Regression by composition

Daniel M. Farewell

Division of Population Medicine, School of Medicine, Cardiff University, UK

E-mail: farewelld@cardiff.ac.uk

Rhian M. Daniel

Division of Population Medicine, School of Medicine, Cardiff University, UK

Mats J. Stensrud

Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Anders Huitfeldt

Division of Mental Health and Addiction, Oslo University Hospital, Norway

[To be read before The Royal Statistical Society at the Discussion meeting held at Imperial College London, on Tuesday, 24 March, 2026, Professor Maria De Iorio in the Chair]

Summary.
We describe a modular regression framework in which covariate-dependent transformations are composed together and act on probability distributions. This framework is based on group actions of vector spaces, which are computationally convenient families of transformations that are well suited to model building and inference. We quantify covariate contributions to each group action through corresponding linear maps, and these are the only model parameters to be estimated. Algebraic features of group actions—notably, their invariant subsets—are informative about local statistical properties of the regression model. Vector space actions on affine spaces also provide a minimal geometric structure for comparing distributions, with affine transformations characterizing collapsible contrasts. In two substantive data analyses, we illustrate how unconventional models may be expressed as regressions by composition. We exhibit and extend existing nonlinear models for interpolating infant growth curves for individuals, and for producing standard population growth charts. We also use regression by composition to specify and fit a bespoke, mechanistically-motivated binary regression model for antiretroviral therapies in the treatment of HIV.

1. A compact characterization of the model, and three examples

Regression models parameterize the conditional distribution of a variate of interest given covariates. Covariate dependence is often expressed in terms of linear predictors: that is, linear combinations of covariate contrasts. Let Y be a variate of interest, and let $\mathcal{P}$ be a set of candidate distributions of Y. A *regression by composition* consists of

(a) a reference distribution $p_0 \in \mathcal{P}$,
(b) a finite sequence $\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3, \ldots$ of vector spaces, each with a group action on $\mathcal{P}$, and
(c) a corresponding sequence $\eta_1, \eta_2, \eta_3, \ldots$ of covariate-dependent linear predictors.

The reference distribution p_0 and the combined group actions define a parametric model $\mathcal{M} \subseteq \mathcal{P}$ indexed by the vector spaces $\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3, \ldots$, each parameter of which is amenable to regression modelling. Writing group action on the right, regression by composition specifies the conditional distribution P of Y, given covariates, as

$$P = ((p_0 \cdot \eta_1) \cdot \eta_2) \cdot \eta_3 \cdots ,$$

where $(p_0 \cdot \eta_1) \in \mathcal{P}$ is the result of the action of the linear predictor $\eta_1 \in \mathbb{V}_1$ on the distribution p_0, and so on. The conditional distribution P takes values in the composite parametric family $\mathcal{M}$.

McCullagh has long affirmed the conceptual and practical benefits of understanding treatment contrasts as group actions on probability distributions (see, for example, McCullagh, 1999, 2002), and more recently offered a compelling Socratic defense of this perspective (McCullagh, 2022, Section 21.1.6). Beginning with regression models for continuous and binary data, we show how *composing* group actions expands explanatory possibilities beyond what is possible in standard regression models, and envisage how calls to a modelling function might express these extensions in illustrative computer code. We then exhibit an oft-encountered nonlinear model that enjoys a parsimonious representation as a regression by composition.

1.1. Real-valued variate of interest

Let `y` refer to a real-valued variate, or response, of interest. In this imaginary clinical example, the only measured covariates are a binary treatment indicator `trt`, and `age` recording participant age at study entry. Gaussian linear regression of `y` on `trt`, written as a regression by composition, could be specified thus:

```
model y = N(0, 1) | Scale(1) | Translate(1 + trt)
```

In calling this modelling subroutine, the three terms to the right of '=' mirror those on the right-hand-side of the algebraic expression $P = (p_0 \cdot \eta_1) \cdot \eta_2$. Like the Unix pipe (Kernighan & Pike, 1984, p. 31), the symbol '`|`' expresses function composition, connecting output of one function to input of the next. Gaussian regression begins by setting p_0 to be the standard normal law $\mathrm{N}(0, 1)$ with mean zero and unit variance. The action on probability distributions (McCullagh, 2022, pp. 231 sqq.) of the linear predictor η_1—taking values in the vector space $(\mathbb{R}_+, \times)$ of positive real numbers under multiplication, with real exponentiation as scalar multiplication—is made explicit through a *flow* (another name for a group action) that we shall call `Scale()`. For concreteness and for reasons of familiarity, here we imagine specifying a real-valued covariate contrast linked with η_1 via a Wilkinson and Rogers (1973) formula (in this case, simply `1`, corresponding to a single,

common, variance term), but this is not essential. This associates to the flow `Scale()` a one-dimensional linear map $\beta_1 : \mathbb{R} \to \mathbb{R}_+$ to be estimated. As its name implies, this flow scales distributions, to varying degrees: each $v \in \mathbb{R}_+$ acts on a cumulative distribution function $F : \mathbb{R} \to [0, 1]$ as

$$(F \cdot v)(y) = F(y/v).$$

Here $F \cdot v$ does not denote multiplication, but instead the *action* of $v \in (\mathbb{R}_+, \times)$ on the distribution function F. In this paper, groups act on the right (Robinson, 1996, p. 34), mimicking piping; left action is conventional in other settings. Scaling a distribution that has zero mean changes its variance (but not its mean) and at this intermediate stage of model construction $p_0 \cdot \eta_1 = \mathrm{N}(0, \eta_1^2)$, the normal distribution with zero mean and variance η_1^2.

The real-valued linear predictor $\eta_2 \in (\mathbb{R}, +)$ acts via the flow `Translate()`. Its associated covariate contrast `1 + trt` incorporates into the model an unknown two-dimensional linear map $\beta_2 : \mathbb{R}^2 \to \mathbb{R}$ to be estimated, specifying a grand mean and a treatment contrast. For $v \in \mathbb{R}$, the action of the flow `Translate()` on cumulative distribution functions F is

$$(F \cdot v)(y) = F(y - v).$$

Translating distributions changes means (but not variances). Since $p_0 \cdot \eta_1 = \mathrm{N}(0, \eta_1^2)$, we deduce that the full model is

$$P = (p_0 \cdot \eta_1) \cdot \eta_2 = \mathrm{N}(\eta_2, \eta_1^2),$$

having 1+2 = 3 scalar quantities (specifying the two linear maps) to be estimated, analogous to the usual residual standard deviation and two regression coefficients.

This basic specification can be tweaked in many ways. A heteroskedastic model emerges by modifying the covariate contrast formula associated with the scaling action:

```
model y = N(0, 1) | Scale(1 + age) | Translate(1 + trt)
```

The standard deviation of the resulting normal distribution is still η_1, but now η_1 has attached to it a two-dimensional linear map from the covariate contrast space $(\mathbb{R}^2, +)$ into the multiplicative space $(\mathbb{R}_+, \times)$ of the linear predictor, allowing the residual standard deviation to vary continuously with age. In all, this model has four scalar parameters to be estimated.

Scaling and translating have here combined to yield a *transformation model* (Hothorn et al., 2014): a parameterized model of the conditional cumulative distribution function. Both flows arise from direct transformation of the variate of interest Y. More formally, they are induced by the action of a vector space $\mathbb{V}$ on the support Υ of the variate of interest through the relationship

$$(p \cdot v)(A) = p(A \cdot -v)$$

between laws $p, p \cdot v \in \mathcal{P}$ for all measurable $A \subseteq \Upsilon$, where $-v$ denotes the inverse element of $v \in \mathbb{V}$. In the language of optimal transport (see, for instance, Panaretos & Zemel, 2020,

p. 2), $p \cdot v = (\cdot v)\#p$ is the *pushforward* of p under the transformation $(\cdot v) : \Upsilon \to \Upsilon$. As another example of a transformation model, Jones and Pewsey (2009) describe a hyperbolic scaling transformation given by

$$y \cdot v = \sinh(v \sinh^{-1} y)$$

for $v \in (\mathbb{R}_+, \times)$ and $y \in \mathbb{R}$, inducing an action on $\mathcal{P}$ that changes the tail weight of distributions over the reals. Let us call the associated flow a `Cinch()`, by analogy with belt-tightening, and in a nod to the hyperbolic sine function; the three-flow regression by composition

```
model y = N(0, 1) |
            Cinch(1) | Scale(1 + age) | Translate(1 + trt)
```

specifies a model whose heavy-tailedness, two scale parameters admitting age-dependent dispersion, and two treatment-specific location parameters are all to be determined. If desired, treatment- or age-dependent kurtosis may equally be incorporated.

A rich class of models arises from composing data transformations. The expressive power of transformation models is underlined by the proven flexibility of *normalizing flows*: compositions of parameterized, differentiable, data transformations (Papamakarios et al., 2021). Many of the generalized additive models for location, scale and shape of Rigby and Stasinopoulos (2005) are also of this type. However, other group actions (such as convolution, compounding, and exponential tilting) cannot so easily be characterized in closed form in terms of data transformations. Several familiar statistical models, including proportional and additive hazards, Poisson regression, and binary regression—to which we now turn—are most naturally thought of as compositions of actions on distributions, not data.

1.2. Binary variate of interest

Imagine `y` is now a binary variate of interest, and again `trt` and `age` are covariates. To specify a logistic regression (Berkson, 1944; Cox, 1958; but see also Hanley, 2025, who traces its prehistory to the 1930s) of `y` on `age` and `trt`, we might write:

```
model y = Ber(1/2) | ScOdds(1 + age + trt)
```

Our more-or-less arbitrary starting point is a Bernoulli distribution with $p_0(Y = 1) = 1/2$. The flow `ScOdds()` scales by an odds ratio $v \in \mathbb{R}_+$ the odds $\omega = p(Y = 1)/p(Y = 0)$ of the event $\{Y = 1\}$, acting as

$$\omega \cdot v = \omega v.$$

As in standard logistic regression, there is a single linear map $\mathbb{R}^3 \to \mathbb{R}_+$ to be estimated, and hence three scalar coefficients.

Again the model can be amended easily. It is well known that odds ratios are not *collapsible* contrasts (meaning that the marginal odds ratio can be nearer to the null than

any conditional odds ratio; for a definition, see Greenland et al., 1999, or Section 4 for details), and a collapsible contrast may be preferred. When treatment lowers the outcome risk, it has also been argued (though not universally accepted) that risk ratios may be more *transportable* to different clinical settings than odds ratios (Huitfeldt et al., 2022; Piccininni & Stensrud, 2025). Let us imagine moving `trt` to a collapsible risk ratio flow `ScRisk1()`, acting on the probability $p = p(Y = 1) \in \mathbb{R}$ as $p \cdot v = pv$ for $v \in (\mathbb{R}_+, \times)$:

```
model y = Ber(1/2) | ScOdds(1 + age) | ScRisk1(0 + trt)
```

Here we have one linear map $\mathbb{R}^2 \to \mathbb{R}_+$ associated with `ScOdds(1 + age)` and another $\mathbb{R} \to \mathbb{R}_+$ associated with `ScRisk1(trt)` to be estimated, so again three scalar model parameters overall. A distinctive and sometimes appealing feature of the `ScRisk1()` flow is that it has only one *fixed* or $\mathbb{R}_+$*-invariant point*, namely $p = 0$. By this we mean that $0 \cdot v = 0$ for all $v \in \mathbb{R}_+$, where the model "constrain[s] the benefits [or harms] of treatment to be zero" (Deeks, 2002). The single fixed point of `ScRisk1()` contrasts with the two of the `ScOdds()` flow, namely $\omega = 0$ and $\omega = \infty$, or equivalently $p = 0$ and $p = 1$. A rather less advantageous property of `ScRisk1()` is that the set $\triangle$ of *probability* laws is not an invariant subset of the group action—because $[0, 1] \cdot \mathbb{R}_+ = [0, \infty) \neq [0, 1]$—so the modelled distribution P may stray outside the set $\triangle$. Section 3 discusses invariance at greater length.

Also unlike the odds ratio, the risk ratio is asymmetric in the sense that the complementary flow `ScRisk0()` given by $q \cdot v = qv$ for $q = 1 - p$ and for $v \in (\mathbb{R}_+, \times)$, or in terms of p as

$$p \cdot v = 1 - (1 - p)v$$

yields a group of (also collapsible) transformations that *differs* from `ScRisk1()`. The index v is sometimes called a *survival ratio*, to distinguish it from the risk ratio. In some contexts, there may be scientific reasons to choose one or the other (to "count the living or the dead", as Sheps, 1958, put it vividly), but regression by composition affords us the luxury to choose both. The model

```
model y = Ber(1/2) |
            ScOdds(1 + age) | ScRisk0(0 + trt) | ScRisk1(0 + trt)
```

allows treatment to scale the probability p of $Y = 1$, or the probability $q = 1 - p$ of $Y = 0$, or any combination of the two, including risk ratios and survival ratios as special cases, and therefore also both 'sides' of the *switch relative risk* model (van der Laan et al., 2007; Baker & Jackson, 2018). This collapsible, two-parameter treatment summary sits atop the two odds-ratio parameters and yields a binary regression model with a total of four scalar parameters to be estimated. There are no fixed points for the implied treatment contrast: two subjects differing only in their treatment group are free to have neither, one or both event probabilities equal to 0 or 1.

That we may associate two regression parameters with a single binary treatment is a feature only made possible because of the presence of another explanatory covariate (here,

age) in the model: without it, the model is overparameterized and not identifiable. Flexible and collapsible binary regression models such as these cannot be expressed as generalized linear models (Nelder & Wedderburn, 1972) for any choice of link function (Daniel et al., 2024), and were our original motivation for exploring regression by composition. In Section 7, we go further, and show how binary regression models with three or more treatment parameters can admit curvilinear (and hence noncollapsible), fixed-point-free relationships between untreated and treated event probabilities, including the possibility that treatment may benefit high-risk individuals while potentially harming lower-risk subgroups.

