drug within such a region, but this quantity gives no information about drug distribution within the region, and by itself it is therefore insufficient to characterize the region pharmacokinetically. We cannot speak of the quantity of drug at a single point, except as an infinitesimal. The drug concentration in such a region is a function of both time and location within the region, *i.e.*, a "time-space function." We can speak of the overall injection rate into such a region, but this parameter is pharmacokinetically inadequate because it does not specify the distribution of the injected substance within the region. We cannot speak of the injection rate into a single point, except as an infinitesimal. We can, however, conceive of injection as a flux, possibly of non-uniform density, into the region through some boundary surface. The injection rate for such a region is therefore a function of time and location on the boundary surface, *i.e.*, a time-space function.

We shall take the time 0 as the starting point for observing the output of a pharmacokinetic system. The output will then be defined as a set of time-functions or time-space functions on the interval $0 \leq t < \infty$ giving the drug quantities or concentrations in one or more compartments or at one or more anatomical points in the system subsequent to time $t = 0$. The input to the system comprises one or more of the following elements:

1) Initial conditions: A set of drug levels at $t = 0$, either single values or functions over regions of space;

2) Boundary conditions: A set of conditions on the drug levels at the boundaries of spatial regions at times after $t = 0$;

3) Injection rates: A set of time-functions or time-space functions on the interval $0 \leq t < \infty$ giving the rates of injection into the several regions;

4) Applied concentrations: A set of time-functions or time-space functions giving drug levels maintained or enforced in certain compartments; and

5) Past history: A set of time-functions or time-space functions on the interval $-\infty < t \leq 0$ giving drug levels, boundary conditions, applied concentrations and injection rates prior to time $t = 0$.

Our fundamental assumption here is that of a fixed anatomical frame of reference. That is, all those anatomical points of the system which are not lumped into compartments retain their identity at all times; and

[3] Since animals move, we must consider the possibility that such a region is moved or deformed, and we must assume the anatomical points of the region defined in such a way that they retain their identity through all such motions or deformations. In general, this means that the points are not necessarily defined by fixed coordinates in a rigid spatial frame of reference.

any anatomical points that are lumped into compartments stay put, *i.e.*, compartments do not appear or disappear or change their boundaries. Under this assumption both the inputs and the outputs, as defined here, are elements of so-called *linear spaces*. "Linear space" is the somewhat misleading name given to a mathematical abstraction which may bear little resemblance to a "space" in the usual sense. A linear space is in fact merely a *set* of things with the special property that they can be added together or multiplied by a number and the resulting sum or product will also be a member of the same set.[4] An example of a linear space is the set of all single-valued functions of time on the interval $t_1 \leq t \leq t_2$. The sum of any two single-valued functions of time on this interval is itself a single-valued function of time on the same interval, and is itself therefore a member of the set. Similarly, multiplication of any function by a number gives another function on the same interval. Another example of a linear space is the set of all functions of time and space on some time interval and space region. These can be added together or multiplied by a number to give new functions on the same time interval and space region. Still another example is the set of all pairs of numbers $\{a, b\}$. The sum of two such pairs $\{a_1, b_1\}$ and $\{a_2, b_2\}$ can be defined as the pair $\{(a_1 + a_2), (b_1 + b_2)\}$; and the product of the pair $\{a, b\}$ by the number α can be defined as the pair $\{\alpha a, \alpha b\}$. More generally, sets of n elements of linear spaces may themselves form a linear space, if the elements within a set can be arranged in some sort of order (*e.g.*, if they are elements of a vector or a matrix[5]). In that case, the i-th element of a set can be multiplied by a fixed number, or added to the i-th element of a second set, to give the i-th element of a new set, for $i = 1$ to n.

The concept of a linear space forms the basis for the most general interpretation of the principle of superposition, as it has been stated above. In order for the principle to be meaningful, it is only necessary that the inputs ξ be elements of a linear space, and that the outputs ζ also be elements of a linear space, not necessarily the same one as the inputs. This means that the principle of superposition can be meaningfully stated in terms of inputs and outputs that are numbers, time-functions, time-space functions, discontinuous functions, sets of numbers, sets of functions, mixed sets containing both numbers and functions, *etc.* Since, as we have shown above, a wide variety of pharmacokinetic inputs and outputs are elements of linear spaces, the principle of superposition can be applied to a correspondingly wide range of pharmacokinetic systems. This we shall now illustrate with some specific examples.

The simplest case is a one-compartment system with drug administered by injection, which obeys the equation

$$\frac{dy}{dt} = -ky + x. \tag{1}$$

Here x and y are functions of time on the interval $0 \leq t < \infty$, representing respectively the injection rate and the quantity of drug in the system; and k is a first-order rate coefficient, which can be either constant or varying with time. (If k varies with time, the product ky of the two time-functions k and y is defined in the usual way as a time-function whose value at any time is equal to the product of the values of k and y at that same time.) The input ξ to this system comprises the injection rate x and the initial value $y(0)$, and may be written as $\xi = \{x, y(0)\}$. Each such input uniquely determines an output ζ $(= y)$. Addition of two inputs ξ_1 and ξ_2 is defined as follows, for any numbers α and β :

$$\begin{aligned} &\alpha\xi_1 + \beta\xi_2 \\ &\quad = \alpha\{x_1, y_1(0)\} + \beta\{x_2, y_2(0)\} \\ &\quad = \{[\alpha x_1 + \beta x_2], [\alpha y_1(0) + \beta y_2(0)]\}. \end{aligned} \tag{2}$$

[4] For a more formal definition of a linear space, see Birkhoff and MacLane (3), p. 152, or Collatz (6) pp. 1–2.

[5] See the APPENDIX for an explanation of these terms, if necessary.

It is important to remember that, whereas an ordinary equation defines a *number* as its solution, a differential equation like equation (1) defines not a number but a *function*. It is the entire function y, and not any single value, which must be thought of as the solution to equation (1). Accordingly, any given input function x_1, together with an initial condition, defines a certain output function y_1. Similarly, an input x_2 with an initial condition defines an output function y_2. The functions y_1 and y_2 satisfy the equations

$$\frac{dy_1}{dt} = -ky_1 + x_1 \qquad (3a)$$

and

$$\frac{dy_2}{dt} = -ky_2 + x_2, \qquad (3b)$$

respectively. In the first case, the output ζ_1 is the function y_1, and the input ξ_1 comprises the function x_1 and the initial condition $y_1(0)$, *i.e.*, $\xi_1 = \{x_1, y_1(0)\}$. Similarly, in the second case, $\zeta_2 = y_2$ and $\xi_2 = \{x_2, y_2(0)\}$.

If we now multiply equations (3a) and (3b) by α and β respectively, and then add them together, we obtain

$$\frac{d(\alpha y_1 + \beta y_2)}{dt} = -k(\alpha y_1 + \beta y_2) + (\alpha x_1 + \beta x_2). \qquad (4)$$

This is a differential equation defining a new function, namely the function $\alpha y_1 + \beta y_2$. In terms of this function, and the new input function $\alpha x_1 + \beta x_2$, equation (4) is of exactly the same form as equation (1). It states, therefore, that the function $\alpha y_1 + \beta y_2$ is a solution of equation (1), with the injection rate $x = \alpha x_1 + \beta x_2$. Moreover, the initial value of the function $\alpha y_1 + \beta y_2$ is obviously $\alpha y_1(0) + \beta y_2(0)$. Therefore $\alpha y_1 + \beta y_2$ is *the* solution or output associated with the input $\{[\alpha x_1 + \beta x_2], [\alpha y_1(0) + \beta y_2(0)]\}$. This input is $\alpha\xi_1 + \beta\xi_2$, according to equation (2). Therefore, for the system defined by equation (1), $\alpha\zeta_1 + \beta\zeta_2$ (*i.e.*, $\alpha y_1 + \beta y_2$) is the output for the input $\alpha\xi_1 + \beta\xi_2$; and therefore the system obeys the principle of superposition.

When we turn to multicompartment systems, we open the door to infinite mathematical complexities; but all these can be made to disappear by using a completely general analysis based on matrix algebra (see APPENDIX). A linear multicompartment system (that is, one in which all unidirectional compartmental effluxes obey first-order kinetics) obeys equation (1), provided y and x are interpreted as vector functions of time ($t \geq 0$) giving respectively the drug quantities in, and the injection rates into, the various compartments. Then k is a matrix of rate coefficients (or a matrix function of time, if the rate coefficients vary with time), and $y(0)$ is the vector of initial compartment contents or concentrations. With these definitions, the argument of equations (3, 4) can be followed through again to show that linear multicompartment systems obey the principle of superposition.

Sometimes a drug, instead of being injected into the system, is maintained in a controlled concentration in a source compartment, from which it passes into the system by a first-order process. This is the situation, for example, when an inhalation anesthetic is administered in a controlled concentration in the inspired gas mixture. In such cases, x is a drug concentration (or a vector of drug concentrations) in the compartment (or compartments) where the drug concentration is controlled, and the system obeys an equation of the form

$$\frac{dy}{dt} = -ky + bx, \qquad (5)$$

where b is a first-order rate coefficient or a vector or matrix of first-order rate coefficients. Here again, an argument analogous to that of equations (3, 4) shows that these systems obey the principle of superposition.

A system with diffusion may be approximated by a linear multicompartment system with a very large number of very small compartments, the approximation being made to any desired degree of accuracy by increasing the number and decreasing the size of compartments. Alternatively, one may take the drug concentration y as a function on

three-dimensional space and time, and employ the methods of vector analysis (7, 28). For generality, we assume the diffusion medium to be anisotropic; then, for rectangular coordinates s_1, s_2, and s_3,

$$\begin{aligned}\frac{dy}{dt} = D_{11}\frac{\partial^2 y}{\partial s_1^2} + D_{22}\frac{\partial^2 y}{\partial s_2^2} + D_{33}\frac{\partial^2 y}{\partial s_3^2} \\ + (D_{23} + D_{32})\frac{\partial^2 y}{\partial s_2 \partial s_3} \\ + (D_{31} + D_{13})\frac{\partial^2 y}{\partial s_1 \partial s_3} \\ + (D_{12} + D_{21})\frac{\partial^2 y}{\partial s_1 \partial s_2} \\ = \text{div } (D \text{ grad } y),\end{aligned} \tag{6}$$

where the D_{ij} are the elements of the matrix of components of the tensor D. For an isotropic medium, $D_{11} = D_{22} = D_{33}$, the other D_{ij} are zero, and the tensor D is simply the diffusion coefficient.

With diffusion systems there is usually no input other than the initial and boundary conditions. The simplest case is where the boundary conditions $y(B)$ simply specify the value of y at all points and at all times on the boundary B. Since

$$\begin{aligned}\alpha \text{ div } (D \text{ grad } y_1) + \beta \text{ div } (D \text{ grad } y_2) \\ = \text{div } (D \text{ grad } [\alpha y_1 + \beta y_2]),\end{aligned} \tag{7}$$

we can show by adding equations, as we have done to obtain equation (4) from equations (3a) and (3b), that if y_1 and y_2 are solutions to equation (6) with initial and boundary conditions $\{y_1(0), y_1(B)\}$ and $\{y_2(0), y_2(B)\}$, respectively, then $\alpha y_1 + \beta y_2$ is the solution with initial and boundary conditions $\{[\alpha y_1(0) + \beta y_2(0)], [\alpha y_1(B) + \beta y_2(B)]\}$; *i.e.*, the principle of superposition holds.