1.3. Positive-valued variate of interest

We illustrate how regression by composition can help us to understand and fit a somewhat delicate nonlinear model (Colquhoun, 1969; Blunck & Mommsen, 1978) that nevertheless sees widespread use in the physical, biological and human sciences (Keating & Quinn, 1998; Reeve & Turner, 2013; Cornish-Bowden, 2015; Walters et al., 2024). The Michaelis–Menten model (Michaelis & Menten, 1913; Johnson & Goody, 2011; Briggs & Haldane, 1925) describes a relationship between an explanatory variable $x \geq 0$ (often, time) and a variate of interest $y \geq 0$ (yield, volume, size, ...) characterized by the rectangular hyperbola

$$y = \frac{ax}{x + b}$$

for $a > 0$ and $b \in \mathbb{R}$. This hyperbola passes through $(0, 0)$ and has a horizontal asymptote at $y = a$ (with a vertical asymptote at $x = -b$), reaching half its maximum value at $x = b$. It is widely used to describe the expansion of some quantity from a more-or-less negligible amount towards a stable, limiting value. As in its original enzyme kinetics context, the functional form of the Michaelis–Menten model can often be motivated theoretically, and it has provided empirically satisfactory descriptions of growth in a wide variety of scientific applications.

To represent the Michaelis–Menten model as a regression by composition, we exploit a connection between rectangular hyperbolae and certain Möbius transformations:

```
model y = LogN(0, 1) | Moebius(0 + 1/x) | Scale(1)
```

We write `LogN(0, 1)` for the standard lognormal distribution, and have already encountered the transformation flow `Scale()` acting on distributions over the real numbers. The `Moebius()` flow—required for our formulation of the Michaelis–Menten model—is also of the transformation type, consisting of a translation sandwiched between reciprocals (see McCullagh, 1996). McCullagh points out close connections between Möbius transformations on real-valued and circular quantities, and indeed this particular group may very helpfully be thought of as acting on distributions over the projectively extended real line $\mathbb{R} \cup \{\infty\}$, which has the topology of a circle. The action of $v \in (\mathbb{R}, +)$ on $y \in \mathbb{R} \cup \{\infty\}$ is

given by

$$y \cdot v = \frac{y}{1 + vy},$$

defining a bijection $\mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ for every $v \in \mathbb{R}$. In particular, $\infty \cdot v = 1/v$, so for $v > 0$ the infinite 'maximum' of the lognormal distribution is mapped to the finite value $1/v$, and hence the resulting distribution has bounded support. For each v, the bijective map $y \mapsto y \cdot v$ is 'monotonic' in the sense that as y moves clockwise around $\mathbb{R} \cup \{\infty\}$, $y \cdot v$ also moves consistently clockwise or anticlockwise in accord with the sign of v.

To see that the Michaelis–Menten model results, consider the passage of the median of p_0 through the flows `Moebius()` and `Scale()`. The median of a standard lognormal distribution is 1, and so the median of the modelled distribution P is

$$(1 \cdot \eta_1) \cdot \eta_2 = \left(\frac{1}{1 + \beta_1/x}\right) \cdot \eta_2 = \frac{\beta_2}{1 + \beta_1/x} = \frac{\beta_2 x}{x + \beta_1}.$$

Here we identify the constant $\beta_1 \in \mathbb{R}$ with the image of the basis vector $1 \in \mathbb{R}$ under the linear map $\beta_1 : \mathbb{R} \to \mathbb{R}$ (and similarly for $\beta_2 \in \mathbb{R}_+$). A notable departure from other regression formulations is that in this case it is the *median* of P, not the mean, that is recognisably of the Michaelis–Menten form. Choosing p_0 to be lognormal is not crucial to the construction, and it could be replaced by any distribution with positive support and unit median.

Some off-the-shelf approaches to fitting Michaelis–Menten models—using `nls()` in R (Pinheiro & Bates, 2009, pp. 520–521), for example—take the dispersion of the modelled distribution to be variation-independent of its mean. Raaijmakers (1987) and Cornish-Bowden (2014) suggest that this assumption may be inappropriate in growth contexts, where typically variation is small (even zero) initially but increases with x. A better option is to fit a generalized linear model using an inverse link function $g : \mu \mapsto 1/\mu$, which implies a Möbius action on the mean μ, and the double-reciprocal structure $g(\mu) = 1/\beta_2 + (\beta_1/\beta_2)/x$ first recognized by Lineweaver and Burk (1934). Employing a gamma distribution then incorporates by default a variance that grows in proportion to the square of the mean. A feature of our regression by composition formulation is that the median and range similarly rise and fall together.

Additional flexibility in describing the variability in y can be achieved by prepending another flow to the existing composition:

```
model y = LogN(0, 1) | Power(1) | Moebius(0 + 1/x) | Scale(1)
```

The `Power()` flow is also of the transformation type, and is induced by the action

$$y \cdot v = y^v$$

of $v \in (\mathbb{R}_+, \times)$ on $y \in \mathbb{R}_+$. Placed here, the `Power()` flow does not affect the modelled median (nor the minimum, nor the maximum), but does alter the concentration of the resulting distribution around its median.

In situations where the initial quantity of the variate of interest is too large to be ignored, a modified Michaelis–Menten equation

$$y = \frac{ax}{x + b} + c$$

for $a > 0, b \in \mathbb{R}$ and $c \in \mathbb{R}$ is a natural generalization, translating the hyperbola vertically so that it passes through $(0, c)$. Appending the `Translate()` flow

```
model y = LogN(0, 1) |
            Power(1) | Moebius(0 + 1/x) | Scale(1) | Translate(1)
```

yields a modified Michaelis–Menten model in which all four linear predictors could be expanded to understand relationships with other measured quantities. Because the linear predictors correspond directly to the usual Michaelis–Menten constants, their estimates represent physically interpretable quantities. Alongide parsimony, parameter interpretability is one of the principal advantages of mechanistically-motivated nonlinear models (Pinheiro & Bates, 2009, p. 274). We use regression by composition to fit modified Michaelis–Menten models in Section 6, where we follow Walters et al. (2024) in applying it to the study of infant growth.

2. A more detailed description of the model

We use the term *regression model* to describe any mathematical formulation for the conditional distribution of a variate of interest given covariates. Let Ω be a measurable space on which random variables representing observations on subjects or experimental units may be defined. (Appendix A describes some of our notational choices.) The variate $Y : \Omega \to \Upsilon$ is sometimes called a response variable, and takes values in a measurable space Υ. Covariates generate a sigma algebra $\mathcal{F}$ on Ω; variate and covariates can in principle be of any type, and could for instance be tuples, images, or stochastic processes. We interpret Y as the variate of interest from a representative observational unit, with $\mathcal{F}$ describing their covariate information. Observations on a sample of units will be used to draw inferences about some aspect of the conditional law of Y given $\mathcal{F}$. Appendix B says more about the fundamental relationship between sampling, statistical modelling and scientific induction.

Denote by $\mathcal{P}$ the affine space of signed laws p on Υ having total measure $p(\Upsilon) = 1$; Appendix C explains this choice, and exhibits some familiar features of affine spaces. The subset $\triangle$ of $\mathcal{P}$ consisting of probability distributions is strictly smaller than $\mathcal{P}$. In admitting *signed* laws, we are acknowledging that beyond $\triangle$ we encounter invalid probability measures: as an example, linear-in-probability models can yield set measures outside the unit interval for some combinations of covariate values (Battey et al., 2019), as can the risk ratio and survival ratio flows of Section 1.

Mathematically, we understand a regression model to be a function $P : \Omega \times \Pi \to \mathcal{P}$, an $\mathcal{F}$-measurable law-valued random variable indexed by a set Π. We interpret P as a

(parameterized, by Π) conditional law of the variate Y given covariates $\mathcal{F}$. Appendix D shows that such interpretations are heritable under arbitrary marginalization: if an $\mathcal{F}$-measurable, law-valued P is the conditional law of Y given $\mathcal{F}$, then for any coarser covariate information $\mathcal{G} \subseteq \mathcal{F}$ and any probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$, the $\mathcal{G}$-measurable, law-valued conditional mean $\overline{P} = \mathbb{P}(P \mid \mathcal{G})$ is the related conditional law of Y given the sub-sigma algebra $\mathcal{G}$, averaged over the measure $\mathbb{P}$.

In regression by composition, we construct our parameterized conditional law P sequentially. Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \cdots$ be an increasing sequence of subsets of the covariate information $\mathcal{F}$, and define associated random variables $X_1, X_2, X_3, \ldots$ describing covariate *contrasts* that take values in vector spaces $\mathbb{U}_1, \mathbb{U}_2, \mathbb{U}_3, \ldots$, whose denizens and dimensions can vary; see, for example, Halmos (1974, pp. 3–4) for a reminder of the formal definition of a vector space. The sequence $X_1, X_2, X_3, \ldots$ is assumed to be *adapted* to $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots$ (see Andersen et al., 1996, pp. 59 sqq.). Writing X for a contrast and $\mathbb{U}$ for its contrast space, a value $X = 0$, or however the identity in $\mathbb{U}$ is denoted, means that covariate information is deemed equal to some global reference value in terms of this particular contrast (Appendix E gives an example). When we use the term *covariate*, we are imagining use cases including both purely predictive modelling and causal inference. Causal inference comes with the usual requirements that structural assumptions are carefully justified and suffice to identify the estimand of interest (see, for example, Chapter 3 of Hernán & Robins, 2023).

To each such covariate contrast space $\mathbb{U}$ is associated a corresponding vector space $\mathbb{V}$ that has a group action on the laws $\mathcal{P}$. For $p \in \mathcal{P}$ and $v \in \mathbb{V}$, these act and are written on the right as $p \cdot v$. Writing v, v' for arbitrary elements of $\mathbb{V}$, and 0 for the identity element thereof, the defining properties of a group action are that (i) for all $p \in \mathcal{P}$, $p \cdot 0 = p$, and (ii) $(p \cdot v) \cdot v' = p \cdot (v + v')$, where $v + v'$ denotes addition in $\mathbb{V}$. Both requirements are natural in a regression context. The first property characterizes an *identity* transformation, permitting us to represent the idea that the conditional law of Y does not vary at all for a particular covariate contrast. The second property captures the idea of homogeneous modifications to conditional distributions, defining a *local* parametric model within the regression pipeline. We shall place one further technical restriction on the group action, and assume that it is *faithful*, meaning that if two elements of $\mathbb{V}$ have the same action on $\mathcal{P}$ then they are, in fact, equal in $\mathbb{V}$. This sidesteps some identifiability problems: imagine the difficulties in estimation that could arise if, for example, $p \cdot v = p$ for all $p \in \mathcal{P}$ and *all* $v \in \mathbb{V}$ (a valid group action, but not a faithful one). Group actions on laws are often more conveniently expressed in terms of densities, distribution functions or generating functions; Appendix F gives details.

Some authors use the term *flow* to describe the actions of vector spaces (usually $\mathbb{R}$) on sets (Halmos, 1956, p. 2), while others call them *dynamical systems* (Brown, 2018, p. 8). We find the *flow* terminology an appealing shorthand (e.g. the translation flow, the Möbius flow), and especially useful for distinguishing two *different* actions of the *same* vector space within the same model. Because it applies to a *single* group of transformations,

our use of the term *flow* diverges from the machine learning literature, where it often describes the passage of a probability distribution through a sequence of differentiable transformations (Papamakarios et al., 2021). There is a an extensive intersection between normalizing flows and regression by composition—both are pipelines of indexed sets of transformations—but differences beyond semantics, too: our apparently severe restriction that transformations to be composed form a *group* pays immediate dividends. Where computing inverses of arbitrary transformations can be difficult or impossible, inverting flow-based transformations $p \mapsto p \cdot v$ amounts simply to replacing the index v with its group inverse $-v$.

The group actions described by flows are central to regression by composition: they express a continuous notion of function iteration (i.e. composition with self), and are themselves composed together to create more complex model classes. Every parameter of a regression by composition is a linear map of the form $\beta : \mathbb{U} \to \mathbb{V}$. The function spaces $\mathcal{L}(\mathbb{U}, \mathbb{V})$ of linear maps from $\mathbb{U}$ to $\mathbb{V}$ parameterize the whole regression by composition; that is $\Pi = \mathcal{L}(\mathbb{U}_1, \mathbb{V}_1) \times \mathcal{L}(\mathbb{U}_2, \mathbb{V}_2) \times \mathcal{L}(\mathbb{U}_3, \mathbb{V}_3) \times \cdots$. Since $\mathbb{U}$ and $\mathbb{V}$ are vector spaces, the linear maps $\mathcal{L}(\mathbb{U}, \mathbb{V})$ also form a vector space. These maps between the contrast spaces of covariates and groups transforming variate distributions are the linear heart of regression by composition. Although it is not strictly necessary that distributional contrasts be parameterized by vector spaces (Nelder, 1965; McCullagh, 2022, p. 177), it is computationally advantageous to differentiate with respect to model parameters (the linear maps $\mathbb{U} \to \mathbb{V}$), and suitably smooth functions between normed vector spaces have accompanying Fréchet derivatives. Importantly for our purposes, the chain rule for functions defined by composition has a natural analogue for Fréchet derivatives. In Section 5, we make extensive use of this fact in constructing and fitting regressions by composition.