A more complicated case would be where the boundary conditions are of the form

$$[D_s \text{ grad } y]_B = k[y(B) - x(B)], \tag{8}$$

where $[D_s \text{ grad } y]_B$ is the flux component at the boundary in a direction normal to the boundary surface, k is the permeability coefficient of the boundary surface, and $y(B)$ and $x(B)$ are the concentrations just inside and just outside the boundary surface, respectively. The permeability coefficient k may differ at different points on the boundary and at different times, as may the external concentration $x(B)$ and the tensor D_s. Equation (8) represents a boundary surface of limited permeability; and the special case where $k = 0$ represents a reflecting (impenetrable) boundary. The input to the system comprises the initial conditions $y(0)$ and the external concentration x (of which only the boundary function $x(B)$ actually contributes input). It is easily verified, then, that if y_1 and y_2 satisfy equations (6) and (8) with inputs $\{y_1(0), x_1\}$ and $\{y_2(0), x_2\}$ respectively then $\alpha y_1 + \beta y_2$ is the function which satisfies equations (6) and (8) with the input $\{[\alpha y_1(0) + \beta y_2(0)], [\alpha x_1 + \beta x_2]\}$.

A system with delayed transport, such as transport in the circulation (29, 43, 52) obeys an equation of the form

$$\begin{aligned}\frac{dy}{dt} = -ky + \int_{-\infty}^{t} k\phi_1(t-\theta)y(\theta)\,d\theta \\ + \int_{-\infty}^{t} \phi_2(t-\theta)x(\theta)\,d\theta,\end{aligned} \tag{9}$$

where y is the drug level at some point of observation, x is the rate of injection (at some other point in the system), $y(\theta)$ and $x(\theta)$ are respectively the values of these functions at the time θ, ϕ_2 is a statistical density function[6] of transit times from the site of injection to the point of observation, ϕ_1 is the sum of statistical density functions

[6] Although we refer to the functions ϕ_2 and ψ [equations (9–12)] as statistical density functions, they differ from ordinary statistical density functions in that their integrals from 0 to ∞ may be less than 1, *i.e.*, there may be losses in transmission, or there may be permanent retention of part of the input to a discriminating system. The function ϕ_1 has the structure of a renewal density function (52) based on a single-passage recirculation time density function whose integral from 0 to ∞ may be less than 1 because of losses from the recirculation pathway.

of transit times for one, two, *etc.*, recirculations, and $\phi_1(t-\theta)d\theta$ and $\phi_2(t-\theta)d\theta$ are respectively the statistical frequencies of recirculation times and transit times between $t-\theta$ and $t-\theta+d\theta$.

We have chosen to consider as the "response" of a pharmacokinetic system only that portion of the output function where $t \geq 0$. We therefore partition the function y in equation (9) into a part prior to $t=0$ (the "past history"), which will be denoted by h, and a part on the interval $t \geq 0$, to which we will now restrict the notation y. We make a corresponding partition of the first integral in equation (9) into an integral from $-\infty$ to 0 and an integral from 0 to t, and obtain

$$\begin{aligned}\frac{dy}{dt} = -ky &+ \int_0^t k\phi_1(t-\theta)y(\theta)\,d\theta \\ &+ \int_{-\infty}^0 k\phi_1(t-\theta)h(\theta)\,d\theta \qquad (10) \\ &+ \int_{-\infty}^t \phi_2(t-\theta)x(\theta)\,d\theta,\end{aligned}$$

where y is a function of time on the interval $0 \leq t < \infty$. The input ξ then comprises the "past history" h on the interval $-\infty < t < 0$, the function x on the interval $-\infty < t < \infty$, and the initial condition $y(0)$. We define addition of such inputs as follows, for any numbers α and β:

$$\begin{aligned}&\alpha\xi_1 + \beta\xi_2 \\ &= \alpha\{h_1, x_1, y_1(0)\} + \beta\{h_2, x_2, y_2(0)\} \\ &= \{[\alpha h_1 + \beta h_2], [\alpha x_1 + \beta x_2], \\ &\quad [\alpha y_1(0) + \beta y_2(0)]\}.\end{aligned} \qquad (11)$$

It is now easily verified that if y_1 and y_2 are the outputs for the inputs ξ_1 and ξ_2, respectively, then $\alpha y_1 + \beta y_2$ is the output for $\alpha\xi_1 + \beta\xi_2$.

A "discriminating system" obeys an equation of the form

$$\frac{dy}{dt} = x - \int_{-\infty}^t \psi(t-\theta)x(\theta)\,d\theta, \qquad (12)$$

where ψ is the statistical density function[6] of survival times. Here the input is $\{x, y(0)\}$; and the principle of superposition is easily verified by the same general method used in the preceding cases.

Systems combining two or more of the foregoing features, for example multicompartment systems with statistically distributed intercompartmental delay times, or multicompartment systems with intracompartment diffusion gradients due to incomplete mixing, can also be shown to obey the principle of superposition. Suppose we have two linear systems, each of whose outputs forms part of the input to the other. Let y and y' be the respective outputs, and let $\{y', x\}$ and $\{y, x'\}$ be the respective inputs. (In the present context, x and x' are assumed to comprise all inputs other than y and y', including initial and boundary conditions, past history, *etc.*) Assuming that the overall system is physically realizable and determinate, there is a unique solution $\{y, y'\}$ for any given external input $\{x, x'\}$. Let $\{y_1, y_1'\}$ and $\{y_2, y_2'\}$ be the solutions for the external inputs $\{x_1, x_1'\}$ and $\{x_2, x_2'\}$, respectively. Then from the linearity of the first subsystem it follows that $\alpha y_1 + \beta y_2$ is the solution for the input $\{[\alpha y_1' + \beta y_2'], [\alpha x_1 + \beta x_2]\}$, for any numbers α and β. Similarly, for the second subsystem, $\alpha y_1' + \beta y_2'$ is the solution for the input $\{[\alpha y_1 + \beta y_2], [\alpha x_1' + \beta x_2']\}$. Therefore for the overall system, $\{[\alpha y_1 + \beta y_2], [\alpha y_1' + \beta y_2']\}$ is the output for the external input $\{[\alpha x_1 + \beta x_2], [\alpha x_1' + \beta x_2']\}$. Since any such combination of two linear systems is thus linear, it follows that any such combination of any number of linear systems is linear.

The concept of linearity actually implies two distinct properties, namely *homogeneity* and *additivity* (*cf.* 56, pp. 132–137). A homogeneous system obeys the principle that if ζ is the output for the input ξ then, for any number α, $\alpha\zeta$ is the output for the input $\alpha\xi$; but it does not necessarily obey the superposition principle as regards *addition* of inputs. Of course, if the inputs are simple numbers, there is no distinction between inputting the sum of two numbers

(*e.g.*, 1.32 + 2.64 = 3.96) and inputting the product of one of them by an appropriate α (*e.g.*, 1.32 × 3 = 3.96). This is not so, however, for the more general types of input we are considering, *i.e.*, elements of linear spaces. For example, the sum of the pairs (1, 2) and (1, 0) is the pair (2, 2), which cannot be obtained by multiplying either of the original two pairs by any number. If a system is such that when inputs are added the outputs add, it is said to have the property of additivity. For all practical purposes, a physical system which is additive is also homogeneous and hence linear; but a system which is homogeneous may not be additive.

An artificial example of a homogeneous but non-linear system is the system

$$\frac{dz_1}{dt} = -k_{11}z_1 + k_{12}\sqrt{z_1 z_2} + x_1 \quad (13)$$

$$\frac{dz_2}{dt} = k_{21}\sqrt{z_1 z_2} - k_{22}z_2 + x_2 \quad (14)$$

Multiplication by α shows that if z (the column vector of z_1 and z_2) is the solution for $\{x, z(0)\}$ then αz is the solution for $\{\alpha x, \alpha z(0)\}$. On the other hand, if z and z' are solutions for $\{x, z(0)\}$ and $\{x', z'(0)\}$, respectively, and we multiply by numbers α and β and add the differential equations as we did with equations (3a) and (3b) to obtain equation (4), we find that $\alpha z + \beta z'$ is not a solution for $\{\alpha x + \beta x', \alpha z(0) + \beta z'(0)\}$.

In point of fact, it seems rather unlikely that the usual pharmacokinetic mechanisms will ever give rise to a system which is homogeneous but non-linear; however, the possibility may have to be considered in some cases.

It should be noted that, while in some respects we have been very general in our description of pharmacokinetic inputs and outputs, in other respects we have laid down some very specific requirements. A pharmacokinetic input or output must be defined for a fixed set of anatomical points and a fixed set of points in time, in order for us to use it in the principle of superposition. Therefore, we can use as the output, for example, the plasma level at some fixed time t; but we cannot use the peak plasma level, because this may occur at different times with different routes and schedules of administration. In a linear pharmacokinetic system the peak drug level at any point will be directly proportional to the dose, as long as the route and schedule of administration are unchanged, *i.e.*, as long as the input is unchanged except for multiplication by a dosage factor α. However, since peak levels for different routes and schedules of administration will not necessarily coincide in time, the peak level resulting from the sum of two inputs will not necessarily equal the sum of the peak levels resulting from those inputs applied separately, but may be less. If they were to be viewed as output, then, peak drug levels in a linear system might be said to have the property of homogeneity but not that of additivity.

Discussion

1. Concept of Linearity in Analysis of Pharmacokinetic Data

In the analysis of data on pharmacokinetic systems, the principle of superposition provides a simple test for detecting the presence of non-linear processes, *i.e.*, processes which do not obey first-order kinetics. Examples of such processes are metabolic transformations with Michaelis-Menten kinetics, active renal tubular transport with a characteristic transport maximum, and the binding of drug to saturable sites on plasma proteins or in tissues. In addition, there are processes whose rates are affected by pharmacological actions of the drug itself. These would include processes such as metabolism of the drug by an enzyme which is induced by the drug, or distribution of the drug by tissue blood flows which are affected by the drug. On the other hand, time-dependent processes (*i.e.*, processes in which the rate coefficients or other parameters are functions of time) do not introduce non-linearity, though they may produce pharma-

cokinetic behavior that cannot be simulated by any multicompartment model with constant rate coefficients.

In some respects, linear and non-linear systems may closely resemble one another. A few years ago, Wagner (48) showed that a linear four-compartment model could account for a period of apparent zero-order elimination kinetics, such as that which several investigators had reported for salicylates. Responding to Wagner's suggestion, however, Levy (25) and Cummings and Martin (9) cited the evidence that the rate of salicylate elimination does not increase in proportion to the dose, or in proportion to the total quantity present in the body, and that the fraction of the dose which is eliminated as salicyluric acid is less after large doses than after small doses. We draw attention now to the fact that these arguments indicate failure of the principle of superposition, and are therefore absolutely conclusive in ruling out not only Wagner's particular multicompartment model but also all other linear models, including those that are not conventional multicompartment models at all.

To be more explicit, let y be a vector function of time, giving the quantities of drug in all compartments (including the urine) as functions of time. The input—a single dose at time $t = 0$—is most conveniently represented by an initial condition vector $y(0)$, whose components are all zero except for the one corresponding to the compartment where the drug is administered (the gastrointestinal tract). There is no other input. Increasing the dose by a factor α corresponds to multiplying the input by α. According to the principle of superposition, the output would then be changed from y to αy. It follows from this that the quantity of drug which has been eliminated (*i.e.*, is present in the urine compartment) at any time t would be directly proportional to the dose, and hence that the rate of elimination at any time would be directly proportional to the dose. Moreover, the quantity eliminated in any particular *form* (*e.g.*, salicyluric acid) would be proportional to the dose, so that the *proportion* eliminated in any particular form—that is, the ratio of the quantity eliminated in that form to the sum of the quantities eliminated in all forms—would be dose-independent. Since, as Levy and Cummings and Martin pointed out, these rules are not followed by salicylates, all linear pharmacokinetic models for salicylates are ruled out.

Nevertheless, Wagner had a valid point, namely that a linear model can sometimes be made to simulate the time-course of behavior of a non-linear system. The same point was recently demonstrated again by DiSanto and Wagner (12). It is therefore of considerable importance to determine the pharmacokinetic behavior at different doses, so as to test the principle of superposition. To do this, it is not necessary to go through the tedious procedure carried out by DiSanto and Wagner, fitting a linear model to the data at each dose to determine whether the parameters must be changed with the dose; nor is it necessary to determine the area under the plasma concentration-time curve (49, pp. 242–246), or the apparent biological half-life, as a function of the dose. One need only plot out graphically the observed plasma levels (or cumulative urinary excretion), divided by the dose given (*cf.* 50, p. 20). If the system is linear, the curves for different doses will superimpose. This simple procedure allows one to detect failure of superposition, and hence non-linearity, even before taking the first steps toward formulating a detailed pharmacokinetic model.