The linear maps β (which are to be estimated) are one argument of the linear predictors $\eta : \Omega \times \mathcal{L}(\mathbb{U}, \mathbb{V}) \to \mathbb{V}$, defined by $\eta(\omega, \beta) = \beta(X(\omega)) \in \mathbb{V}$. The actions of the linear predictors η are composed together to create the full regression by composition

$$P = ((p_0 \cdot \eta_1) \cdot \eta_2) \cdot \eta_3 \cdots .$$

This chain or pipeline of group actions has arguments $\omega \in \Omega$ and $(\beta_1, \beta_2, \beta_3, \ldots) \in \Pi$ so, as intended, $P : \Omega \times \Pi \to \mathcal{P}$ is a parameterized, $\mathcal{F}$-measurable, law-valued random variable, and interpretable as a modelled conditional distribution.

3. Invariance, and examples of its statistical implications

We now show how certain algebraic features of flows—their invariants—correspond to important local properties of regression models, deferring to the next section the geometric considerations needed to discuss collapsibility. We also take the opportunity to showcase a handful of versatile flows in order to illustrate these algebraic properties. Our selection of flows is eclectic, and far from comprehensive.

Because the identity element of $\mathbb{V}$ must act on $\mathcal{P}$ as the identity transformation, we have $\mathcal{S} \cdot \mathbb{V} \supseteq \mathcal{S}$ for every subset $\mathcal{S} \subseteq \mathcal{P}$. An *invariant* subset $\mathcal{S}$ of a group action $(\cdot\mathbb{V})$ is one that satisfies the stronger condition $\mathcal{S} \cdot \mathbb{V} = \mathcal{S}$. Trivially, the empty set $\emptyset$ and $\mathcal{P}$ are always invariant subsets. Examples of nontrivial invariant subsets of a group action might include singleton sets, the set $\triangle$ of all probability laws, or a particular parametric model $\mathcal{M} \subset \mathcal{P}$. Some group actions have no nontrivial invariant subsets, some have singleton invariants, others larger invariants; we give examples of each case.

3.1. Invariant points

A law $p \in \mathcal{P}$ is an *invariant point* or *fixed point* of the action of $\mathbb{V}$ if $\{p\} \cdot \mathbb{V} = \{p\}$. Deeks (2002) calls such points *zero constraints* (referring to zero benefit, and zero harm). A good example is the flow that acts multiplicatively on the odds (used in logistic regression), which is "constrained to predict absolute benefits of zero both when the control group event rate is 0 per cent and when it is 100 percent" (Deeks, 2002).

Invariant points are usually extreme, and so likely to be approached only at extreme covariate values. Nevertheless, they can indicate a certain rigidity near the fixed point that may have important consequences. As alluded to in Section 1.2, it is difficult for logistic regression (with two fixed points) to capture the effect of an intervention that is of substantial benefit for people at high risk of experiencing the event of interest, but is of diminishing value as the untreated risk decreases towards zero. By contrast, a log-linear or relative risk model (having only one invariant point) describes such an effect in a straightforward manner. The analyses in Section 7 employ several models with no fixed points.

Invariant points can also have implications for the assessment of model uncertainty. Confidence intervals (or credible intervals, or likelihood intervals) around predictions located well away from the bulk of the data can be apparently precise or vague depending on whether there is, or is not, assumed to be an invariant point in the vicinity of the prediction being made (see Daniel et al., 2024, for several examples, and Figure 5 for another).

Convolution is a class of flows without any invariant points. Let $\mathbb{V}$ be a vector space of characteristic functions with group operation piecewise multiplication, corresponding to addition of the implied random variables. Writing Υ' for the continuous dual of Υ, $\mathbb{V}$ acts on characteristic functions $\phi : \Upsilon' \to \mathbb{C}$ as

$$(\phi \cdot v)(t) = \phi(t)v(t),$$

adding to the input random variable Y (whose characteristic function is ϕ) an independent random variable with characteristic function v. Only the identity element $v = (t \mapsto 1)$ leaves any elements unchanged following application of the corresponding transformation, so convolution has no fixed points.

Rather than convolving an independent random variable with the input, we can instead add (or *compound*) together independent realizations from the input distribution itself.

The number of copies can be drawn from a probability distribution, and need not be a natural number. Using the well-known result that the characteristic function of the sum of a random number of random variables is a probability-generating (or, more generally, factorial moment-generating) function G_v of the characteristic function ϕ of the summands, we see that a compounding flow will take the form

$$(\phi \cdot v)(t) = G_v(\phi(t))$$

for $t \in \Upsilon'$ and $v \in \mathbb{V}$. In order to respect the defining flow property $(\phi \cdot v) \cdot v' = \phi \cdot (v + v')$, the family of compounding distributions indexed by $v \in \mathbb{V}$ must itself be *closed under compounding*, satisfying $G_{v'}(G_v(\cdot)) = G_{v+v'}(\cdot)$. In general, compounding flows have a single invariant point at random variables that equal zero almost surely, corresponding to the characteristic function $\phi = (t \mapsto 1)$.

Consider first deterministic compounding: that is, compounding by the family of constant random variables with factorial moment-generating functions $\{t \mapsto t^v : v \in (\mathbb{R}_+, \times)\}$. The vector space $(\mathbb{R}_+, \times)$ acts on characteristic functions ϕ by compounding as

$$(\phi \cdot v)(t) = (\phi(t))^v,$$

and canonical Poisson regression (Nelder, 1974) arises from compounding a Poisson distribution in this way:

```
model y = Poi(1) | Comp(1 + trt)
```

To describe a related negative binomial regression (Lawless, 1987), we may begin with a geometric distribution 'counting failures' with support $\{0, 1, 2, \ldots\}$ and unit mean, compound this with a geometric distribution 'counting trials' with support $\{1, 2, 3, \ldots\}$ and success probability $v \in (\mathbb{R}_+, \times)$, and finally compound deterministically as before:

```
model y = Geom(1/2) | CompGeom(1 + trt) | Comp(1 + trt)
```

A feature of this model is that the full two-parameter flexibility of the negative binomial distribution is made amenable to regression modelling. This contrasts with flavours of negative binomial regression that instead extend generalized linear models by allowing the estimation of a scalar dispersion parameter (Venables & Ripley, 2002, pp. 206 sqq.).

Like the geometric distribution, Bernoulli distributions are also closed under compounding. Subsequent compounding by a Bernoulli distribution with success probability $v \in (\mathbb{R}_+, \times)$ corresponds to zero-inflation (Lambert, 1992), with an additional point mass $1 - v$ placed at zero:

```
model y = Geom(1/2) |
            CompGeom(1 + trt) | Comp(1 + trt) | CompBer(1 + trt)
```

Such a regression is analogous to a two-part or *hurdle* model, in which group-specific proportions of excess zeros could be estimated.

3.2. Closure on probability laws

It can also be helpful to know if a flow is invariant on $\triangle \subset \mathcal{P}$, the set of laws $p \in \mathcal{P}$ that are *probability* laws (satisfying $p(A) \in [0, 1]$ for all events $A \subseteq \Upsilon$). A group action of $\mathbb{V}$ for which $\triangle$ is an invariant subset cannot turn valid probabilities into invalid ones for any combination of covariates, even extreme covariate values beyond those present in the data used to fit the model. If all flows in a regression by composition are *closed* in this sense, then all modelled conditional distributions are guaranteed to be valid.

Many—perhaps most—commonly-used statistical models have such global invariance properties. For example, transformation models (Hothorn et al., 2014) are necessarily invariant on $\triangle$. The predominance of logistic regression over linear probability models, or of the Cox (1972) relative risk model over the Aalen (1989) additive hazards model, can in part be explained by an understandable preference among the statistical community for closed flows. For a suggested counterargument with particular reference to binary regression, see Hellevik (2009).

A broad class of flows (including the one giving rise to the Aalen model) for which $\triangle$ is *not* an invariant subset are those realized by taking the maximum or minimum of the input random variable and another, independent, random variable. For example, let $\mathbb{V}$ be a vector space of functions from the domain Υ of Y (with Υ here being assumed totally ordered) to the positive reals, with group operation pointwise multiplication. The flow that acts on cumulative distribution functions $F : \Upsilon \to [0, 1]$ as

$$(F \cdot v)(y) = F(y)v(y)$$

for $y \in \Upsilon$ and $v \in \mathbb{V}$ returns the cumulative distribution function of the maximum of two independent random variables whose respective cumulative distribution functions are F and v. The output $F \cdot v$ is certain to be a valid cumulative distribution function if v is a valid cumulative distribution function, but otherwise no guarantees are available. Any v that is a valid cumulative distribution function has an inverse element $y \mapsto 1/v(y)$ that is invalid, with the identity $(y \mapsto 1) \in \mathbb{V}$ being the only exception. If v is not a valid cumulative distribution function, the transformed cumulative distribution function $F \cdot v$ may be nonincreasing, or range outside the unit interval, leading to invalid probabilities.

Sometimes invariance on a subset of $\triangle$ may also be of interest. Consider for example the affine space $\mathcal{P}$ of signed measures on $\Upsilon = \mathbb{R}$: the translation flow given by $(F \cdot v)(y) = F(y - v)$ is invariant on the probability laws $\triangle \subset \mathcal{P}$, but not on the subsimplex of probability laws with support only on $(0, \infty)$. When modelling strictly positive-valued quantities (height, weight, blood pressure, etc.) it may be undesirable to place any support on negative values, so this kind of 'nonclosure' can be relevant to scientific inference, even though it does not result in invalid likelihoods.

3.3. Parametric invariances

Some parametric families $\mathcal{M}$ are invariant to the action of certain flows. The set of Weibull distributions is invariant to the action of scaling cumulative hazard functions

$(\Lambda \cdot v)(y) = \Lambda(y/v)$ by $v \in \mathbb{R}_+$ and also powering $(\Lambda \cdot v)(y) = \Lambda(y^{1/v})$ for $v \in \mathbb{R}_+$. Looked at in reverse, these parametric invariances show us how to construct Weibull-type regressions: since a regression by composition of the form

```
model y = Exp(1) | Power(1) | Scale(1)
```

produces the usual two-parameter Weibull distribution from an exponential p_0, variants like

```
model y = Exp(1) | Power(1 + age) | Scale(1 + trt)
```

produce Weibull conditional distributions whose parameters depend on covariates. Unlike in standard implementations of Weibull regression (`survreg` in the `survival` package for R, for example)—but like the mean-and-variance model of Section 1.1, and the negative binomial model of Section 3.1—here again *both* model parameters can be connected to covariates via linear predictors.

More generally, regression by composition provides a compact characterization of existing parametric families of distributions, and the flexibility to construct new ones. The model space $\mathcal{M} \subseteq \mathcal{P}$ is the set of laws reachable from p_0 through the prescribed actions of vector spaces $\mathbb{V}$: that is,

$$\mathcal{M} = ((p_0 \cdot \mathbb{V}_1) \cdot \mathbb{V}_2) \cdot \mathbb{V}_3 \cdots .$$

For example, the class of normal distributions parameterized by their standard deviation and mean is precisely the set $(p_0 \cdot \mathbb{R}_+) \cdot \mathbb{R}$, where p_0 is the standard normal law, $\mathbb{R}_+$ acts by scaling, and $\mathbb{R}$ acts by translation. Different parametric families result if we replace p_0 by another law; if p_0 is the uniform distribution on the interval $[0, 1]$, then $\mathcal{M}$ is a collection of uniform distributions indexed by interval width and minimum.

If the scaling and translation actions are permuted, we obtain parameterized families of normal and uniform distributions whose locations and dispersions are not variation independent (Dawid, 2001). In other words, scaling $(\cdot\mathbb{R}_+)$ and translation $(\cdot\mathbb{R})$ fail to *commute strongly*: it is not the case that $(\cdot v) \cdot v' = (\cdot v') \cdot v$ for all $v \in \mathbb{R}_+$ and $v' \in \mathbb{R}$. However, when (as here) flows can be interchanged so that $(\cdot\mathbb{R}_+) \cdot \mathbb{R} = (\cdot\mathbb{R}) \cdot \mathbb{R}_+$ without changing the set of implied transformations of $\mathcal{P}$, we say that they *commute weakly*.

3.4. Transitivity

If fixed points describe a lack of model flexibility, the other extreme arises when any (or almost any) law can be reached from any (or almost any) input. A group action is *transitive* on an invariant set $\mathcal{S}$ if, for any $p, q \in \mathcal{S}$, there exists $v \in \mathbb{V}$ such that $p = q \cdot v$. By extension, we shall say that a group action is transitive on the (not necessarily invariant) set $\triangle$ of probability laws if it is transitive on an invariant set $\mathcal{S}$ containing $\triangle$, and *almost transitive* on $\triangle$ if it is transitive on an invariant set $\mathcal{S}$ for which $\triangle \setminus \mathcal{S}$ is a null set; this caveat makes allowance for a countable number of invariant points. Transitivity on $\triangle$ therefore characterizes a flow as being locally 'nonparametric'.

Transitivity is closely related to a flow's invariances (McCullagh, 2022, p. 238). The *orbit* $p \cdot \mathbb{V} = \{p \cdot v : v \in \mathbb{V}\}$ of every law p must be (a) nonempty and (b) an invariant subset, so if $\mathcal{P}$ has no nontrivial invariant subsets, then the orbit of every element of $\mathcal{P}$ must be $\mathcal{P}$. The risk difference flow is transitive on $\mathcal{P}$ and hence on $\triangle$, while the odds ratio, risk ratio and survival ratio flows are almost transitive on $\triangle$. An example of a flow acting on Bernoulli distributions that is neither transitive nor almost transitive on $\triangle$ follows Theorem 4.2. For real-valued variates of interest, translation is a simple example of a flow that is not transitive in general: for example, only normal distributions with the same variance can be reached by translating a given normal distribution.

4. Collapsibility

4.1. Motivation and notation

In observational studies, regression models are often used to 'adjust' for confounders. For example, suppose age is the only common cause of both treatment choice and a variate of interest Y. A simple contrast of the distributions of Y in the treated and untreated populations will not be causally interpretable: between-group differences may be attributable not only to the effects of treatment, but also to systematic differences in age distributions, because one or other treatment may be preferentially selected by some age groups. However, under the stated, strong, assumptions, distributional comparisons between treatment groups within age strata *are* causally meaningful, and a regression model of Y on age and treatment group can then be used to estimate these age-specific (conditional) treatment effects.

Regression by composition offers one way to describe conditional distributional comparisons. For a treatment effect modelled by composition, we can use the associated flow to compare conditional laws P_0 (for the untreated) and P_1 (for the treated) since, by construction, they share the same orbit, whatever the value of the confounders. Let us describe such a comparison using a binary 'subtraction' operator $\ominus : \mathcal{P} \times \mathcal{P} \to \mathbb{V}$, satisfying $p \ominus q = v$ if and only if $p = q \cdot v$. The operation $p \ominus q$ is very much like subtraction in an affine space, as in Appendix C, and is well-defined for any $p, q \in \mathcal{P}$ belonging to the same principal homogenous space of $\mathbb{V}$ (an orbit over which $\mathbb{V}$ acts freely and transitively).

We might also want to compute and compare marginal laws $\overline{P_0} = \mathbb{P}(P_0)$ (for the untreated) and $\overline{P_1} = \mathbb{P}(P_1)$ (for the treated), averaging out confounders over some relevant measure $\mathbb{P}$. Unfortunately, the marginal laws $\overline{P_0}$ and $\overline{P_1}$ typically do not belong to the same orbit, so an extension of the comparison $\ominus$ is needed. This problem is the same one addressed by Kalbfleisch and Prentice (1981, for relative risk models) through averages of well-defined local comparisons, by Greenland et al. (1999, for generalized linear models) through consideration of probability limits in misspecified regression models, and by Vansteelandt and Dukes (2022, again for generalized linear models) by judicious weighting of conditional contrasts.

4.2. Characterizing collapsible contrasts

We take a different approach, and extend the definition of $\ominus$ by introducing a 'summary' function $p \mapsto [p]$ that simplifies (often substantially) the information content of p while respecting the way in which $\mathbb{V}$ acts on $\mathcal{P}$. Specifically, we require that the map $\mathcal{P} \to [\mathcal{P}]$ is an affine *morphism of vector space actions*, or *morphism of* $\mathbb{V}$*-sets*. In the language of category theory (Mac Lane, 1971), the structure-preserving map $p \mapsto [p]$ is a natural transformation (thinking of $\mathbb{V}$-sets as functors), of which the crucial consequence for our purposes is that $p \mapsto [p]$ commutes with the group action; that is, $[p \cdot v] = [p] \cdot v$ for all $p \in \mathcal{P}$ and all $v \in \mathbb{V}$. This allows us to extend the definition of $\ominus$ to a much broader class of comparisons, for we can now write $p \ominus q = v$ if and only if $[p] = [q] \cdot v$. Our extended definition of $p \ominus q$ is well-defined under the weaker condition that $[p], [q] \in [\mathcal{P}]$ belong to the same principal homogeneous space of $\mathbb{V}$. It is still the case that if $p = q \cdot v$ (that is, if p and q are in the same orbit in $\mathcal{P}$) then $[p] = [q \cdot v] = [q] \cdot v$ and so $p \ominus q = v$, and in particular $p \ominus p = 0$ for all $p \in \mathcal{P}$. The square bracket notation $[p]$ is intentionally suggestive of an equivalence class, with p and q deemed equivalent if $[p] = [q]$. If odds or cumulant generating functions are used to express a group action on $\mathcal{P}$, a degree of care is needed to ensure that its affine structure is respected (Appendix F).

With this extended definition of $\ominus$, we can ask whether the comparison $\overline{P_1} \ominus \overline{P_0}$ of the marginal laws lies within the convex hull $\langle P_1 \ominus P_0 \rangle = \mathrm{Conv}(P_1 \ominus P_0)$ of $P_1 \ominus P_0$; $\langle P_1 \ominus P_0 \rangle$ is the smallest convex subset of $\mathbb{V}$ in which $P_1 \ominus P_0$ is guaranteed to fall. For this question to be meaningful, $[P_0]$ and $[P_1]$ should belong to a fixed, convex, principal homogenous space, for then so also do $\overline{[P_0]} = [\overline{P_0}]$ and $\overline{[P_1]} = [\overline{P_1}]$, making $\overline{P_1} \ominus \overline{P_0}$ well-defined. We describe the comparison $P_1 \ominus P_0$ as *weakly collapsible* if $\overline{P_1} \ominus \overline{P_0} \in \langle P_1 \ominus P_0 \rangle$, and *strongly collapsible* if $\overline{P_1} \ominus \overline{P_0} = \overline{P_1 \ominus P_0}$. Neither kind of collapse can be guaranteed in general. Strong collapsibility implies weak collapsibility, because the mean $\overline{P_1 \ominus P_0} = \mathbb{P}(P_1 \ominus P_0)$ must lie in the convex hull $\langle P_1 \ominus P_0 \rangle$. If $P_1 \ominus P_0$ happens to be a constant ($\mathbb{V}$-valued) random variable, then $\langle P_1 \ominus P_0 \rangle = \{P_1 \ominus P_0\} = \{\overline{P_1 \ominus P_0}\}$, making weak and strong collapsibility equivalent conditions (although of course still not guaranteed). Greenland et al. (1999) call this special case *strict collapsibility*. Historically, the term *collapsible* referred to the collapsing of contingency tables (Simpson, 1951), in which confounding can be a dramatic cause of noncollapsibility; here we use it in a more precise sense, similar in spirit to definitions from causal inference, where the presence or absence of confounding does not affect collapsibility.

THEOREM 4.1. *The comparison* $\ominus$ *is strongly collapsible if the group action* $(\cdot) : [\mathcal{P}] \times \mathbb{V} \to [\mathcal{P}]$ *is affine.*

Proof. By definition $[P_1] = [P_0] \cdot (P_1 \ominus P_0)$, so $\overline{[P_1]} = \overline{[P_0] \cdot (P_1 \ominus P_0)}$. Because $p \mapsto [p]$ is affine, we have $\overline{[P_1]} = [\overline{P_1}]$ and, because $(\cdot) : [\mathcal{P}] \times \mathbb{V} \to [\mathcal{P}]$ is affine, we have

$$[\overline{P_1}] = \overline{[P_0] \cdot (P_1 \ominus P_0)} = \overline{[P_0]} \cdot (\overline{P_1 \ominus P_0}) = [\overline{P_0}] \cdot (\overline{P_1 \ominus P_0}).$$

□

Examples of strongly collapsible comparisons include differences in arithmetic moments (including arithmetic means), ratios of geometric moments (including geometric means), and differences in survivor and cumulative distribution functions; Section 4.3 gives a concrete illustration.

THEOREM 4.2. *The comparison* $\ominus$ *is weakly collapsible if, for each* $v \in \mathbb{V}$*, the transformation* $(\cdot v) : [\mathcal{P}] \to [\mathcal{P}]$ *is affine.*

Proof. The proof is very similar to the previous one. Once again we have by definition that $[P_1] = [P_0] \cdot (P_1 \ominus P_0)$, and we know that $[P_0] \in \langle [P_0] \rangle$ and $P_1 \ominus P_0 \in \langle P_1 \ominus P_0 \rangle$ $\mathbb{P}$-almost surely. The set

$$\{([p], [p] \cdot v) : [p] \in \langle [P_0] \rangle, v \in \langle P_1 \ominus P_0 \rangle\}$$

is convex: it is a dense union of nonintersecting line segments (dense by convexity of $\langle P_1 - P_0 \rangle$, nonintersecting because $\langle [P_0] \rangle$ is limited to a single principal homogenous space, lines because each transformation $(\cdot v)$ is affine). So we can apply Lemma G.1 (which is a slightly more general version of Jensen's inequality) and conclude that

$$[\overline{P_1}] = \overline{[P_1]} = \overline{[P_0] \cdot (P_1 \ominus P_0)} \in \overline{[P_0]} \cdot \langle P_1 \ominus P_0 \rangle = [\overline{P_0}] \cdot \langle P_1 \ominus P_0 \rangle.$$

□

Selected examples of weakly collapsible comparisons include, necessarily, all strongly collapsible comparisons, together with risk ratios, rate ratios and ratios of survivor and cumulative distribution functions. Two such cases are discussed further in the next section. The requirement that $[P_0]$ and $[P_1]$ belong to a single principal homogeneous space is not vacuous: the flow acting on Bernoulli distributions given by $p \cdot v = (1 - v)/2 + pv$ for $v \in (\mathbb{R}_+, \times)$ has a fixed point at $p = 1/2$ and *two* principal homogeneous spaces, making half of all possible distributional comparisons undefined.

Because they must be mutually-consistent, 'averageable' quantities, we only consider collapsibility within the restricted class of contrasts taking values in vector spaces that act on distributional summaries $[p]$. However, it is not necessary that the comparison $\ominus$ and its associated group action of $\mathbb{V}$ on $[\mathcal{P}]$ should relate to a flow used in modelling the conditional law P. The conclusions of Theorem 4.1 and Theorem 4.2 are properties of the comparison $\ominus$, and quite independent of any modelling decisions that may have been taken in the construction of P_0 and P_1.

A further implication of Lemma G.1 is that, if marginal and conditional causal effects are respectively identified by marginal and conditional contrasts, it is not possible for conditional contrasts all to have one sign and the marginal contrast to have the other. This would be the case, for example, in a study in which receipt of treatment is randomized, and is true even for noncollapsible comparisons. More generally, suppose a pair $([P_0], [P_1])$ of conditional distributions or summaries thereof belongs almost surely to some (possibly nonconvex)

set contained in a halfspace of $[\mathcal{P}]^2$ bounded by a hyperplane in which lies the identity line. Because the halfspace is convex, the average $(\overline{[P_0]}, \overline{[P_1]}) = ([\overline{P_0}], [\overline{P_1}])$ necessarily belongs to the same halfspace. Thus, of the two sources of Simpson's 'paradox'—noncausal comparisons and noncollapsibility (Hernán et al., 2011)—only noncausal comparisons can produce sign reversal, or indeed its higher-dimensional equivalents.

4.3. Survival examples

Taking as our context the analysis of survival data, we now give three examples of contrasts whose collapsibility or otherwise has given (us) pause for thought (Daniel et al., 2021; Didelez & Stensrud, 2022). The first is weakly collapsible, the next strongly so, and the last not even weakly collapsible.

There is a justified hesitancy about the conditioning on survival implicit in hazard-based models (Hernán, 2010). An important first step is to make explicit how hazard-based contrasts should be extended to compare survival curves not sharing an orbit. For instance, let us express the Aalen flow in terms of survivor functions as

$$(S \cdot v)(y) = S(y)v^y$$

for all $y \geq 0$ and $v \in (\mathbb{R}_+, \times)$. The logarithm $\log v$ of the flow index v is precisely the hazard difference, here assumed constant through time. There are many possible ways to extend Aalen-type contrasts while respecting the action of $\mathbb{R}_+$ on $\mathcal{P}$: for any time $\tau > 0$, we may set $[S] = S(\tau)$ and specify $[S] \cdot v = [S]v^\tau$, defining an affine morphism of $\mathbb{V}$-sets. The transformation $(\cdot v) : \mathbb{R} \to \mathbb{R}$ given by $[S] \mapsto [S]v^\tau$ is affine for every $v \in \mathbb{R}_+$, so this implied contrast associated with the Aalen model is weakly (but not strongly) collapsible.

Now consider two seemingly equivalent survival models, both expressed as simple one-flow regressions by composition. In either case, we begin with a 'standard' exponential distribution whose survivor function satisfies $S_0(y) = \exp(-y)$ for $y \geq 0$. The first model is of the accelerated-failure type (Wei, 1992)

```
model y = Exp(1) | Accelerate(1 + age * trt)
```

where the `Accelerate()` flow given by $(S \cdot v)(y) = S(vy)$ is just a trivially inverted version of `Scale()`, while the second model has a relative-risk form (Cox, 1972)

```
model y = Exp(1) | PowerSurv(1 + age * trt)
```

in which `PowerSurv()` acts on survivor functions as $(S \cdot v)(y) = (S(y))^v$. (Both group actions can, of course, equivalently be characterized in terms of hazard functions, as in McCullagh, 2022, p. 238.) Despite employing radically different flows, the likelihood functions implied by the two models are identical in every respect: as is well known, conditional treatment effects for Weibull (and hence exponential) distributions have *both* an acceleration factor and a hazard ratio interpretation (Kalbfleisch & Prentice, 2002, p. 42). Nevertheless, if we consider attempting to collapse conditional treatment effects $P_1 \ominus P_0$

over a given age distribution, we find that the former is strongly collapsible while the latter is not collapsible at all.

To see this, consider first the flow `Accelerate()`, which is induced by the data transformation $y \cdot v = y/v$. Since scaling is a 'multiplicative translation', we can conveniently summarize a law p by its geometric mean $[p] = p(Y;\times)$ which is, like its arithmetic counterpart, an affine summary $\mathcal{P} \to (\mathbb{R}_+, \times)$. For $p \in \mathcal{P}$ and $v \in (\mathbb{R}_+, \times)$, we have $[p \cdot v] = (p \cdot v)(Y;\times) = p(Y/v;\times) = p(Y;\times)/v = [p]/v$, the last of which we may take as our definition of $[p] \cdot v$. Since the map $\mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ given by $([p], v) \mapsto [p]/v$ is affine—that is to say, in both arguments simultaneously—we conclude that the implied comparison is strongly collapsible.

By contrast, the flow `PowerSurv()`, which is not so naturally expressible as a transformation model, may be extended to a comparison $\ominus : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ via a choice similar to the one available for the Aalen flow: set $[S] = S(\tau)$ and define $[S] \cdot v = [S]^v$. The map $S \mapsto [S]$ is an affine, natural transformation of $\mathbb{R}_+$-sets but, of the transformations $(\cdot v) : [S] \mapsto [S]^v$, only the identity map is affine, and we conclude that hazard ratios are in general not even weakly collapsible.