Figure 4 illustrates a simple superposition plot of curves computed for the non-linear model of DiSanto and Wagner (12). The failure of superposition is obvious without further analysis. Data can also be tested for linearity by an alternative plotting method, used by Krüger-Thiemer (23). This method is illustrated in figure 5, where the logarithm of the plasma level is plotted against time for a non-linear model of salicylate kinetics (27) and for a linear model. In this type of plot, a change of dose shifts the response

curve up or down in a parallel fashion, if the system is linear. This type of plot gives a good display of the data for linear systems, but it is difficult to judge visually whether the parallelism is exact, and one cannot tell immediately whether the vertical shifts are the right distance for superposition. For the detection of non-linearity, therefore, one should choose a plotting method, such as that of figure 4, which makes the data at different doses superimpose for linear systems. Wagner (50, p. 20) has already suggested the simple superposition plot as a test for linearity; but he has validated it only in terms of a conventional two-compartment open model. Here we

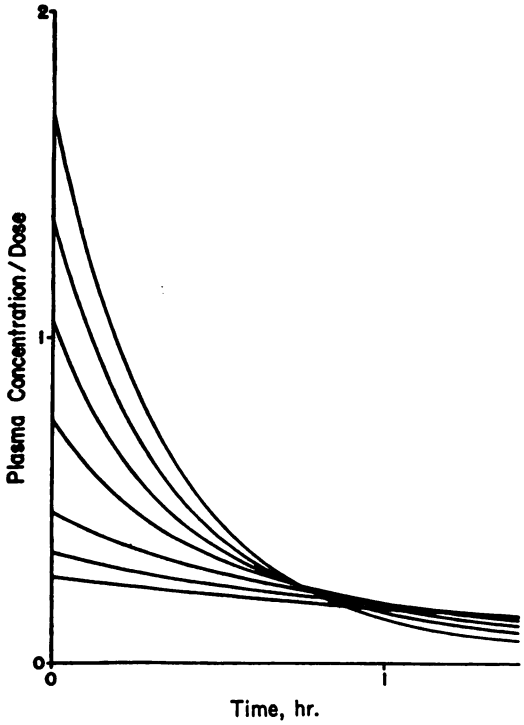

FIG. 4. Superposition plot for the non-linear pharmacokinetic model of DiSanto and Wagner (12). The model has the equation

$$\frac{d}{dt}[C + AC/(B + C)] = -KC,$$

or

$$\frac{dC}{dt} = -\frac{KC(B + C)^2}{AB + (B + C)^2},$$

where C is the plasma concentration, and A, B, and K are parameters having the values 10, 1, and 2.75, respectively. The curves in this figure (and in figs. 5 and 7) were computed by numerical integration by a fourth-order Runge-Kutta method on the Dartmouth Time-Sharing System. On the left-hand side of the figure, the curves lie in order of dose, that for the highest dose being at the top. The dose D is related to the initial plasma concentration C_0 as follows:

$$D = V[C_0 + AC_0/(B + C_0)],$$

where V (= 0.5) is the volume of distribution. Doses for the seven curves were 1.917, 3.00, 4.33, 6.67, 9.55, 14.76, and 29.90.

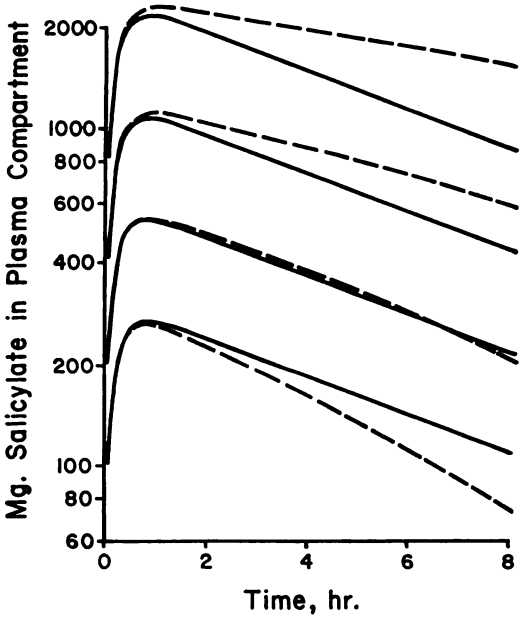

FIG. 5. Comparison of linear and non-linear models for salicylate excretion. Dashed curves were computed by numerical integration of equations (4–14) of Levy *et al.* (27). These equations describe a non-linear model with two saturable metabolic transformations. Doses (from bottom) were 300, 600, 1200, and 2400 mg. Solid curves were computed for corresponding doses in a linear model with three compartments (gastrointestinal tract, plasma, and urine), with parameter values adjusted to give a good fit to the model of Levy *et al.* at a dose of 600 mg (absorption rate constant 4.2; excretion rate constant 0.13). At the dose of 600 mg the fit is quite good for 8 hr (though not for much longer); but changes in dose cause the two models to diverge. The curve for the linear model is shifted up or down in parallel fashion, whereas for the non-linear model the shifts are non-parallel.

stress that the concept of linearity is much broader, and can apply to systems with time-dependent rate coefficients, time-delays, diffusion gradients, and an unrestricted number and arrangement of compartments. Figure 6 shows an artificial model containing a time-dependent metabolic transformation process and an "enterohepatic circulation" with a fixed time-delay; and figure 7 shows computed curves of plasma levels in this system in response to oral doses. The system can be seen to obey the principle of superposition. In real life, the kinetics of enterohepatic circulation are probably somewhat more complicated than a simple fixed time delay. Jusko and Levy (20) have suggested that the double peak in urinary excretion rates of riboflavin reflects enterohepatic circulation, with accumulation of the drug in the gallbladder over a period of time, abrupt emptying of the gallbladder at mealtime, and an ensuing period of rapid absorption of the riboflavin thus liberated into the gastrointestinal tract. Even such a complicated system as this, however, could be effectively tested for linearity by a simple superposition plot, provided the schedules of drug administration were carefully controlled in relation to a fixed schedule of meals.

Figures 8 to 10 illustrate superposition plots applied to real pharmacokinetic data taken from the literature. Figure 8 shows data on plasma levels of *d*-tubocurarine in the dog after each of two intravenous doses (5). With the possible exception of two points we could conclude that these data show superposition and that the pharmacokinetics of *d*-tubocurarine are linear. However, the range of doses is rather narrow, and perhaps a wider range of doses would turn up more definite departures from superposition. Figure 9 shows that in fact the pharmacokinetics of *d*-tubocurarine are *not* strictly linear, because the renal levels attained with different doses do not satisfy the principle of superposition. This effect is apparently too small to affect the plasma levels noticeably. (It should be noted that the investigators who did this work did not overlook or fail to appreciate this finding, although they did not do superposition plots.)

Figure 10 shows a superposition plot of data on three intravenous doses of bishydroxycoumarin in man (30). From it one can see immediately that (as has long been known) the rate of elimination does not increase in proportion to the dose or the plasma level. One can also see that with increasing dose there is a slight increase in the early apparent volume of distribution (*i.e.*, a downward displacement of the curve of plasma concentration/dose). This would suggest that binding to saturable sites on plasma protein may be affecting the kinetics; but in these experiments the larger doses were infused over a longer

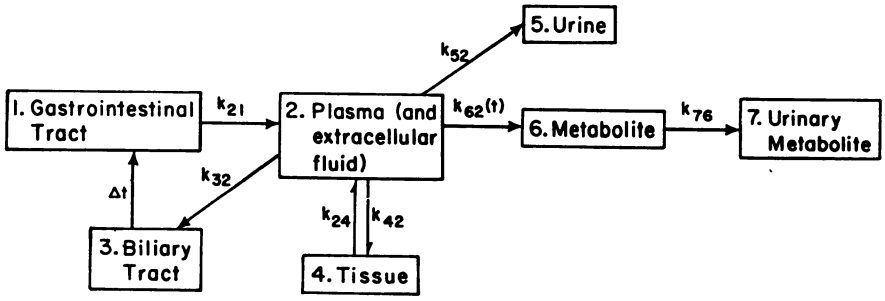

FIG. 6. Schematic diagram of an artificial pharmacokinetic model containing a time-dependent metabolic process and an "enterohepatic circulation" with a fixed time-delay. Parameter values were as follows: $k_{21} = 4\ hr^{-1}$, $k_{32} = 0.5\ hr^{-1}$, $\Delta t = 0.5$ hr, $k_{42} = 0.5\ hr^{-1}$, $k_{24} = 1\ hr^{-1}$, $k_{52} = 0.02\ hr^{-1}$, $k_{62} = 0.2 + 0.002t^2\ hr^{-1}$ (t = time in hr), $k_{76} = 0.3\ hr^{-1}$, plasma + extracellular fluid volume 0.2 liter/kg body weight.

time period (up to 30 min), and this may account for some of the discrepancy. The investigators here analyzed these data in terms of a three-compartment model, and found that just one of the several parameters changed consistently with the dose. The attempt to thus localize the non-linearity is certainly worth while; but the conclusion is necessarily subject to whatever doubts we may have about the correctness of the three-compartment model itself; and the correctness of a pharmacokinetic model is hard to be sure of when only observations on a single compartment are available (35, 54). On the other hand, the dose-dependency of the elimination rate and (possibly) the apparent volume of distribution are important empirical facts about bishydroxycoumarin that do not depend upon any particular compartmental model; and these facts are easily brought out by a superposition plot.

The test of simply increasing or decreasing the dose is of course a test of homogeneity rather than linearity. As pointed out above, this is probably sufficient proof of linearity for most pharmacokinetic systems; but in some cases it may be desirable to test further for additivity. This might be done by using either multiple dosage sched-

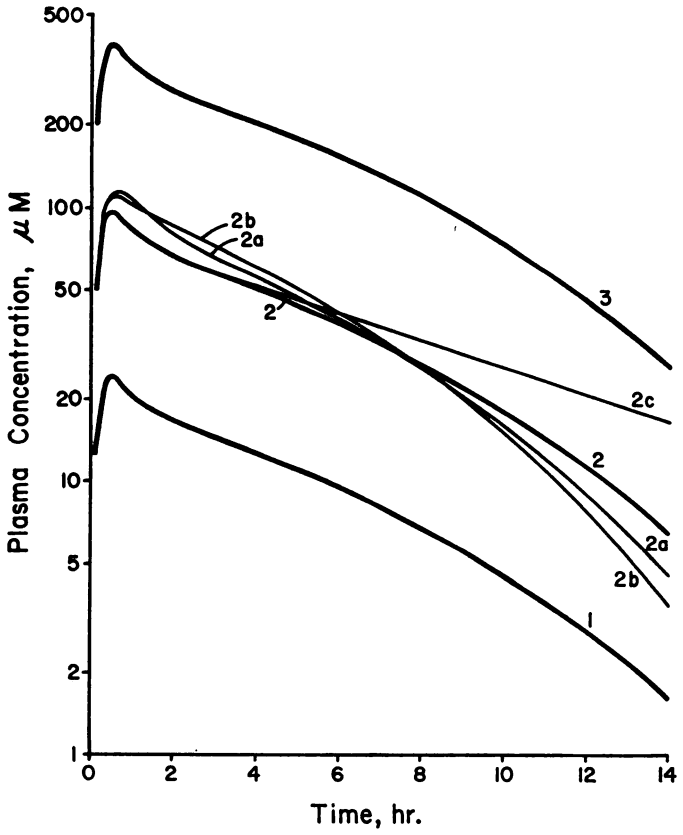

FIG. 7. Computed responses (plasma level) of the model shown in figure 6. Curves 1, 2, and 3 were computed for doses of 8, 32, and 128 μmol/kg, respectively. On the logarithmic ordinate scale these curves are parallel and shifted by a distance exactly equal to the logarithm of 4 (the ratio of the doses). If the computed levels for each curve were divided by the dose, these three curves would superimpose. Curves 2*a*, 2*b*, and 2*c* were computed to illustrate the roles played by each of the special features of the model, by deleting each feature in turn. For curve 2*a*, the "enterohepatic circulation" was deleted. For curve 2*b*, the tissue compartment (no. 4) was deleted. For curve 2*c*, the time-dependent metabolic transformation was deleted.