More concretely, suppose that the (modelled, say) conditional contrast

$$P_1 \ominus P_0 = P_1(Y;\times)/P_0(Y;\times) = \log(S_1(\tau))/\log(S_0(\tau))$$

was $\mathbb{P}$-almost surely equal to 2, in other words that the same acceleration factor or hazard ratio applied across all age groups. Averaging distributions across the age groups, we may deduce that $\overline{P_1}(Y;\times)/\overline{P_0}(Y;\times) = 2$, but in general $\log(\overline{S_1}(\tau))/\log(\overline{S_0}(\tau)) \neq 2$.

4.4. Generalized linear models

Generalized linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) are central to much applied statistical practice; Dobson and Barnett (2018) offer many examples. An immediate advantage of framing generalized linear models as regressions by composition is that dispersion parameters are incorporated in the same framework.

A binary generalized linear model with link function g can be written in terms of the flow

$$p \cdot v = g^{-1}(g(p) + v)$$

in which $v \in (\mathbb{R}, +)$ acts on probabilities $p = p(Y = 1)$. Daniel et al. (2021) call the transformation $(\cdot v)$ the *characteristic collapsibility function*; its unconstrain-shift-reconstrain structure clearly has the required flow properties, and dates back at least to Abel (1826). Indeed, McCullagh (1999) describes the flow structure underlying all generalized linear models as "Abelian". He also points out that the baseline law p_0 in a generalized linear model $P = p_0 \cdot \eta_1$ is a *choice* for the modeller to make, and need not come from an exponential family, since exponential tilting (Esscher, 1932) may be defined for a much broader class of distributions: Luo and Tsai (2012) use the term *proportional likelihood ratio models*. The flow corresponding to exponential tilting (and hence to canonical-link

generalized linear models) is easily expressed in terms of moment-generating functions $M : \Upsilon' \to \mathbb{R}$: for $v \in (\mathbb{R}, +)$,

$$(M \cdot v)(t) = M(t + v)/M(v).$$

McCullagh (1999) provides details that allow this flow to be modified for non-canonical link functions.

To study the collapsibility of the contrasts implied by generalized linear models for other types of variates of interest, we may take the arithmetic mean $[p] = p(Y)$ as our summary measure. Contrasts collapse strongly when the flow given by

$$[p] \cdot v = g^{-1}(g([p]) + v)$$

is affine in the pair $([p], v)$; this occurs only when g is the identity link. Weakly collapsible contrasts arise when $(\cdot v)$ is affine, which (of the commonly-used link functions) happens only if g is the log or complementary log link, as shown by Gail et al. (1984) and Neuhaus and Jewell (1993).

5. Fitting regressions by composition

Regression by composition is amenable to likelihood-based modes of statistical inference. Here we describe a frequentist approach, but pure-likelihood or Bayesian perspectives are also feasible. Likelihoods built by composition will not typically be expressible in closed form; nevertheless, from modular components (primitive models, and a library of flows), users can automate assembly and analysis for regression models of arbitrary complexity. The versatility of regression by composition comes from flows having a common interface, allowing them to be connected together in different ways without needing to expose the modeller to the details of likelihood construction.

Our primary tool for model building is known in functional programming as a *fold* (Hutton, 1999), accepting as arguments an aggregating function, an initial value, and a list, ultimately returning an object of the same type as the initial value. For us, the initial value is a zero-parameter model m_0 (returning p_0), and the list is a finite sequence of flows $((\cdot), (\cdot), (\cdot), \ldots)$. The aggregating function α takes as inputs a model with zero or more parameters (a function $\mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots \to \mathcal{P}$) and a flow (a function $\mathcal{P} \times \mathbb{V} \to \mathcal{P}$), outputting an enhanced model with an additional parameter corresponding to the flow's index space $\mathbb{V}$; that is

$$\alpha : (\mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots \to \mathcal{P}) \times (\mathcal{P} \times \mathbb{V} \to \mathcal{P}) \to (\mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots \times \mathbb{V} \to \mathcal{P}),$$

having explicit definition

$$\alpha : (m, (\cdot)) \mapsto ((v_1, v_2, v_3, \ldots, v) \mapsto m(v_1, v_2, v_3, \ldots) \cdot v).$$

The zero-parameter model $m_0 : \{()\} \to \mathcal{P}$ with $m_0 : () \mapsto p_0$ is a function from the set containing the empty tuple $()$ to $\mathcal{P}$. The Fold operator recursively applies α to 'the

previous model' (starting with m_0) and 'the next flow', where at each stage of aggregation another parameter is added to the model. The final model $M : \mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots \to \mathcal{P}$ is

$$M = \cdots \alpha(\alpha(\alpha(m_0, (\cdot)), (\cdot)), (\cdot)) \cdots = \mathrm{Fold}(\alpha, m_0, ((\cdot), (\cdot), (\cdot), \ldots)),$$

and we call such a construction a *composite parametric family*.

Fréchet derivatives (see Appendix C) ∇M of the model M with respect to its vector-valued parameters may likewise be computed iteratively. Each flow has two relevant partial derivatives (corresponding to its $\mathcal{P}$- and $\mathbb{V}$-valued arguments), which we write symbolically as $(\cdot)' : \mathcal{P} \times \mathbb{V} \to \mathcal{L}(\mathcal{Q}, \mathcal{Q})$ and $\nabla(\cdot) : \mathcal{P} \times \mathbb{V} \to \mathcal{L}(\mathbb{V}, \mathcal{Q})$, respectively. Flow derivatives $(\cdot)'$ and $\nabla(\cdot)$ can often be evaluated analytically, and are especially straightforward when the maps $(\cdot v)$ or $(p \cdot)$ are affine. Both partial flow derivatives are needed to aggregate model derivatives: the total derivative $\nabla(\alpha(m, (\cdot)))$, taking values in $\mathcal{L}(\mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots \times \mathbb{V}, \mathcal{Q})$, is expressible as the direct sum

$$\nabla(\alpha(m, (\cdot))) = ((m \cdot)' \circ \nabla m) \oplus \nabla(m \cdot)$$

of two partial derivatives (with respect to its $\mathbb{V}_1 \times \mathbb{V}_2 \times \mathbb{V}_3 \times \cdots$- and $\mathbb{V}$-valued arguments), the first of these following from an application of the chain rule. The updated model derivative $\nabla(\alpha(m, (\cdot)))$ thus described in terms of its predecessor ∇m, m itself, and the partial derivatives $(\cdot)', \nabla(\cdot)$ of the new flow, we may again use the Fold operator to recursively calculate the global model derivative ∇M. For efficiency, both recursions (defining M and ∇M) may be subsumed into a single fold, making use of the operator's so-called 'banana split' property (Hutton, 1999).

The vector-valued parameters of a composite parametric family M enter a regression by composition P indirectly, as outputs of corresponding linear predictors $\eta_1, \eta_2, \eta_3, \ldots$; that is

$$P = M \circ (\eta_1 \oplus \eta_2 \oplus \eta_3 \oplus \cdots).$$

A further application of the chain rule yields the derivative ∇P of the regression by composition P with respect to $\Pi = \mathcal{L}(\mathbb{U}_1, \mathbb{V}_1) \times \mathcal{L}(\mathbb{U}_2, \mathbb{V}_2) \times \mathcal{L}(\mathbb{U}_3, \mathbb{V}_3) \times \cdots$:

$$\nabla P = \nabla M(\eta_1 \oplus \eta_2 \oplus \eta_3 \oplus \cdots) \circ \nabla(\eta_1 \oplus \eta_2 \oplus \eta_3 \oplus \cdots),$$

the second term simplifying to $\nabla\eta_1 \oplus \nabla\eta_2 \oplus \nabla\eta_3 \oplus \cdots$. Since each η is (by design) a linear map, we have $\nabla\eta : \beta \mapsto \eta$; the constant derivative $\nabla\eta$ is analogous to the factors $\partial\eta/\partial\beta_j = x_j$ in the computation of likelihoods for generalized linear models (McCullagh & Nelder, 1989, p. 41).

The simplest approach to likelihood evaluation is to encode and update information about laws directly in terms of log-densities $\log \mathrm{d}p/\mathrm{d}q : \Upsilon \to \mathbb{R}$, computed relative to some suitable reference measure q. However, not all flows can easily be represented in terms of their action on log-densities. If moment-generating functions or characteristic functions are used to represent laws in a fitting procedure, then a likelihood conversion is needed, and amounts to inversion of Laplace or Fourier transforms, perhaps by way of

a saddlepoint approximation (Daniels, 1954). At present, the R package `rbc` uses direct manipulation of log-densities to implement continuous transformation flows, and binary regressions by composition.

6. Modelling infant growth

Our first worked example concerns the measurement and modelling of infant growth. Antenatal monitoring of child health through regular recording of height and weight allows assessment of an individual's current measurements relative to previous observations, to model-based predictions, or to standard reference populations (Cole, 1988). Where large deviations from expected growth are observed, further investigations or interventions may be warranted.

6.1. STORK data

Walters et al. (2024) advocate the use of modified Michaelis–Menten models

$$y = \frac{ax}{x + b} + c$$

(where $a > 0, b \in \mathbb{R}$ and $c \in \mathbb{R}$) for smoothing and interpolating the weight or height (y) of children from birth to three years of age (x). They employ the R function `nls()` to fit individual growth curves by nonlinear least squares. Using the same data arising from the STORK study (Ley et al., 2016), we compare their nonlinear least squares fits to the comparator regression by composition specified in Section 1.3. While the two models both have conditional medians (given age) of the modified Michaelis–Menten form, they lead to rather different variance structures. As its name implies, nonlinear least squares minimizes squared deviations from the median curve, and in particular implicitly assumes that these deviations are of equal importance at every age. A consequence is that the choice of time origin is irrelevant for nonlinear least squares estimation of Michaelis–Menten models.

By contrast, at $x = 0$ our regression by composition specification necessarily implies a degenerate distribution (all mass being concentrated at $y = c$), with variance growing with increasing x. Therefore choice of time origin *is* an important consideration for this flavour of Michaelis–Menten model. Given the microscopic size of a fertilized ovum, conception seems a defensible origin, as for example in Watkins et al. (2020). In the absence of more specific gestational information, we take as our explanatory variable x the child's age plus 38 weeks, approximately the average time from conception to birth in humans. (The conventional 40 weeks average gestation is timed from the first day of the last menstrual period.)

Nonlinear least squares optimization can be sensitive to starting values of the parameters. We reproduce the fits of Walters et al. (2024) to the STORK data by first noting that modified Michaelis–Menten models would be ordinary linear models if the parameter b were known. One-dimensional profile likelihood may be used to estimate b (starting iteration at $b = 0$)

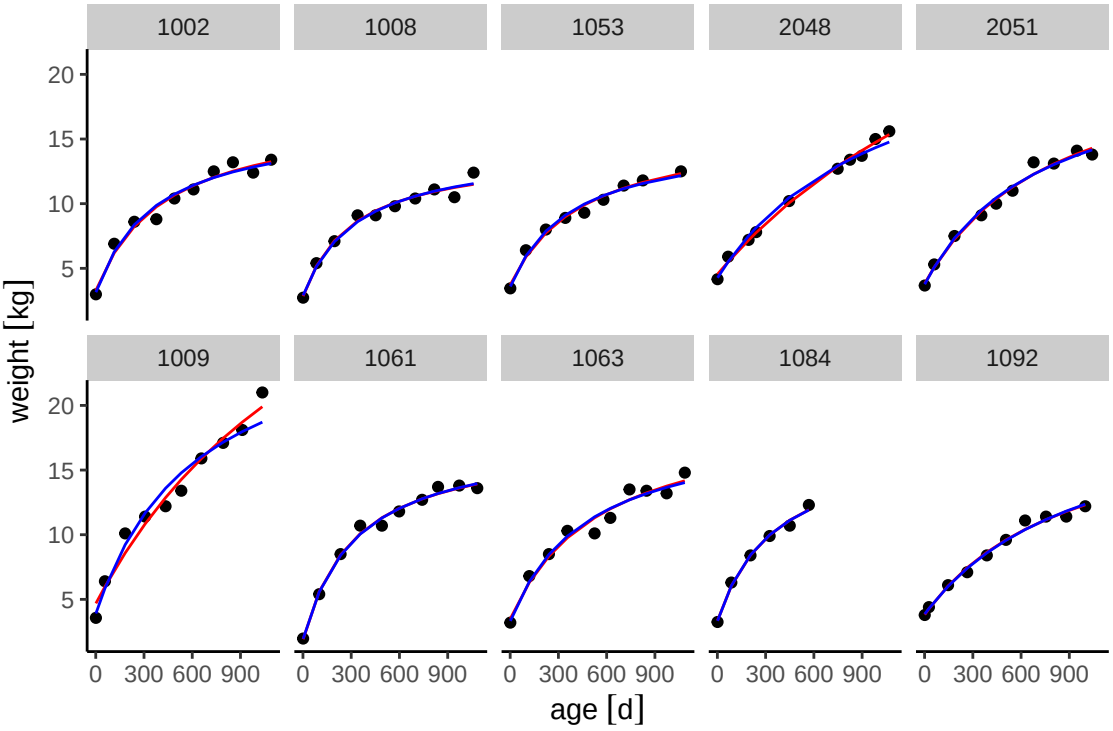

Fig. 1: Reproduction of Figure 1A–B from Walters et al. (2024). The red lines are nonlinear least squares fits, with regression by composition fits shown in blue.

and, because ordinary least squares equations can be solved in closed form, the user is not required to choose starting values for a or c. This approach circumvents the fitting problems experienced by Walters et al. (2024) in a small number of cases, and elsewhere reproduces their results. Our regression by composition approach (which is necessarily iterative) converged successfully for all 97 individuals in the STORK database, initially setting all regression coefficients to zero.

Figure 1 reproduces Figure 1A–B from Walters et al. (2024) (black points and red curves), and overlays (in blue) the fitted regression by composition models. The individual plots represent ten selected children (the top row are boys, the bottom row are girls) from the STORK study. In eight of these ten children, the two fits are virtually indistinguishable.

In two children (2048 and 1009) there are modest discrepancies between the fits. These arise directly from the different variance structures being assumed: nonlinear least squares penalizes deviations equally at all ages, while this particular regression by composition recognises greater variability (and hence applies a lower penalty in fitting) with increasing age.

6.2. STARR data

Longitudinal records of infant growth are also stored in the STAnford medicine Research data Repository (STARR, Weber et al., 2024), including 14,695 children followed from birth for up to three years. We focus here on a subset of 11,655 children for whom measurements of height were available. As shown by Walters et al. (2024), modified Michaelis–Menten models fitted to STARR data by nonlinear least squares provide smoothed individual-specific growth curves for the purposes of interpolation, and can also offer plausible short-term predictions of a child's onward growth trajectory. However, the uniform variance structure implied by the least squares model is not well suited to modelling population quantiles of the kind needed to develop reference growth charts.

The World Health Organization publish growth charts for boys and girls created using a procedure known as LMS (Lambda-Mu-Sigma; Cole, 1988). For a variate of interest y, this model takes the form

$$y = \mu(\lambda\sigma z + 1)^{\frac{1}{\lambda}},$$

expressed in terms of a standard normal variate z. Each of the three parameters λ, μ, σ is allowed to vary with covariates, crucially including flexible dependence on age (here denoted x). The index λ is the transformation parameter introduced by Box and Cox (1964).

Since it is constructed via a series of data transformations, the LMS model can easily be expressed using regression by composition:

```
model y = N(0, 1) | Scale(ns(x)) | Translate(0 + offset(1)) |
            Power(ns(x)) | Scale(ns(x))
```

Here `ns()` is used to indicate a basis for a smooth curve, such as a natural cubic spline. The argument `offset(1)` expresses a known transformation; no coefficients are to be estimated in the `Translate()` flow, and $v_2 = 1$ for every observation. The smooth curves v_1, v_3, v_4 corresponding to the arguments of the flows `Scale()`, `Power()`, and another `Scale()` relate directly to LMS parameters:

$$\lambda = 1/v_3 \qquad \mu = v_4 \qquad \sigma = v_1 v_3.$$

Usually some form of likelihood penalty is imposed to ensure that the flexible terms are indeed appropriately smooth (Cole & Green, 1992).

We illustrate a different approach, allowing only linear dependence of v_1 and v_3 on age but replacing the final `Scale()` flow with three further flows expressing a modified Michaelis–Menten model:

```
model y = N(0, 1) | Scale(1 + x) |
            Translate(0 + offset(1)) | Power(1 + x) |
              Moebius(0 + 1/x) | Scale(1) | Translate(1)
```

In fact, we also allow each linear predictor to vary with sex, but suppress this for compactness.

Figure 2 shows nine reference centile curves overlaid with STARR height data. The ten individuals highlighted by Walters et al. (2024) in their Figure 1C–D are shown in coloured

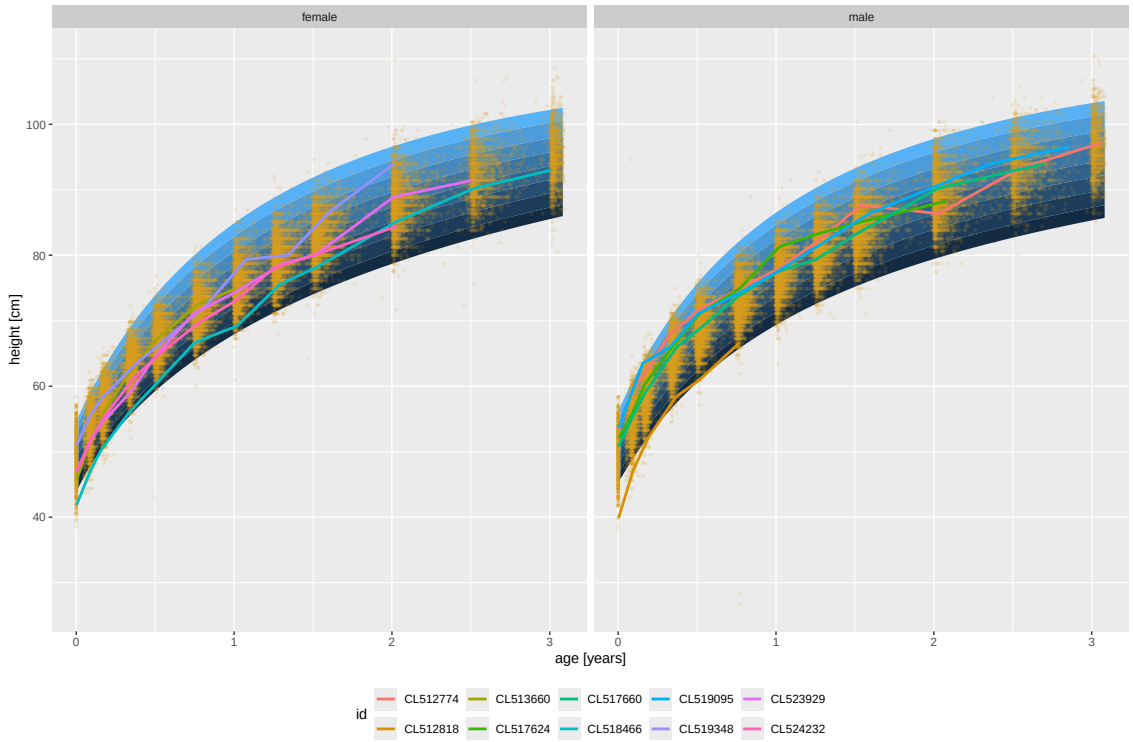

Fig. 2: Growth charts for height derived from the STARR data. The blue bands delimit the nine fitted centiles proposed by Cole (1994). Yellow dots show the raw data. The coloured lines show the growth of ten children selected for display by Walters et al. (2024).

lines, and their presence calls attention to one immediate (and valid) objection to our analysis, namely that we have treated longitudinal data as though they were cross-sectional. This will certainly lead to meaningful underestimation of standard errors; we gloss over this point simply by neglecting to report them. First-order biases may also result if, as seems at least plausible, there is a degree of informative sampling (as discussed in Farewell et al., 2017, for example): perhaps exceptionally tall or short individuals are seen more often as part of routine care.

These potential problems deserve a more considered approach than we give them here, but the anticipated correlation between a child's measurements of height could be introduced in at least two different ways. First, child-specific random effects could in principle be included inside the flow indices: that is, in the linear predictors η. Linear mixed models (Laird & Ware, 1982) can be understood as incorporating a random intercept and random slope (for example) into the linear predictor of the `Translate()` flow. In this way, a child's observations are correlated, and much more likely to 'track' a centile line. As usual, additional distributional specifications for the random effects are needed to complete the likelihood specification, and likelihood contributions must now be calculated for each child's observations collectively. It is certainly possible in principle to entertain the inclusion of random effects in more than one flow.

Alternatively, and arguably more in the spirit of regression by composition, we could begin by choosing p_0 as a suitable multivariate (normal, say) distribution, and combining it with flows acting on multivariate distributions. Analogues of existing flows operating observation-by-observation are relatively simple to implement; the crucial additional ingredients are one or more flows replacing `Scale()` with vector space actions inducing *covariance* between observations. In this way, linear random effects are replaced by suitable random walks, as for example in Diggle et al. (2007) or McCullagh (2022, Section 4.2). This more direct approach helpfully avoids some of the subtleties that can accompany latent variable models, including the interpretation of causal effects (Martinussen et al., 2020; Sarvet & Stensrud, 2022), and negative variances (Nelder, 1954; Bridge et al., 2024).

7. Modelling the effects of HIV treatments

Our second example concerns the estimation of treatment effects in a randomized controlled trial, particularly when enough is understood about the treatments being compared that a plausible shape for the relationship between the conditional risks of the two trial arms can be postulated.

7.1. The ACTG175 trial

ACTG175 was a randomized, controlled, double-blind clinical trial to evaluate the efficacy of combination antiretroviral therapies compared with monotherapy in individuals with HIV (Hammer et al., 1996). For this analysis, we focus on 477 participants randomized to

receive a combination of didanosine and zidovudine (multitherapy) and 474 participants assigned to zidovudine alone (monotherapy). The dataset, available in the `speff2trial` package in R, includes data from 1,054 participants across both treatment arms. Our analysis is based on 951 participants, excluding 103 individuals who were lost to follow-up within two years of randomization.

The primary outcome in the original trial was the time to the earliest occurrence of a reduction from baseline CD4 count of at least 50%, progression to AIDS, or death. For simplicity, we focus here on a binary outcome: whether this composite event occurred within two years of randomization. This two-year timeframe was chosen to capture sufficient events while minimizing participant loss to follow-up. Hammer et al. examined the potential for bias due to differential loss to follow-up and argued that the negligible association between baseline disease markers (such as CD4 count) and subsequent loss to follow-up supported the robustness of their conclusions.

The published estimated hazard ratio for the primary outcome, comparing multitherapy with monotherapy, was 0.50 (95% confidence interval 0.39 to 0.63), with a median follow-up of 143 weeks. In our subset, the proportions experiencing the primary outcome within two years of randomization were 14% in the multitherapy arm and 28% in the monotherapy arm.

Concerns have been raised about the generalizability of this result to other populations. The trial recruited patients with CD4 counts between 200 and 500 cells per cubic millimetre and demonstrated the superiority of multitherapy over monotherapy within this range. However, other trials conducted around the same time reported hazard ratios in the same direction but closer to the null for patients at higher risk, such as those with lower CD4 counts (Darbyshire, 1996; Saravolatz et al., 1996). Additionally, around the time these trials were published, studies suggested that adherence to combination therapy was lower among healthier, asymptomatic patients (those with higher CD4 counts). This was largely attributed to the severe gastrointestinal and other side effects of combination therapy, which asymptomatic individuals—perceiving their risk as relatively low—may have been less willing to tolerate (Mehta et al., 1997).

7.2. Outline of approach

Similar to Huitfeldt et al. (2022) and Daniel et al. (2024), we start by encoding the information above in a mechanistic causal model. However, while Huitfeldt et al. and Daniel et al. adopt models resembling Rothman's *causal pies* (Rothman, 1976), we instead express our model as a Markov chain with covariates. This avoids using joint counterfactuals, since their associated assumptions can be hard to verify (Robins & Richardson, 2011; Richardson, 2013).

Our mechanistic model, with its Markov memoryless assumption (conditional on baseline covariates) leads us to consider what shape of relationship may plausibly be present between the two outcome risks, given baseline covariates, under the conditions to be

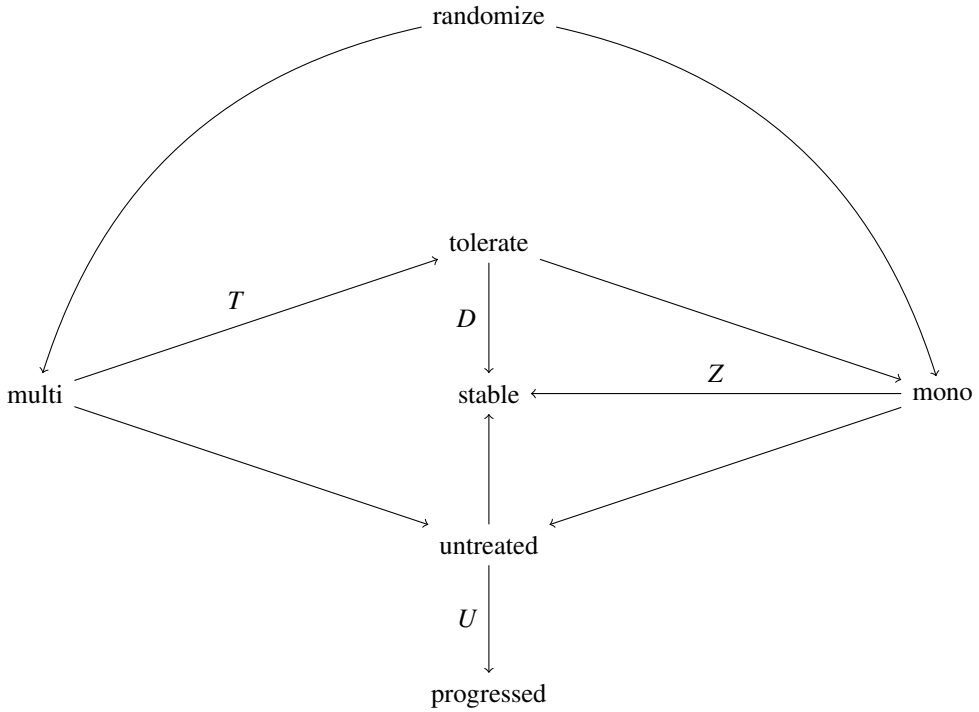

Fig. 3: The state space (and allowed transitions) of a Markov chain model representing the actions of the two treatments regimes (multitherapy and monotherapy) in the ACTG175 trial.

compared: randomization to multitherapy or monotherapy. We then consider a regression by composition model that permits such a relationship between the two conditional risks, and fit it to the data, comparing the results with those from more standard approaches.