ules or multiple routes of administration. To explore superposition with different routes of administration, for example, one might study the pharmacokinetics of intravenous and intramuscular doses, given first separately and then together. To test superposition with different dosage schedules, one might study in the same way the

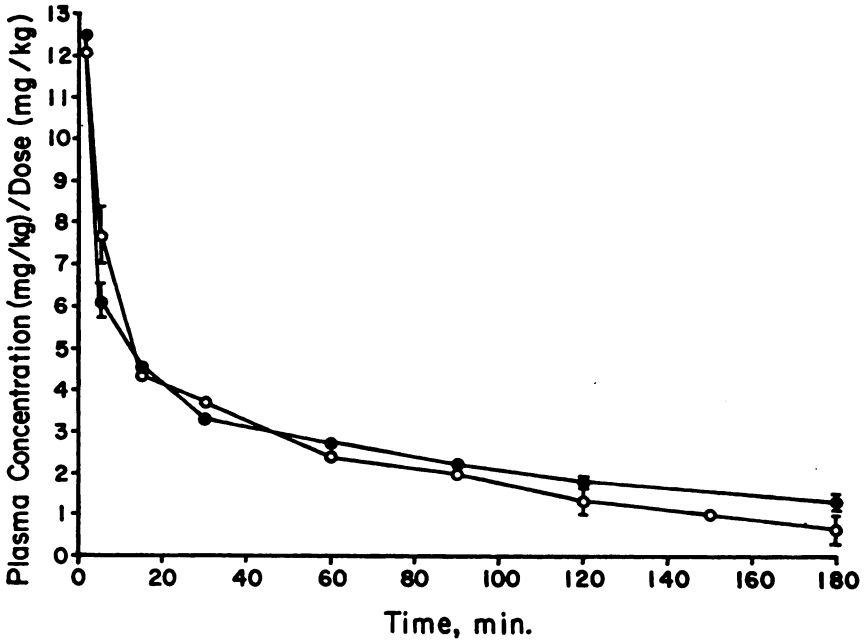

FIG. 8. Superposition plot of plasma concentrations of *d*-tubocurarine in the dog after each of two intravenous doses [data of Cohen *et al.* (5)]. O, 0.3 mg/kg; ●, 1.0 mg/kg.

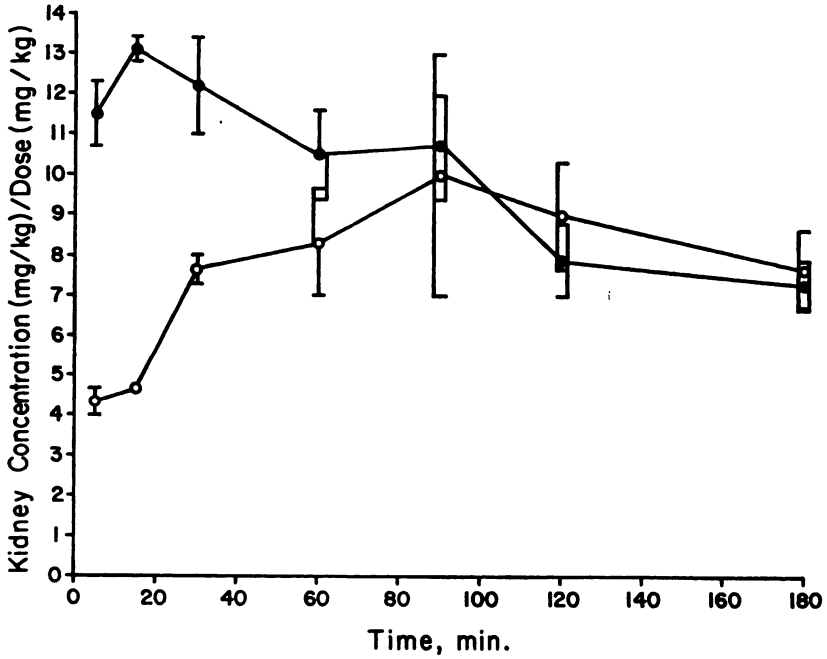

FIG. 9. Superposition plot of renal concentrations of *d*-tubocurarine in the dog after each of two intravenous doses [data of Cohen *et al.* (5)]. O, 0.3 mg/kg; ●, 1.0 mg/kg.

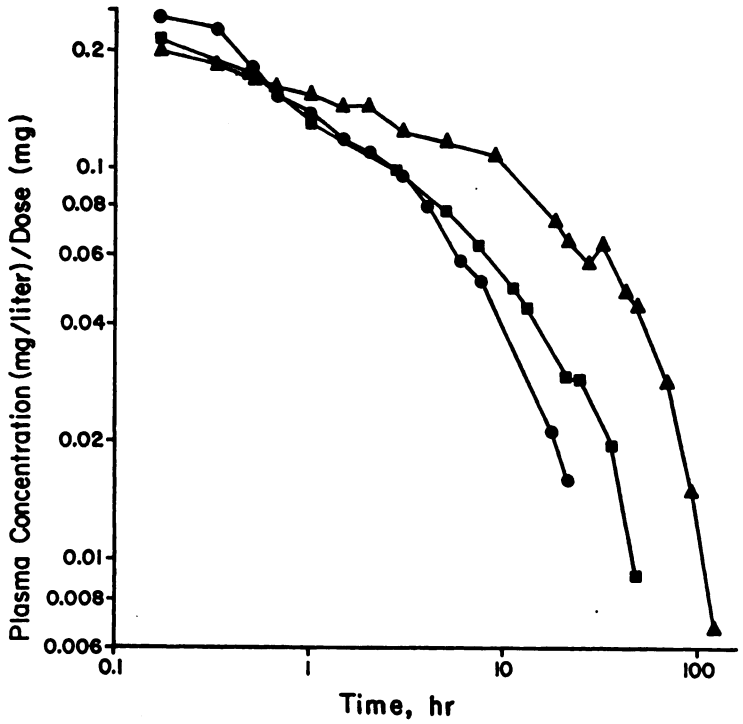

FIG. 10. Superposition plot of plasma bishydroxycoumarin concentrations (mg/liter) in man after each of three intravenous doses [data of Nagashima *et al.* (30)]. ●, 150 mg; ■, 286 mg; ▲, 600 mg.

effect of combining single intravenous doses with constant intravenous infusions, or the effects of a series of doses spaced in time, as compared to the effects of a single dose. In this last type of study, one must be careful not to be led astray by temporal variations in the pharmacokinetic parameters. If the pharmacokinetics are not the same at noon as at 8 A.M., for example, then one must compare the effect of giving both an 8 A.M. and a noon dose with the sum of the effects of the 8 A.M. and noon doses given separately, and not with the sum of two 8 A.M. responses staggered in time.

2. *Application of the Concept of Linearity in Theoretical Studies*

a. Drug Accumulation on Continuous Infusion or Repeated Dosage. The principles of drug accumulation on continuous infusion or repeated dosage have been discussed for one-compartment systems by many authors (14, 36, 44, 55); and they have also been extended to more general types of multicompartment system (37, 51), for which the only assumptions are that the first-order rate coefficients are constant and that all elimination takes place directly from the "central" compartment. Some of these principles of drug accumulation, however, can easily be shown to hold for *all* linear systems, including those with time-delays, time-varying rate coefficients, and elimination from peripheral compartments. For example, since multiplication of the input by a factor α multiplies the output by the same factor, it follows that if the drug level tends asymptotically to a fixed level with constant drug infusion[7] then that asymptotic level is directly proportional to the infusion rate. Similarly, if asymptotic maximum, mini-

[7] Necessary and sufficient conditions for such asymptotic behavior, which might be termed "pharmacokinetic stability," have been elucidated for linear multicompartment systems (13, 45) but not for all possible types of pharmacokinetic system.

mum and mean levels are attained on prolonged repetition of the same dose, then these levels are directly proportional to the dose. The same holds for a repeated *pattern* of doses, *e.g.*, a daily pattern consisting of oral doses at 8 A.M., 12 noon, 4 P.M. and 8 P.M.: if asymptotic maximum, minimum and mean levels are attained, they will be directly proportional to the dose. The dosage pattern may be even more complicated, involving two or more different doses, drug preparations or routes of administration. To concoct an elaborate example, suppose a subject is taking at 8 A.M. a 10-mg oral dose of a standard preparation and a 40-mg oral dose of a sustained-release preparation, then at 4 P.M. another 40 mg of the sustained-release preparation, and at 8 A.M., 12 noon, 4 P.M. and 8 P.M. a metered dose of 2 mg by inhalation from a nebulizer. If all doses in such a regimen are multiplied by some common factor α, without changing the routes or schedule of administration, then in the resulting asymptotic daily pattern of plasma and tissue concentrations the concentrations will all be multiplied by the factor α.

These conclusions hold for all linear systems. If, in addition to being linear, a system is "time-invariant" (*i.e.*, it contains no parameters which depend directly upon the time t), or if all time-dependent parameters are periodic with the same period, then additional generalizations can be made. First we observe that, for any linear system, if a dose is given repeatedly at an interval τ, the cumulative response during the interval from the n-th to the $(n + 1)$-th repetition is equal to the sum of the first segment of length τ of the response to the n-th dose, plus the second segment of length τ of the response to the $(n - 1)$-th dose, *etc.* (*cf.* fig. 11). If now the system contains no time-dependent parameters, or if all time-dependent parameters vary periodically with the same period as the dosage cycle, then the responses to all

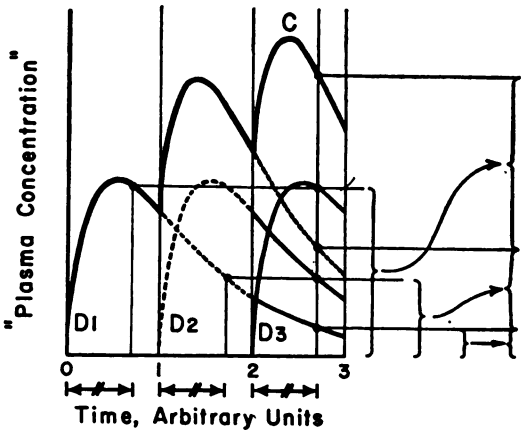

FIG. 11. Repetitive dosage: The cumulative response as the summation of segments of individual responses. Curves *D*1, *D*2, and *D*3 show the responses to a single dose given at times 0, 1 or 2, respectively. Curve *C* shows the cumulative response to all three doses, assuming the system is linear. Consider the curves *D*1, *D*2, *D*3, and *C* to be divided into segments 1 time unit in duration. Then the segment of the cumulative response (curve *C*) after the third dose is the sum of the first segment of curve *D*3, plus the second segment of curve *D*2, plus the third segment of curve *D*1. In this particular case, the responses *D*1, *D*2, and *D*3 are identical except for a time-shift; and therefore the segment of the cumulative response after the third dose can alternatively be equated to the sum of the first three segments of curve *D*1.

the individual doses will be the same. In that case the cumulative response during the interval from the n-th to the $(n + 1)$-th dose is equal to the sum of the first n segments, superimposed one upon another, of the response to a single dose (*cf.* figs. 11, 12). If this sum converges as $n \to \infty$, then it can be concluded that the system approaches asymptotic drug levels on repetitive drug administration, which levels are given by the limit to which this sum converges.[8]

The asymptotic mean levels at the various anatomical points of the system are given by the integrals of the corresponding drug levels over one dosage cycle, divided by the cycle duration τ. At any given anatomical point in a linear system which is time-

[8] To form this sum one theoretically needs to know the response to a single dose out to time $t \to \infty$. In practice this will require extrapolation from data obtained over a finite time period, and this extrapolation may be a source of error. (*cf.* section 4 below).