7.3. A mechanistic model in the form of a Markov chain with covariates

7.3.1. *An explanation of the Markov chain model*

The assumed Markov chain model is depicted in Figure 3 and may be understood as follows. Each patient enters the study in a state to be randomized (S_1 = randomize), either to the multitherapy (S_2 = multi) or monotherapy (S_2 = mono) arm. A patient randomized to the monotherapy arm subsequently either experiences the intended viral response to zidovudine to an extent sufficient to protect them against disease progression for the next two years (S_3 = stable), an absorbing state, or otherwise zidovudine does not work sufficiently for them and they effectively enter a state of being untreated (S_3 = untreated). From that effectively untreated state, they either experience the primary outcome within two years (S_4 = progressed), also an absorbing state, or they do not (S_4 = stable). A patient randomized to the multitherapy arm (S_2 = multi), either tolerates the multitherapy

(S_3 = tolerate) or does not, and if they do not, they effectively enter a state of being untreated with either drug (S_3 = untreated). From this effectively untreated state, they either experience the primary outcome within two years (S_4 = progressed) or they do not (S_4 = stable). A patient randomized to multitherapy who does tolerate it (S_3 = tolerate) subsequently either experiences the intended metabolic response to didanosine to an extent sufficient to protect them against disease progression for the next two years (S_4 = stable), or otherwise didanosine does not work sufficiently for them and they effectively enter a state of being treated with zidovudine alone (S_4 = mono). From this state, similarly to those randomized to monotherapy, their disease is either stabilised by zidovudine (via the intended metabolic response, S_5 = stable) or zidovudine does not work sufficiently for them either and they effectively enter a state of being untreated (S_5 = untreated), from which state either they experience the primary outcome within two years (S_6 = progressed) or they do not (S_6 = stable).

7.3.2. *The time-homogeneous and memoryless assumptions conditional on covariates*
Some important assumptions are already implicit in the formulation above. First, the notion of treatment 'working' is considered binary: either the intended metabolic response is sufficient, in which case the two-year progression outcome is avoided, or the treatment response is insufficient and the patient enters a state of effectively not being treated with that drug. A more realistic model would allow for a continuum between these two extremes. Likewise, being able to tolerate multitherapy is simplistically considered binary. Also, the arrow from 'tolerate' to 'mono' represents an assumption of no biological interaction between the drugs in the sense that someone who takes multitherapy but for whom the didanosine component does not work sufficiently is comparable to someone (with the same covariate values) taking zidovudine alone. This too could be relaxed.

The remaining assumptions enter via the time-homogeneous and memoryless properties of the assumed transition probabilities, conditional on covariates $\mathbf{C}$, which are shown on the arrows in Figure 3. Let S_t be the state occupied by the Markov chain at stage $t = 1, 2, 3, \ldots$, and let $\bar{S}_t = (S_1, \ldots, S_t)$ denote the history of the Markov chain up to and including stage t. The probability Z that zidovudine 'works', conditional on baseline covariates $\mathbf{C}$, is

$$\text{pr}(S_t = \text{stable} \mid S_{t-1} = \text{mono}, \bar{S}_{t-2}, \mathbf{C}) = \text{pr}(S_t = \text{stable} \mid S_{t-1} = \text{mono}, \mathbf{C}). \qquad (1)$$

This probability is assumed not to depend on t (the time-homogeneous assumption) nor on $\bar{S}_{t-2}$ (the memoryless property) but it is a random variable, via its dependence on $\mathbf{C}$. Imagine two patients, A and B, who share the same baseline covariates $\mathbf{C}$. Patient A is randomized to multitherapy, is able to tolerate it, but didanosine does not work for them; Patient B is randomized to monotherapy. The time-homogeneous and memoryless assumptions (1) encode the important assumption that patients A and B, upon reaching the 'mono' state via two different paths, are identical with respect to their subsequent transition probabilities.

Similarly, the conditional probability U of experiencing the outcome from the 'untreated' state

$$\text{pr}(S_t = \text{progressed} \mid S_{t-1} = \text{untreated}, \bar{S}_{t-2}, \mathbf{C}) = \text{pr}(S_t = \text{progressed} \mid S_{t-1} = \text{untreated}, \mathbf{C})$$

can again depend on $\mathbf{C}$, but is assumed not to depend on whether the 'untreated' state was reached either (a) on the monotherapy arm, with zidovudine not working, or (b) on the multitherapy arm with multitherapy not being tolerated, or (c) on the multitherapy arm with neither zidovudine nor didanosine working.

We label the other two transition probabilities as $T = \text{pr}(S_3 = \text{tolerate} \mid S_2 = \text{multi}, \mathbf{C})$ and $D = \text{pr}(S_4 = \text{stable} \mid S_3 = \text{tolerate}, \mathbf{C})$. Again dependent on $\mathbf{C}$, T and D are the probabilities of tolerating dual therapy and of didanosine 'working', respectively.

7.3.3. *The consequences of the model*

Let Y denote the final state occupied by the chain (either 'progressed', which can occur at $t = 4$ or 6, or 'stable', which can occur at $t = 3$, 4, 5 or 6), and let X denote the randomized arm: that is, the second state occupied by the chain (S_2, either 'multi' or 'mono'). The two conditional risks of interest are $\text{pr}(Y = \text{progressed} \mid X = \text{multi}, \mathbf{C})$ and $\text{pr}(Y = \text{progressed} \mid X = \text{mono}, \mathbf{C})$. Under the assumptions of the Markov chain, we can evaluate both by multiplying conditional probabilities:

$$\text{pr}(Y = \text{progressed} \mid X = \text{multi}, \mathbf{C}) = (T(1 - D)(1 - Z) + (1 - T))U, \tag{2}$$

and

$$\text{pr}(Y = \text{progressed} \mid X = \text{mono}, \mathbf{C}) = (1 - Z)U. \tag{3}$$

If none of Z, D or T depended on $\mathbf{C}$, then the ratio of these two conditional risks would be a constant, and a generalized linear model with a log link would be a sensible choice for modelling the conditional distribution of Y given X and $\mathbf{C}$. When at least one of these three does depend on $\mathbf{C}$, the relationship is more complicated, as we now explore.

The ACTG175 dataset includes the following baseline covariates: age, gender, race (white or not), weight, CD4 count, and whether the patient has received antiretroviral therapy in the past; for disease progression, the most prognostic of these is CD4 count. For simplicity, suppose first that this is the only baseline covariate, and let $C \in [0, 1]$ represent the centiles of CD4 count. If an individual has a CD4 count of 350 cells/mm^3 then their C is the proportion of individuals with a CD4 count less than or equal to 350 cells/mm^3.

In what follows we will posit some plausible relationships between C and each of U, Z, D, and T. These are not modelling assumptions, but rather are used here to illustrate a plausible minimal level of complexity for the relationship between the two conditional risks of interest. Our aim will then be to accommodate at least this level of complexity in our chosen model.

To this end, we posit that U decreases with C: a higher CD4 count is associated with a lower probability of disease progression or death from an untreated state. We also posit that

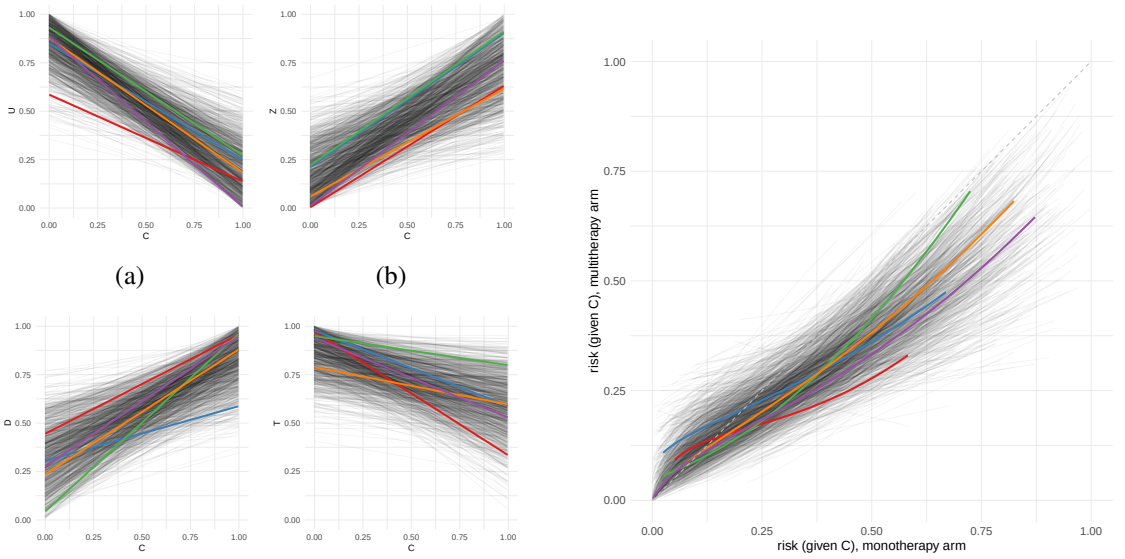

Fig. 4: Plots (on the left) showing a random collection of (presumed linear) relationships between the conditional transition probabilities in the Markov chain and C. These are shown in panels (a)–(d), respectively, showing the risk of progression from 'untreated', the propensity for Z to 'work', the propensity for D to 'work' and the propensity to tolerate multitherapy. Panel (e) (on the right) shows the resulting L'Abbé plots. Five sets of plots have been chosen at random and highlighted in colour, so that the nature of the individual plots can be more clearly visualized.

both Z and D increase with C: a higher CD4 count, and thus less frailty, is associated with a higher probability of a normal metabolic response to either drug treatment. Finally, we follow Mehta et al. (1997) in positing that T decreases with C: healthier individuals with higher CD4 counts (and hence a lower perceived risk) are less likely to willingly weather the unpleasant side effects of multitherapy.

In Figure 4, a random selection of linear relationships for each of U, Z, D and T with C is plotted, with the directions as discussed in the previous paragraph, along with the resulting *L'Abbé plot* for each selection (L'Abbé et al., 1987). For our purposes, a L'Abbé plot shows all pairs (p, p') of untransformed and transformed conditional risks across the range of possible values of p, where p represents the conditional outcome risk given CD4 count and $X = \text{mono}$, and p' is the same conditional outcome risk given $X = \text{multi}$.

Notable features of the curves on the right include almost all of them crossing the null line ($p' = p$) at a low risk, many of them meeting the $p = 0$ axis strictly above zero ($p' > 0$), most of the curves being convex across much of the range of p (although often concave at low p) and, for higher p, the curves being rather straight.

Of course, there is no reason to believe that the dependences on C chosen on the left of the plot should be linear. If the plots on the left were expanded to include non-linear possibilities, an even greater variety of shapes would be seen in the L'Abbé plots on the right. But Figure 4 suggests that, minimally, a regression model should be chosen that allows for relationships at least as complex as those seen in the right-hand plot.

7.4. A regression by composition model

As discussed in Section 1.2, risk transformations with no fixed points arise by composing a risk scaling transformation (`ScRisk1()`) with a transformation that scales the complement of the risk (the survival probability), `ScRisk0()`. To allow for the anticipated and largely convex nonlinearity in the L'Abbé plot (Figure 4), we also include treatment in the odds-scaling transformation (`ScOdds()`) used for the baseline covariates. For greater flexibility on the form of the nonlinearity (to allow for lower curvature at higher p, for example) we also include a flow (`PowOdds()`) that powers the odds ω: for $v \in (\mathbb{R}_+, \times)$, $\omega \cdot v = \omega^v$, leading to the following 11-coefficient regression by composition model:

```
model y = Ber(1/2) |
            ScOdds(1 + age + sex + race + wt + cd4 + past + trt) |
              ScRisk1(0 + trt) |
                ScRisk0(0 + trt) | PowOdds(0 + trt)
```

As we shall see, the fitted model makes appreciable use of the flexibility offered by each of the four different treatment flows.

7.5. Comparing the regression by composition model fit to logistic regression

Figure 5 compares the fit of the four-flow regression by composition model to what would be obtained from a simple logistic regression model

```
model y = Ber(1/2) |
            ScOdds(1 + age + sex + race + wt + cd4 + past + trt)
```

with a single parameter for treatment, and then a 14-coefficient logistic regression model (one grand mean, seven main effects, and six product terms)

```
model y = Ber(1/2) |
            ScOdds(1 + (age + sex + race + wt + cd4 + past) * trt)
```

that includes a product term between the treatment arm and each baseline covariate.