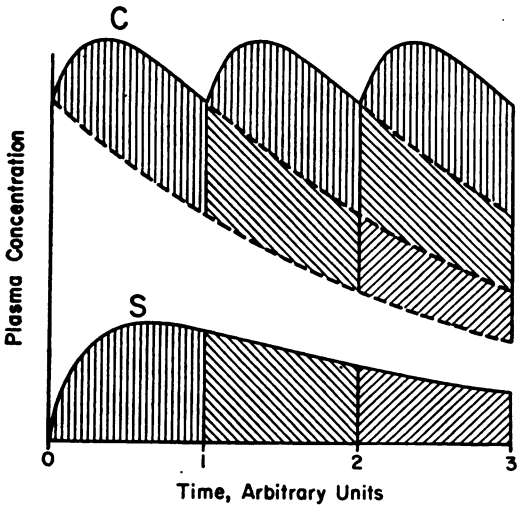

FIG. 12. On repetitive dosage in a linear, time-invariant system, the integral of (area under) the asymptotic cumulative response over one dosage cycle equals the integral of (area under) the response to a single dose from 0 to ∞. Curve S is the response to a single dose; curve C is the asymptotic cumulative response after a long series of regularly-repeated doses. The dashed lines show how the cumulative level would decay if dosing were stopped at various times. In the time interval from 0 to 1, the vertical distance between the dashed line and curve C at any time is equal to the height of curve S. Similarly, in the time interval from 1 to 2, the segment of curve C is equal to the segment of the lowest dashed line plus the first two segments of curve S. Equal areas are indicated by similar crosshatching. The area under curve C over one time unit is equal to the sum of the "stacked" segments of the area under curve S.

invariant or periodic with the same period as the dosage cycle, the asymptotic drug level during each cycle is the sum of all successive segments of length τ of the response to a single dose; and therefore the integral of the asymptotic drug level over one cycle equals the integral of the drug level after a single dose, from time $t = 0$ to ∞ (*cf.* fig. 12). We, therefore, have

$$\bar{y}_\infty = \frac{1}{\tau}\int_0^\infty y\,dt, \qquad (15)$$

where y is the response to a single dose, and $\bar{y}_\infty$ gives the asymptotic mean drug level (or levels, if y and $\bar{y}_\infty$ are vectors) on repeated dosage at the interval τ (*cf.* fig. 12).

b. Irreversible Drug Actions. Jusko (18, 19) has developed a theoretical approach to irreversible chemotherapeutic or teratogenic agents. For a chemotherapeutic agent, Jusko assumes that the rate of malignant cell destruction is proportional to the number of living cells and the local tissue drug concentration X_t. The cells are also assumed either to die spontaneously or to divide, both processes taking place at overall rates proportional to the number of cells. Letting S_t be the fraction of cells surviving after a time t, Jusko has derived the equation (slightly modified here)

$$\log_e S_t = -k\int_0^t X_t\,dt + k_g t. \qquad (16)$$

The term $k_g t$ reflects the net change due to spontaneous cell death and division.

If the dose of chemotherapeutic agent is finite (*i.e.*, administration is not continued indefinitely), then one may reasonably assume that X_t rises to a maximum and declines asymptotically to zero, and that for some sufficiently large time T,

$$\int_0^T X_t\,dt \simeq \int_0^\infty X_t\,dt. \qquad (17)$$

Jusko showed that for a two-compartment open system the latter integral was directly proportional to the dose D; and hence, from equation (16), he arrived at the equation:

$$\log_e S_T = -K_e D + k_g T, \qquad (18)$$

where K_e is a constant. For a fixed value of T, this equation predicts a linear relation between $\log_e S_T$ and the dose D, a relation which can be tested experimentally.

As we have shown in this review, however, a direct proportionality between the dose D and the time-integral of tissue concentration holds for any linear system. Furthermore, it holds not only for a single dose but also for any pattern of successive doses or infusions, provided this pattern remains unchanged except for a scale factor. It follows that equation (18) is not tied to the two-compartment model, or to single-dose drug

administration, but holds for all linear pharmacokinetic systems and for any fixed pattern of drug administration.

c. Competitive Drug Antagonism. According to the most commonly-used theory of competitive drug antagonism (1), the ratio of equieffective concentrations of an agonist A in the presence and absence of a competitive antagonist B is given by

$$\frac{[A]}{[A]_0} = 1 + \frac{[B]}{K_B}, \qquad (19)$$

where K_B is the drug-receptor dissociation equilibrium constant for the antagonist B, and $[A]$ is the concentration of agonist which, in the presence of an antagonist concentration $[B]$, produces a biological response exactly equal to that produced by the concentration $[A]_0$ in the absence of antagonist. Equation (19) can be applied readily to isolated tissues where the known drug concentrations in the bathing medium can be assumed to prevail at the receptors after diffusion equilibrium has been established, but in intact animals the situation is considerably more complicated. The drug concentrations at the receptors are usually not only unknown but also inconstant, rising to a maximum soon after injection of the drug, and then falling. Nevertheless, if the system is pharmacokinetically linear, and if the routes and schedule of drug administration are not changed, then the drug concentrations at the receptors are directly proportional to the doses given; and if, at any given time, all receptors can be assumed to be exposed to the same concentration of drug A and to the same concentration of drug B,[9] then the concentrations in equation (19) can be replaced by the corresponding doses, and the equation will still hold. Of course, K_B will not then be a true dissociation equilibrium constant, but rather an empirical constant, with the same units as the dose of B; and it will reflect drug distribution as well as affinity for the receptor. (The fact that the apparent K_B is influenced by drug distribution is important to keep in mind when evaluating studies such as those undertaken to determine, by comparing K_B values, whether *beta*-adrenergic receptors in different organs differ.)

This method of analysis is invalid if the system is pharmacokinetically non-linear, or if the receptors are distributed in two or more pharmacokinetically different compartments or in a pharmacokinetically non-uniform region. The assumption of pharmacokinetic linearity is especially vulnerable because the drug-receptor reaction itself involves saturable receptor sites and therefore necessarily introduces non-linearity. Moreover, competitive antagonism requires that the degree of receptor site saturation be high, and therefore the drug-receptor interaction cannot be even approximately linear. The validity of this approach to competitive antagonism *in vivo* therefore rests on the assumption that the number of receptors is too small for the drug-receptor reaction to affect significantly the kinetics of the agonist and antagonist concentrations in the receptor region. This assumption is probably valid in many cases, but it is probably not valid for low concentrations of drugs with very high affinities for the receptor, such as atropine (46, 47).

The direct proportionality between the dose and the drug concentration at the receptors holds only if we always measure the concentration at the same time-interval after starting the administration of the drug, and only if we use always the same routes and schedule of drug administration, changing only the dose. Strictly speaking, therefore, the two drugs A and B must always be administered in the same time-relationship, and the response must be measured at a fixed time-interval after starting the drugs. These requirements can often be relaxed somewhat, however; for if the concentration of one of the drugs remains reasonably constant for a prolonged

[9] Without this assumption one cannot infer equal receptor activation from equal responses (46), which is a necessary step in the derivation of equation (19).

interval, then the exact timing of its administration becomes no longer important. Also, if the antagonist concentration remains constant, then one can simply record peak responses to the agonist, since under the assumptions already made the peak response to the agonist will always follow the injection of the drug by the same time-interval.

d. Delayed-release Pharmaceutical Formulations. Delayed-release pharmaceutical formulations may release drug at a rate which is not a simple exponential, *i.e.*, is not proportional to the quantity of drug remaining unreleased. However, the factor controlling the rate of release is generally the time elapsed, and not the quantity of drug remaining. For example, if one gives two such timed-release capsules, the rate of release will be at all times just twice the rate of release from one capsule; but the rate of release after half the drug has been released from two capsules may not equal the initial rate of release from a single capsule, though the amount of drug remaining is the same. In general, a timed-release preparation can be regarded as providing a fixed time-schedule of drug administration; and, as with other inputs, the resulting drug levels obey the principle of superposition if the system is linear.

Krüger-Thiemer and Eriksen (24) have proved for a one-compartment model of the body that the response to a delayed-release pharmaceutical preparation which released part of its drug immediately was equal to the sum of the responses to the rapidly-released and slowly-released parts given separately. Their laborious proof is unnecessary, however, for this conclusion is nothing but a statement of the principle of superposition, and it therefore holds not only for their one-compartment model, but also for all other linear models.

e. Steady-state Flux Across Linear Membranes. Danielli (10) analyzed passive diffusion across homogeneous membranes and across membranes consisting, in effect, of a succession of identical compartments. Buerger (4) treated a more general type of membrane, consisting of a more or less arbitrary arrangement of compartments. Here we shall consider a still more general type of membrane, within which may occur any of the types of pharmacokinetic process mentioned above, and for which we shall make only the assumptions that the system is linear and that steady states exist for all constant inputs under consideration.

For a membrane separating two phases containing a diffusing substance at fixed concentrations x_1 and x_2, respectively, the steady-state concentration of the diffusing substance at every point within the membrane obeys the principle of superposition with respect to the input $\{x_1, x_2\}$, if the system is linear. Since all steady-state efflux rates from points within the membrane depend linearly either on the concentrations within the membrane or (*e.g.*, in discriminating systems) on the input concentrations themselves, they must also obey the principle of superposition, as must therefore the total efflux from the membrane into, say, the first phase. Since the steady-state influx into the membrane from the first phase is directly proportional to x_1, it follows that the net steady-state flux from the first phase into the membrane, which equals the net steady-state flux across the membrane, obeys the principle of superposition with respect to $\{x_1, x_2\}$. When $x_2 = 0$, the net steady-state flux is therefore directly proportional to x_1, the proportionality constant p_{21} being the effective permeability coefficient for passage from the first to the second phase; and when $x_1 = 0$ the net steady-state flux (in the reverse direction) is $p_{12}x_2$, where p_{12} is the effective permeability coefficient for passage in the reverse direction. In general, then, the net steady-state flux from the first to the second phase is given by

$$J_{2\leftarrow 1} = p_{21}x_1 - p_{12}x_2 . \qquad (20)$$

If the standard free energies and the activity coefficients of the diffusing substance are the same in the two phases, and if there

is no active transport, then p_{12} and p_{21} are equal, and the net steady-state flux is directly proportional to the concentration difference, with a single permeability coefficient.

It may be noted that no geometrical assumptions have been made, so that what we have called a "membrane" could be replaced by any linear system, even one which had no physical resemblance whatever to a true membrane (*e.g.*, a succession of metabolic transformations). The sole requirement for equation (20) is that the connection between the two phases be a linear system.

3. Prediction of System Behavior

As Westlake (53) has emphasized, the mere knowledge that a system is linear allows one to make important practical predictions about its behavior. It has already been indicated above, for example (section 2 a, figs. 11, 12), that the asymptotic drug levels on continuous infusion or repetitive dosage in a linear, time-invariant system can be predicted from the response to a single dose. Again, if one has measured the plasma levels in a linear system after an intravenous injection, and also after an intramuscular injection, then one can predict the responses to various combinations of intravenous and intramuscular doses by simply adding the responses to the intravenous and intramuscular injections given separately, after multiplying each by an appropriate dosage factor (*cf.* fig. 13). A similar procedure gives the response to the simultaneous oral administration of a delayed-release preparation and a standard preparation (*cf.* fig. 14), or to the injection of an intravenous bolus followed by a constant infusion.

4. Usefulness of Specific Models for Linear Systems

The principle of superposition is so powerful by itself that a question begins to arise whether specific pharmacokinetic models have any value at all for linear systems. In fact they do; and their value can be measured by the predictions and insights they provide that are not provided by the principle of superposition alone. Examples of such predictions or insights would include extrapolation in time, results with different routes of administration, and effects of model parameter changes.

Extrapolation in time is always hazardous, because in the course of time effects may begin to appear which were not detected in short-term experiments and which are therefore not provided for in the model used for extrapolation. For example, the data of Okita *et al.* (31) on digitoxin are well fitted by a two-compartment model in which the half-life of the slower exponential component of plasma elimination is 45 to 50 hr. Since the true biological half-life of this drug is much longer, an incautious reliance upon this model could have led to the use of maintenance doses which, on prolonged repetition, would have produced digitoxin accumulation to toxic levels. Another example has been given by Gibaldi and Weintraub (15). As emphasized by these authors and others [*e.g.*, Riggs (35)], the possibility of a prolonged phase of slow elimination, undetected in short-term experiments, should be recognized as a serious limitation on our ability to predict drug accumulation levels from single-dose responses, whether we base those predictions on a specific pharmacokinetic model or simply on the principle of superposition. The effect of such a slow elimination process would be a gradual upward drift in the accumulation levels after a seemingly steady state had been established. As a general rule, therefore, extrapolation in time should not be relied on too heavily, and the stability of drug accumulation levels on prolonged infusion or repetitive administration should be verified by direct observation.

With regard to changes in route of administration or model parameters, on the other hand, it may be possible to establish more satisfactorily the overall reliability of a model. Accumulated experience confirming the validity of the model in a variety of situations increases the confidence we can

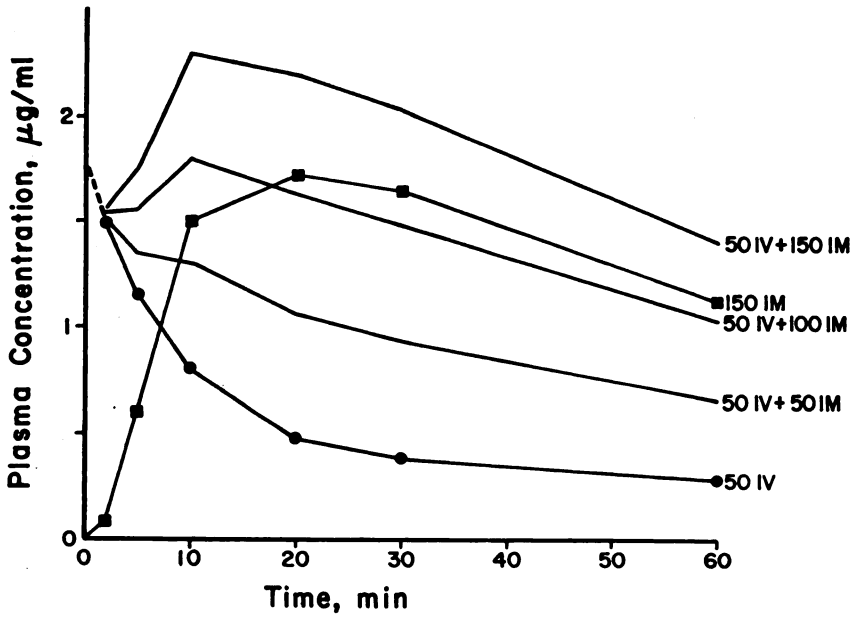

FIG. 13. Application of the principle of superposition to determine an optimal combination of intravenous and intramuscular doses. The curve labeled "50IV" (●) represents hypothetical data on plasma levels of a hypothetical drug after an intravenous dose of 50 mg. (The initial value is estimated by extrapolation, shown by the dashed line.) The curve labeled "150IM" (■) shows the "data" after an intramuscular dose of 150 mg. The other three curves were calculated by the superposition principle, and show the predicted responses to an intravenous dose of 50 mg, combined with intramuscular doses of 50, 100, or 150 mg, respectively. If the objective is to maintain the plasma level between 1 and 2 μg/ml, then the combination of 50 mg intravenously with 100 mg intramuscularly is the best of the three combinations. Note that although the plasma level falls below therapeutic levels in less than 8 min after the intravenous dose, there remains enough after 10 to 30 min to produce "toxic" plasma levels (over 2 μg/ml) when added to the levels produced by an ordinarily safe intramuscular dose. [The "data" in figure 13 are actually based on the data of Rowland *et al.* (38) and Sloman *et al.* (42) on lidocaine. For lidocaine, however, published observations on combined intravenous and intramuscular administration (40) do not agree well with predictions based on superposition of responses to intravenous and intramuscular doses given separately. Possibly the discrepancy is due to a difference in methods or subjects. If not, it indicates that lidocaine pharmacokinetics are non-linear, despite the apparent dose-independence noted by Rowland *et al.* (38). Some of the data (39, 40) suggest the possibility that therapeutic plasma levels of lidocaine somehow slow its absorption from an intramuscular site; but other data (2) indicate that plasma levels after intramuscular administration of different doses obey the principle of superposition.]

place in its predictions for untried situations. For example, if a model is found to fit the results of drug administration by two or three different routes, this tends to increase confidence in its predictions for still other routes of administration; or if a model is found to explain the pharmacokinetic differences between drugs which differ in such parameters as lipid solubility or binding affinity for plasma proteins, then confidence in its predictions for untried drugs is strengthened.

Specific pharmacokinetic models involve much computational labor; they usually oversimplify the actual system; they are often impossible to verify in detail; their parameters often cannot be accurately determined from available data (35, 54); and those parameters which can be determined often cannot be reconciled with values known from other experiments [*e.g.*, body compartment volumes (52)]. Despite these difficulties, there are many examples (*e.g.*, 8, 21, 33) of the effective use of models in

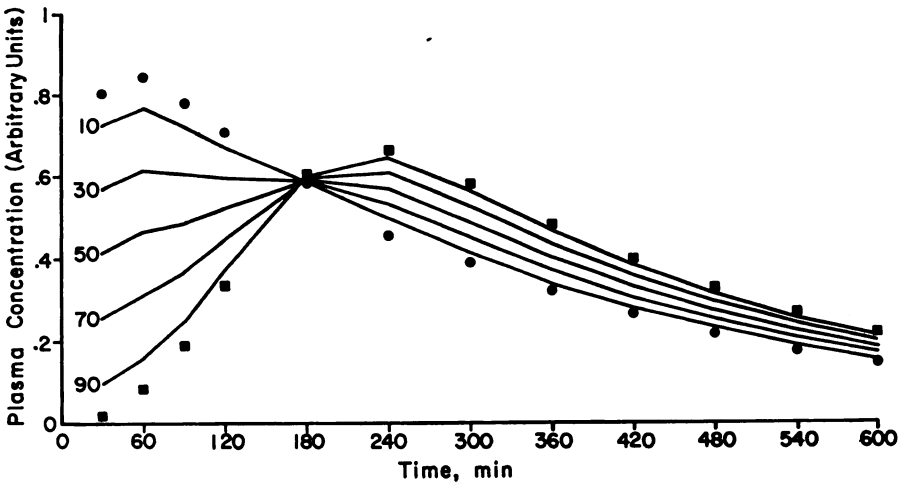

FIG. 14. Application of the principle of superposition to determine the optimal formulation of a hypothetical delayed-release pharmaceutical preparation. The circles (●) and squares (■) represent hypothetical plasma-level "data" on a standard drug preparation and a delayed-release preparation, respectively. The latter releases drug molecules with a normal distribution of release times (mean 120 min, standard deviation 60 min, distribution truncated at 0 so that there are no negative release times). The curves lying between the data points show the plasma-level responses to various combinations of the standard and delayed-release preparation, as calculated by the principle of superposition. The number at the left-hand end of each curve indicates the percentage of the delayed-release preparation in the combination. (The curves cross at about 177 min, and lie in inverted order on the right-hand side of the figure.) The optimal combination would appear to be one containing from 30 to 50% of the delayed-release preparation, the exact percentage depending upon the particular criteria to be met. Once the optimal proportions have been established, the total dosage can be adjusted by a scale factor as necessary. Note that 1) costly clinical studies are kept to a minimum, in that only the two pure preparations need be tested (in addition to final confirmatory studies on the mixture selected); 2) the mathematical complexities of the differential equations are entirely bypassed; 3) exactly the same method is applicable to any linear system, no matter how complicated (*e.g.*, a system with a fixed time-delay in absorption); and 4) for any linear system the method is theoretically an exact one, not an approximation.

explaining the effects of different routes of administration and of different model parameter values. On the other hand, some models seem to amount to little more than empirical data-fitting; and for a system known to be linear, such a model does not tell us any more than we can infer from the principle of superposition alone.

5. Non-linearity

It is unfortunately true that a great many pharmacokinetic systems, perhaps the majority, are non-linear. This is because of the very frequent occurrence in nature of processes whose kinetics are not first-order. Such processes include, as already mentioned, enzymatic transformations obeying Michaelis-Menten kinetics, binding reactions to limited numbers of binding sites on plasma proteins or in tissues, and active transport processes with limited available carrier. Strictly speaking, the principle of superposition does not hold for non-linear systems; but it does not necessarily follow that such systems are entirely beyond the reach of broad general treatments such as we have for linear systems. In the first place, the departures of a non-linear system from the principle of superposition may often be small enough to be ignored. Secondly, even when a system is non-linear from a conventional viewpoint, it is sometimes possible to formulate a particular superposition rule which it obeys.

To illustrate some possible approaches to non-linear systems, we consider a system containing zero-order processes and obeying

$$\frac{dy}{dt} = -ky + z + x, \tag{21}$$

where z is the rate of change of y due to zero-order processes. The variables x, y, and z may here be regarded as either ordinary (scalar) functions of time or vector functions of time.[10] Although equation (21) is technically a linear differential equation, the system is non-linear from our point of view because it does not obey the principle of superposition with respect to the input $\xi = \{x, y(0)\}$. There are, however, at least three ways to derive a superposition rule for this system.

The first method is to take the differences between outputs. According to equation (21) the outputs y_1 and y_2 for any two inputs $\{x_1, y_1(0)\}$ and $\{x_2, y_2(0)\}$ are defined by

$$\frac{dy_1}{dt} = -ky_1 + z + x_1 \tag{22a}$$

and

$$\frac{dy_2}{dt} = -ky_2 + z + x_2, \tag{22b}$$

respectively. Subtracting equation (22b) from equation (22a) we obtain

$$\frac{d(y_1 - y_2)}{dt} = -k(y_1 - y_2) + (x_1 - x_2). \tag{23}$$

Equation (23) has the same general form as equation (1). This system therefore obeys the principle of superposition with respect to *differences* between inputs and *differences* between outputs. Suppose, for example, one had a series of responses $y_1, y_2, \cdots, y_m$ to inputs $\alpha_1 x, \alpha_2 x, \cdots, \alpha_m x$, respectively, with $y_1(0) = y_2(0) = \cdots = y_m(0) = 0$. If the system obeyed equation (21) with $z \neq 0$ then it would not obey the principle of superposition, *i.e.*, the functions $y_1/\alpha_1, y_2/\alpha_2, \cdots, y_m/\alpha_m$ would not superimpose. However, one could subtract the first response from all the others, and the first input from all the others, and in this way obtain a series of curves $(y_2 - y_1)/(\alpha_2 - \alpha_1)$, $(y_3 - y_1)/(\alpha_3 - \alpha_1), \cdots, (y_m - y_1)/(\alpha_m - \alpha_1)$ which would superimpose.

The second method is to observe the outputs with weighted sums of inputs. Unlike the first method, this method requires planning the experiments in advance. If we add equations (22a) and (22b) after first multiplying by the weighting factors $\alpha/(\alpha + \beta)$ and $\beta/(\alpha + \beta)$, respectively, we obtain

$$\frac{d}{dt}\frac{\alpha y_1 + \beta y_2}{\alpha + \beta} = -k\frac{\alpha y_1 + \beta y_2}{\alpha + \beta} + z + \frac{\alpha x_1 + \beta x_2}{\alpha + \beta}. \tag{24}$$

This equation is of the same form as equation (21); and it states, therefore, that if y_1 and y_2 are the outputs for inputs $\{x_1, y_1(0)\}$ and $\{x_2, y_2(0)\}$, respectively, then $(\alpha y_1 + \beta y_2)/(\alpha + \beta)$ is the output for the input $\{(\alpha x_1 + \beta x_2)/(\alpha + \beta), [\alpha y_1(0) + \beta y_2(0)]/(\alpha + \beta)\}$. This obviously can be tested experimentally.

The third method works only for time-invariant or periodic systems. We denote by $y(t + \tau)$ a function whose value at time t equals the value of y at some later time $t + \tau$, where the interval τ is fixed; and, for the discussion of this method only, we use the notation $y(t)$ for y. We define $k(t + \tau)$, $k(t)$, $z(t + \tau)$, *etc.*, similarly. Then from equation (21) we have

$$\frac{dy(t)}{dt} = -k(t)y(t) + z(t) + x(t) \tag{25a}$$

[10] The function z (or its components, if it is a vector function) may be either positive or negative. However, if z is negative or has negative components then equation (21) cannot be assumed to hold generally for a physical system, because under certain conditions it would lead to a physical impossibility, namely negative mass. In such cases, therefore, equation (21) must be assumed to hold only over certain restricted ranges of x and y.

and

$$\frac{dy(t+\tau)}{dt} = -k(t+\tau)y(t+\tau) + z(t+\tau) + x(t+\tau). \quad (25b)$$

If $k(t+\tau) = k(t)$ and $z(t+\tau) = z(t)$ (*i.e.*, if the system is time-invariant or periodic with period τ) then taking the difference between equations (25a) and (25b) we have

$$\frac{d}{dt}[y(t+\tau) - y(t)] = -k(t)[y(t+\tau) - y(t)] + x(t+\tau) - x(t). \quad (26)$$

This equation is of the same general form as equation (1), and therefore the system obeys the principle of superposition with respect to the output $\zeta = y(t+\tau) - y(t)$. In practice, if one had a series of plasma level curves $y_1, y_2, \cdots, y_m$ resulting from inputs $\alpha_1 x, \alpha_2 x, \cdots, \alpha_m x$, respectively, with $y_1(0) = y_2(0) = \cdots = y_m(0) = 0$, then one could choose a suitable fixed time-interval τ and compute a series of differences $y(t+\tau) - y(t)$ at several t values for each output curve. These computed differences could then be divided by the dose and plotted against t, and the curves of $[y_1(t+\tau) - y_1(t)]/\alpha_1$, $[y_2(t+\tau) - y_2(t)]/\alpha_2, \cdots,$ $[y_m(t+\tau) - y_m(t)]/\alpha_m$ would superimpose.

Other non-linear systems may likewise yield to some of these approaches. The method of differences [equation (23)] seems especially likely to be generally useful. For example, for inputs $\{x_i, y_i(0)\}$ which do not differ too much from some standard input $\{x_1, y_1(0)\}$ the output differences $y_i - y_1$ of a non-linear system may approximately obey the principle of superposition with respect to inputs defined as $\{[x_i - x_1], [y_i(0) - y_1(0)]\}$. Another approach is to set up a specific model for the non-linear part of the system, leave the linear part undefined, and try to derive some function of the observed parameters which obeys the principle of superposition without regard to the details of the linear part. An example of this has been given elsewhere [46, equation (11)]. Further theoretical studies may well lead to new and better methods for the general analysis of various types of non-linear pharmacokinetic system.

General Conclusions

The analysis of pharmacokinetic systems in terms of the general concept of linearity allows us to see beyond particular kinetic details and obtain a better insight into overall system behavior. The habit of thinking of system inputs and responses as functions or sets rather than as simple numbers, and the concept of addition of such complicated inputs and responses, greatly improve our conceptual framework for dealing with various dosage schedules and their combinations. Methods of analysis based on linearity and superposition have broad generality, and this generality may often obviate some of the problems posed by the complexity and variability of biological systems and the difficulty of experimentally verifying pharmacokinetic models and evaluating parameters.

From a few examples, one can easily acquire a working intuitive grasp of the conconcept of linearity; and with experience one learns how to apply the concept to an ever broader range of problems. An appreciation of general system properties such as linearity and time-invariance will encourage an experimental scientist to test for such properties; and then even though he may be unwilling to tackle a full mathematical pharmacokinetic analysis himself, he will have obtained the experimental data which another analyst will find essential. The theorist, for his part, should recognize the experimental manifestations of non-linearity, so that he does not waste time trying to contrive a linear model to fit non-linear data.

Pharmacological theories based on particular pharmacokinetic models are subject to doubt when those models are shown to

be inapplicable. In many cases, however, a theory can be based just as easily on the simple assumption of linearity, with no assumptions regarding pharmacokinetic details. A theory resting on this foundation is much less vulnerable than one based on a possibly erroneous specific pharmacokinetic model.

Finally, even though the concept of linearity is very broad and non-specific, it is quite practical, because it provides a basis for utilizing empirical data to make specific practical predictions.

REFERENCES

1. Arunlakshana, O. and Schild, H. O.: Some quantitative uses of drug antagonists. Brit. J. Pharmacol. Chemother. 14: 48–58, 1959.
2. Bellet, S., Roman, L., Kostis, J. B. and Fleischmann, D.: Intramuscular lidocaine in the therapy of ventricular arrhythmias. Amer. J. Cardiol. 27: 291–293, 1971.
3. Birkhoff, G. and MacLane, S.: A Survey of Modern Algebra, 3rd ed., The Macmillan Company, New York, 1965.
4. Buerger, A. A.: A theory of integumental penetration. J. Theor. Biol. 14: 66–73, 1967.
5. Cohen, E. N., Corbascio, A. and Fleischli, G.: The distribution and fate of *d*-tubocurarine. J. Pharmacol. 147: 120–129, 1965.
6. Collatz, L.: Functional Analysis and Numerical Mathematics, Academic Press, New York, 1966.
7. Crank, J.: The Mathematics of Diffusion, Oxford University Press, London, 1956.
8. Crawford, J. S.: Speculation: the significance of varying the mode of injection of a drug. Brit. J. Anaesth. 38: 638–640, 1966.
9. Cummings, A. J. and Martin, B. K.: Interpretation of the kinetics of salicylic acid elimination. J. Pharm. Sci. 57: 891–893, 1968.
10. Danielli, J. F.: The theory of penetration of a thin membrane. *In* The Permeability of Natural Membranes, by H. Davson and J. F. Danielli, pp. 341–352, The Macmillan Company, New York, 1943.
11. Dayton, P. G., Cucinell, S. A., Weiss, M. and Perel, J. M.: Dose-dependence of drug plasma level decline in dogs. J. Pharmacol. Exp. Ther. 158: 305–316, 1967.
12. DiSanto, A. R. and Wagner, J. G.: Potential erroneous assignment of nonlinear data to the classical linear two-compartment open model. J. Pharm. Sci. 61: 552–555, 1972.
13. Fife, D.: Which linear compartmental systems contain traps? Math. Biosci. 14: 311–315, 1972.
14. Gaddum, J. H: Repeated doses of drugs. Nature (London) 153: 494, 1944.
15. Gibaldi, M. and Weintraub, H.: Some considerations as to the determination and significance of biologic half-life. J. Pharm. Sci. 60: 624–626, 1971.
16. Hearon, J. Z.: Theorems on linear systems. Ann. N. Y. Acad. Sci. 108: 36–68, 1963.
17. Jacquez, J. A.: Compartmental Analysis in Biology and Medicine, Elsevier Publishing Co., Amsterdam, 1972.
18. Jusko, W. J.: Pharmacodynamics of chemotherapeutic effects: dose-time-response relationships for phase-non-specific agents. J. Pharm. Sci. 60: 892–895, 1971.
19. Jusko, W. J.: Pharmacodynamic principles in chemical teratology: dose-effect relationships. J. Pharmacol. Exp. Ther. 183: 469–480, 1972.
20. Jusko, W. J. and Levy, G.: Absorption, metabolism, and excretion of riboflavin-5′-phosphate in man. J. Pharm. Sci. 56: 58–62, 1967.
21. Kety, S. S.: The theory and applications of the exchange of inert gases at the lungs and tissues. Pharmacol. Rev. 3: 1–41, 1951.
22. Krüger-Thiemer, E.: Pharmacokinetics and dose-concentration relationships. *In* Physico-Chemical Aspects of Drug Action, ed. by E. J. Ariëns, pp. 63–113, Pergamon Press, Oxford, 1968.
23. Krüger-Thiemer, E.: Nonlinear dose-concentration relationships. Farmaco (Pavia) Ed. Sci. 23: 717–756, 1968.
24. Krüger-Thiemer, E. and Eriksen, S. P.: Mathematical model of sustained-release preparations and its analysis. J. Pharm. Sci. 55: 1249–1253, 1966.
25. Levy, G.: Evidence for nonfirst-order kinetics of salicylate elimination—a rebuttal. J. Pharm. Sci. 56: 1044–1046, 1967.
26. Levy, G.: Dose-dependent effects in pharmacokinetics. *In* Importance of Fundamental Principles in Drug Evaluation, ed. by D. H. Tedeschi and R. E. Tedeschi, pp. 141–172, Raven Press, New York, 1968.
27. Levy, G., Tsuchiya, T. and Amsel, L. P.: Limited capacity for salicyl phenolic glucuronide formation and its effect on the kinetics of salicylate elimination in man. Clin. Pharmacol. Ther. 13: 258–268, 1972.
28. Margenau, H. and Murphy, G. M.: The Mathematics of Physics and Chemistry, D. Van Nostrand Company, Inc., Princeton, N. J., 1956.
29. Meier, P. and Zierler, K. L.: On the theory of the indicator-dilution method for measurement of blood flow and volume. J. Appl. Physiol. 6: 731–744, 1954.
30. Nagashima, R., Levy, G. and O'Reilly, R. A.: Comparative pharmacokinetics of coumarin anticoagulants. IV. Application of a three-compartmental model to the analysis of the dose-dependent kinetics of bishydroxycoumarin elimination. J. Pharm. Sci. 57: 1888–1895, 1968.
31. Okita, G. T., Talso, P. J., Curry, J. H., Smith, F. D. and Geiling, E. M. K.: Blood level studies of C^{14}-digitoxin in human subjects with cardiac failure. J. Pharmacol. Exp. Ther. 113: 376–382, 1955.
32. Piotrowski, J.: The Application of Metabolic and Excretion Kinetics to Problems of Industrial Toxicology, U. S. Government Printing Office, Washington, 1971.
33. Price, H. L., Kovnat, P. J., Safer, J. N., Conner, E. H. and Price, M. L.: The uptake of thiopental by body tissues and its relation to the duration of narcosis. Clin. Pharmacol. Ther. 1: 16–22, 1960.
34. Rescigno, A. and Segre, G.: Drug and Tracer Kinetics. Blaisdell Publishing Company, Waltham, 1966.
35. Riggs, D. S.: The Mathematical Approach to Physiological Problems, The Williams & Wilkins Company, Baltimore, 1963.
36. Rossum, J. M. van: Pharmacokinetics of accumulation. J. Pharm. Sci. 57: 2162–2164, 1968.
37. Rossum, J. M. van and Tomby, A. H. J. M.: Multicompartment-kinetics and the accumulation plateau. Arch. Int. Pharmacodyn. Ther. 188: 200–203, 1970.
38. Rowland, M., Thomson, P. D., Guichard, A. and Melmon, K. L.: Disposition kinetics of lidocaine in normal subjects. Ann. N. Y. Acad. Sci. 179: 383–398, 1971.
39. Scott, D. B., Jesson, P. J., Vellani, C. W. and Julian, D. G.: Plasma-levels of lignocaine after intramuscular injection. Lancet 2: 1209–1210, 1968.
40. Scott, D. B., Jesson, P. J., Vellani, C. W. and Julian, D. G.: Plasma-lignocaine levels after intravenous and intramuscular injection. Lancet 1: 41, 1970.
41. Searle, S. R.: Matrix Algebra for the Biological Sciences, John Wiley & Sons, Inc., New York, 1966.
42. Sloman, G., Isaac, P., Murton, L. and Harper, R.: Plasma levels of lignocaine after intramuscular injection. Med. J. Aust. 2: 655–657, 1971.
43. Stephenson, J. L.: Theory of the measurement of blood flow by the dilution of an indicator. Bull. Math. Biophys. 10: 117–121, 1948.
44. Teorell, T.: Kinetics of distribution of substances administered to the body. II. The intravascular modes of administration. Arch. Int. Pharmacodyn. Thér. 57: 226–240, 1937.

45. THRON, C. D.: Structure and kinetic behavior of linear multicompartment systems. Bull. Math. Biophys. 34: 277-291, 1972.
46. THRON, C. D.: Nonlinear kinetics of atropine action on the pacemaker of the isolated guinea-pig atrium. J. Pharmacol. Exp. Ther. 181: 529-537, 1972.
47. THRON, C. D. AND WAUD, D. R.: The rate of action of atropine. J. Pharmacol. Exp. Ther. 160: 91-105, 1968.
48. WAGNER, J. G.: Fallacy in concluding there are zero-order kinetics from blood level and urinary excretion data. J. Pharm. Sci. 56: 586-594, 1967.
49. WAGNER, J. G.: Biopharmaceutics and Relevant Pharmacokinetics, Drug Intelligence Publications, Hamilton, Ill., 1971.
50. WAGNER, J. G.: Notes supplied for a pharmacokinetic seminar sponsored by J. M. Richards Laboratory and held at Northland Inn, Southfield, Mich., June 19-21, 1972.
51. WAGNER, J. G., NORTHAM, J. I., ALWAY, C. D. AND CARPENTER, O. S.: Blood levels of drug at the equilibrium state after multiple dosing. Nature (London) 207: 1301-1302, 1965.
52. WATERHOUSE, C. AND KEILSON, J.: Transfer times across the human body. Bull. Math. Biophys. 34: 33-44, 1972.
53. WESTLAKE, W. J.: Problems associated with analysis of pharmacokinetic models. J. Pharm. Sci. 60: 882-885, 1971.
54. WESTLAKE, W. J.: Use of statistical methods in evaluation of *in vivo* performance of dosage forms. J. Pharm. Sci. 62: 1579-1588, 1973.
55. WIDMARK, E. M. P.: Studies in the concentration of indifferent narcotics in blood and tissues. Acta Med. Scand. 52: 87-164, 1919.
56. ZADEH, L. A. AND DESOER, C. A.: Linear System Theory, McGraw-Hill Book Company, New York, 1963.

Appendix

1. Matrices. A *matrix* is a rectangular array of numbers or other elements, *e.g.*,

$$\begin{bmatrix} -1 & 3 \\ 0 & 5 \\ 2 & 2 \end{bmatrix}, \begin{bmatrix} a & bx+c \\ gy & h \end{bmatrix}. \tag{A1}$$

When it is desired to refer to the individual elements of a matrix, these are usually identified by doubly-subscripted letters, the subscripts denoting the row and column, respectively:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}. \tag{A2}$$

The entire matrix may be denoted by a single letter, as

$$a = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}. \tag{A3}$$

The use of bold-face type for letters denoting matrices is a common but not universal practice.

A *vector* is a matrix with only a single row (row vector) or column (column vector), *e.g.*,

$$[r_1 \quad r_2 \quad r_3 \quad r_4], \qquad \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}. \tag{A4}$$

As illustrated, only one subscript is necessary for the elements of a vector.

Two matrices are said to be equal if and only if they have the same number of rows, they have the same number of columns, and all corresponding elements are equal. For example:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \tag{A5}$$

means

$$\begin{aligned} a_{11} &= b_{11}\,, \; a_{12} = b_{12}\,, \; a_{13} = b_{13}\,, \\ a_{21} &= b_{21}\,, \; a_{22} = b_{22}\,, \; a_{23} = b_{23}\,. \end{aligned} \tag{A6}$$

The matrices in equation (A5) can be denoted by single letters a and b respectively, and the equation can then be written as

$$a = b. \tag{A7}$$

Equation (A7) is equivalent to the six equations (A6), and illustrates the simplification of notation achieved with matrices.

The *sum* of two matrices can be formed if and only if they have the same number of rows and they have the same number of columns. Their sum is a third matrix with the same number of rows and the same number of columns, in which each element is the sum of the corresponding elements of the two matrices being added. For example:

$$\begin{bmatrix} -1 & 3 \\ 0 & 5 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & z^2 \\ -p & -2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 0 & 5+z^2 \\ 2-p & 0 \end{bmatrix}. \tag{A8}$$

With single-letter symbols for the matrices, equation (A8) might be written

$$a + b = c. \tag{A9}$$

The matrix equation (A9) is then the equivalent of six simple algebraic or arithmetical equations expressing the addition of the individual matrix elements.

The *product* of two matrices can be formed if and only if the first has exactly as many columns as the second has rows. Consider first the so-called *inner product* of two vectors, the first a row vector and the second a column vector. This is defined as the sum of the products of corresponding elements. For examp'e:

$$rc = [r_1 \quad r_2 \quad r_3] \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{A10}$$

$$= r_1c_1 + r_2c_2 + r_3c_3 .$$

The sum on the right-hand side, containing the products of corresponding terms of the two vectors, is a number, not a matrix; therefore the inner product of two vectors is a number, not a matrix.

For matrices other than vectors, the rule for multiplication is as follows: the element in the i-th row and the j-th column of the product is the inner product of two vectors, namely the i-th row of the first matrix and the j-th column of the second. For example,

$$ab = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{bmatrix}. \tag{A11}$$

The product has as many rows as the first matrix and as many columns as the second.

Note that matrix multiplication is not always commutative, *i.e.*, ab may not equal ba. In fact, in the example of equation (A11) the product ba cannot be formed at all, because b has fewer columns than a has rows.

Multiplication of a matrix by a number multiplies every element of the matrix by that number, *e.g.*,

$$\alpha a = \alpha \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \alpha a_{13} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{23} \end{bmatrix}. \tag{A12}$$

This operation is commutative, *i.e.*, $\alpha a = a\alpha$.

A vector or matrix whose elements are functions of time is sometimes referred to as a *vector* or *matrix function of time*. Differentiation of a vector or matrix function of time is equivalent to differentiation of each of its elements individually, *e.g.*,

$$\frac{dy}{dt} = \frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \\ \frac{dy_3}{dt} \end{bmatrix}. \tag{A13}$$

One can easily verify from this definition that the derivative of a sum of vector or matrix functions of time is the sum of the derivatives, *i.e.*,

$$\frac{d(x+y)}{dt} = \frac{dx}{dt} + \frac{dy}{dt}, \tag{A14}$$

and also that, for any number α,

$$\frac{d(\alpha y)}{dt} = \alpha \frac{dy}{dt}. \tag{A15}$$

An important set of general algebraic laws which can readily be shown to hold for matrix addition and multiplication are the distributive laws. These may be stated as follows for any matrices a, b and c such that a can multiply b and c, and b and c can be added, and for any two numbers α and β:

$$a(b + c) = ab + ac, \tag{A16a}$$

$$\alpha(b + c) = \alpha b + \alpha c, \tag{A16b}$$

and

$$(\alpha + \beta)b = \alpha b + \beta b. \tag{A16c}$$

More on matrix algebra can be found in various texts (*e.g.*, 28, 41).

2. Multicompartment systems (16; 17, pp. 48 *ff.*; 45; 46). We consider a system of n compartments, in which a substance is distributed. We assume that all transfers of the substance out of compartments (whether to other compartments, to the outside world, or to oblivion) obey first-order kinetics. Let y_i equal the quantity of substance in the i-th compartment, and let k_{ji} be the first-order rate constant for transfer from the i-th to the j-th compartment, so that the rate of (unidirectional) transfer from the i-th to the j-th compartment is $k_{ji}y_i$. Let everything outside the compartment system be taken as the 0-th compartment, so that k_{0i} is the rate constant for elimination from the i-th compartment. Let x_i be the rate of injection into the i-th compartment. Then the rate of change of the quantity in the i-th compartment is given by

$$\frac{dy_i}{dt} = \sum_{j=1}^{n}{}' k_{ij}y_j - \left(\sum_{j=0}^{n}{}' k_{ji}\right) y_i + x_i, \tag{A17}$$

where the primed summation sign indicates a summation from which the term with $j = i$ is omitted. An equation of this form holds for each compartment ($i = 1, 2, \cdots, n$). These n equations may be written out as follows:

$$\begin{aligned} \frac{dy_1}{dt} &= -\left(\sum_{j=0}^{n}{}' k_{j1}\right) y_1 + k_{12}y_2 \\ &\quad + \cdots + k_{1n}y_n + x_1, \\ \frac{dy_2}{dt} &= k_{21}y_1 - \left(\sum_{j=0}^{n}{}' k_{j2}\right) y_2 \\ &\quad + \cdots + k_{2n}y_n + x_2, \\ &\cdots \end{aligned} \tag{A18}$$

$$\frac{dy_n}{dt} = k_{n1}y_1 + k_{n2}y_2 + \cdots - \left(\sum_{j=0}^{n}{}' k_{jn}\right) y_n + x_n.$$

These equations can be written in matrix form as follows:

$$\begin{bmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \\ \cdots \\ \frac{dy_n}{dt} \end{bmatrix} = \begin{bmatrix} -\left(\sum_{j=0}^{n}{}' k_{j1}\right) y_1 + k_{12}y_2 + \cdots + k_{1n}y_n + x_1 \\ k_{21}y_1 - \left(\sum_{j=0}^{n}{}' k_{j2}\right) y_2 + \cdots + k_{2n}y_n + x_2 \\ \cdots \\ k_{n1}y_1 + k_{n2}y_2 + \cdots - \left(\sum_{j=0}^{n}{}' k_{jn}\right) y_n + x_n \end{bmatrix}. \tag{A19}$$

By the rules of matrix algebra, equation (A19) is equivalent to

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = -\begin{bmatrix} \left(\sum_{j=0}^{n}{}' k_{j1}\right) & -k_{12} & \cdots & -k_{1n} \\ -k_{21} & \left(\sum_{j=0}^{n}{}' k_{j2}\right) & \cdots & -k_{2n} \\ \cdots & & & \\ -k_{n1} & -k_{n2} & \cdots & \left(\sum_{j=0}^{n}{}' k_{jn}\right) \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}. \tag{A20}$$

This can be seen as follows. If we find the inner product of the first row of the first matrix on the right-hand side of equation (A20) with the column vector of the y_i, we obtain the sum $\left(\sum_{j=0}^{n}{}' k_{j1}\right)y_1 - k_{12}y_2 - \cdots - k_{1n}y_n$. The (−) sign preceding the matrix in equation (A20) then changes the signs of all the terms of this sum. Continuing in this way, we find that the result of multiplying out the first two matrices on the right-hand side of equation (A20) is a column vector of algebraic expressions which are identical to the first n terms of the expressions in the right-hand matrix of equation (A19). Addition of the corresponding elements of the column vector of the x_i in equation (A20) then gives a column vector of algebraic expressions identical to the right-hand side of equation (A19).

Equation (A20) can now be written with single-letter symbols y, k, and x for the matrices, as follows:

$$\frac{dy}{dt} = -ky + x. \tag{A21}$$

This equation is formally identical to equation (1). The laws of matrix differentiation, addition and multiplication, as expressed in equations (A12–A16), then validate the derivation of equation (4) from equations (3a) and (3b) for the case when these equations are matrix equations.