From the latter model, it is not immediately obvious how a L'Abbé plot should be constructed. We opted to fit a locally estimated scatterplot smoother to the fitted conditional risks, so that the transformed risk at a given value p of $\text{pr}(Y = \text{progressed} \mid X = \text{mono}, \mathbf{C})$ can be understood as a weighted average of each of the different $\text{pr}(Y = \text{progressed} \mid X = \text{multi}, \mathbf{C})$ corresponding to combinations of covariates $\mathbf{C}$ that lead to $\text{pr}(Y = \text{progressed} \mid$

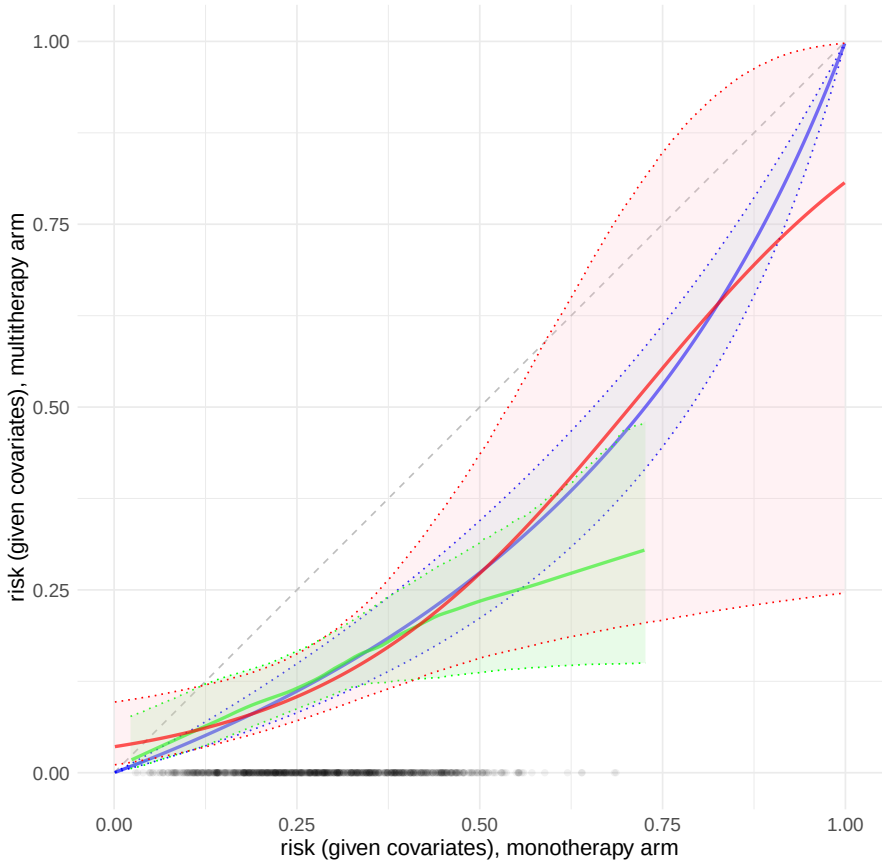

Fig. 5: L'Abbé plots for the fitted treatment comparison, with pointwise 95% confidence intervals, for three different models: our regression by composition model with four flows for treatment (red), simple logistic regression (blue), and logistic regression with a product term included for treatment with each baseline covariate (green). The fitted conditional risks for the monotherapy arm from the logistic regression model are plotted along the horizontal axis.

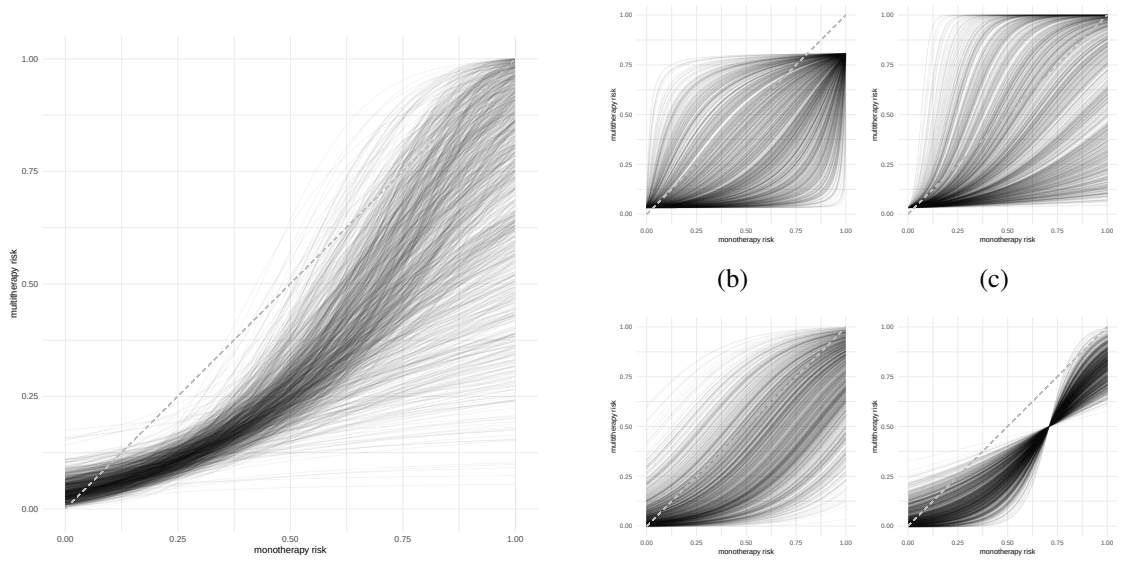

Fig. 6: A collection of L'Abbé plots illustrating the role of each treatment flow in our four-flow regression by composition model. The large plot on the left, panel (a), was obtained by sampling 1,000 times from a multivariate normal distribution with mean and variance–covariance matrix set as those of the four treatment parameter estimators. In the plots on the right, three of the four parameters are fixed at their estimates, and the remaining parameter is sampled independently from a normal distribution with the estimated parameter mean, and a variance four times that of the estimated sampling variance: this is the log odds ratio parameter in panel (b), the log risk ratio parameter in panel (c), the log survival probability ratio parameter in panel (d) and the log odds power parameter in panel (e).

$X = \text{mono}, \mathbf{C}) = p$, weighted by the distribution of $\mathbf{C}$ in the data. The 95% confidence interval was obtained by Monte Carlo sampling from the estimated joint sampling distribution of the estimators for the seven treatment parameters.

The fitted regression by composition (Figure 5, red curve) aligns closely with the illustrative curves from the mechanistic model on the right-hand side of Figure 4. Many key features are present: it crosses the identity line at a very low risk, albeit with some uncertainty, does not pass through $(0, 0)$ or $(1, 1)$, and is predominantly convex, with slight straightening at higher risks. Features of the four individual flows can be understood from Figure 6, which in turn fixes three of the four treatment parameters at their maximum likelihood estimates, while varying the fourth.

Within the range of risks present in the data (as shown by the concentration of points along the horizontal axis), all three models produce similar results. Without product terms and with its fixed points, simple logistic regression yields narrower confidence intervals—

particularly at very low and very high risks. In contrast, the estimated uncertainty in the four-flow regression by composition model increases as it moves away from the data. In particular, there is only a limited range of baseline risks (approximately 0.12 to 0.6) where the usual significance threshold of 5% would be reached to claim a protective treatment effect.

7.6. Concluding remarks on this worked example

We could have presented the results of our regression by composition in a traditional table of estimated coefficients and associated inference, but opted instead to display in a L'Abbé plot the implication of the estimated four treatment parameters for the relationship between conditional outcome risks. This is consistent with the causal inference literature (Rothman, 1975; Cole & Hernán, 2004; Kenah, 2024) where derived risk comparisons are preferred to regression coefficients. The benefit of such a visual aid is arguably greater for models that include more than one parameter for the treatment contrast.

Careful comparison of the four-flow regression by composition model with the logistic regression model with product terms is warranted. The fact that the L'Abbé plots differ (especially at risks greater than 0.5) is not surprising: the form of treatment effect heterogeneity permitted by the two models is different. For example, the logistic regression model with product terms constrains the logarithm of conditional odds ratio for treatment to change linearly with CD4 count, whereas more flexible shapes are implicitly allowed by the regression by composition model. On the other hand, it is perhaps somewhat reassuring that the upper 95% confidence limit crosses the identity line for low-risk patients at almost exactly the same point for both models.

The fact that the L'Abbé plot crosses the identity line at very low risks follows from the blinded nature of the trial; blinding leads to the posited arrow from 'multi' to 'untreated' in Figure 3. In an open-label trial, perhaps patients stop taking didanosine when finding the combination intolerable; an arrow would then instead go from 'multi' to 'mono', with a corresponding change to the set of time-homogenous and memoryless assumptions to be considered. These would imply a L'Abbé plot that does not cross the identity line.

Finally, equations (2) and (3) could also have been derived from a counterfactual approach. Such an approach would involve defining binary random variables intrinsic to each individual such as whether or not the outcome would occur for that individual were they to be unmedicated, whether zidovudine would work for that individual were they to take it, and so on. However, we prefer the derivation based on the Markov chain precisely because it avoids the necessarily cross-world nature of such intrinsic features. Knowing their values would tell us for each individual what their joint outcomes would be under monotherapy and multitherapy, and hence reasoning about assumptions based on them involves reasoning across worlds: assumptions that would be untestable even if the relevant realised features were known in the trial (for example, whether or not someone actually randomized to multitherapy could tolerate it). In contrast, our time-homogeneous

and memoryless assumptions would in principle be testable given information on the path taken by each patient through the Markov chain.

8. Discussion

Studying groups of transformations acting on probability distributions dates back at least to the monograph of Fraser (1968); for an accessible introduction, see Farewell and Prentice (1975). Fraser's structural model is founded on transformations of the space Υ in which the variate of interest takes its values. The same is true of the transformation models described by Barndorff-Nielsen (1983), for which he provides conditional distributions of maximum likelihood estimators given a suitable ancillary statistic.

McCullagh (1999) calls attention to the group actions implicit in generalized linear models (namely, exponential tilting), which cannot typically be expressed in closed form in terms of transformations of Υ. Our perspective is similar: it is not necessarily the *data* that are transformed, but instead the *distribution* of the data. The idea of transforming distributions encompasses and extends what is achievable through traditional pretransformation of the variate Y. Flows act on the distribution of Y expressed on some native scale, and in its native system of units. To our way of thinking, retention of scientifically interpretable scales is important in applied statistics. Explicit group actions on distributions also encourage informed model criticism (Box, 1980), enabling structured model comparison.

Our work concerns models that rely on parametric assumptions, aimed at settings where investigators have a clear rationale for making such assumptions. In some applied problems, especially when there is theoretical or mechanistic understanding, such assumptions are expected to hold, but not necessarily in ways that are easy to describe with conventional regression models. Thus, our goal is to offer a flexible way to encode such assumptions and carry out inference based on them. When these assumptions cannot be justified on scientific grounds, relying on them can result in misleading conclusions. In such cases more flexible nonparametric or semiparametric approaches (perhaps combined with machine learning) are warranted. Flexible models for nuisance components of the problem can still lead to valid inference for estimands of interest, often with robustness to certain kinds of misspecification. Our models are not intended to replace these methods, but serve a different purpose: to support situations where parametric structure is grounded in scientific reasoning or required by the context.

Regression by composition emphasizes modelled contrasts of distributions rather than low-dimensional summaries thereof, such as a difference or ratio of means. If a comparison of means is all that is needed, then fitting a parametric distribution is 'overmodelling', and vulnerable to misspecification. However, we contend that distributional comparisons are important in practice in many applications. It is usually not desirable to base treatment decisions, say, on knowing only that treated individuals on average fare slightly better than their untreated counterparts: like a lottery, a modest change in means might arise as a mixture of substantial gains by a small subset together with moderate losses for the large

majority. Low-dimensional comparisons do not need drastically differing distributional shapes in order to be of questionable relevance: the Behrens–Fisher problem (Fisher, 1935) illustrates the challenge of meaningful inference even when only the variance differs. Structured nonparametric distributional contrasts can sometimes be of scientific interest in functional data analysis, and these may also be framed as regression problems (Ghodrati & Panaretos, 2022).

As with transformation models, quantile regression, dispersion models, and so on, we hope to offer a menu beyond mean models like $E_\beta(Y \mid x) = \mu(x, \beta)$. Quite deliberately, we have avoided additive decompositions of the type $Y = \mu(x, \beta) + \epsilon$. There is a strong temptation to label or interpret this ϵ as (for example) 'noise' or individual-specific 'measurement error' and, perhaps even more dangerously, to characterize the mean $\mu(x, \beta)$ as the 'true Y'. In working with distributions directly, regression by composition highlights not the magnitude of an arbitrarily-chosen *additive* residual ϵ between (say) arithmetic mean prediction and observation, but instead the *surprise* (in the information-theoretic sense) of the observation in the predicted outcome distribution, implicitly adapting the choice of discrepancy metric to the application in question.

We have found that regression by composition improves our understanding of the parametric models we study and apply, and enables extensions thereof. Decomposition breaks up existing approaches into tractable components that can be reassembled as bespoke models, and flows that are old friends continue to work genially alongside more recent acquaintances.

Acknowledgements

We are grateful to Yacine Trad, Vern Farewell, Anthony Davison, and Victor Panaretos for insightful questions on drafts of this paper. We thank Mark Chatfield for helpful discussions about multiplicative comparisons. Tim Morris and Ruth Keogh brought to our attention the collapsibility conundrum in Section 4.3. William Walters and Catherine Ley kindly shared with us the STORK and STARR data, and Tim Cole gave perceptive comments about growth modelling. Simon Schoenbuchner offered invaluable assistance in the creation of the accompanying R package `rbc`. Detailed scrutiny by several referees led to substantial improvements throughout.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, *8*(8), 907–925.

Abel, N. H. (1826). Untersuchung der Functionen zweier unabhängig veränderlichen Größen x und y, wie f(x, y), welche die Eigenschaft haben, daß f(z, f (x,y)) eine symmetrische Function von z, x und y ist. *Journal für die reine und angewandte Mathematik*, *1*, 11–15.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1996). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

Baker, R., & Jackson, D. (2018, June 9). *A new measure of treatment effect for random-effects meta-analysis of comparative binary outcome data*. arXiv: 1806.03471 [stat].

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, *70*(2), 343–365.

Battey, H. S., Cox, D. R., & Jackson, M. V. (2019). On the linear in probability model for binary data. *Royal Society Open Science*, *6*(5), 190067.

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227), 357–365.

Blunck, M., & Mommsen, T. P. (1978). Systematic Errors in Fitting Linear Transformations of the Michaelis-Menten Equation. *Biometrika*, *65*(2), 363–368.

Bochner, S. (1933). Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind. *Fundamenta Mathematicae*, *20*(1), 262–176.

Bogachev, V. I. (2007, January 15). *Measure Theory*. Springer Science & Business Media.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Bridge, H., Morgan, K. E., & Frost, C. (2024). Negative variance components and intercept-slope correlations greater than one in magnitude: How do such "non-regular" random intercept and slope models arise, and what should be done when they do? *Statistics in Medicine*, *43*(14), 2747–2764.

Briggs, G. E., & Haldane, J. B. S. (1925). A Note on the Kinetics of Enzyme Action. *Biochemical Journal*, *19*(2), 338–339.

Brown, R. (2018, June 21). *A Modern Introduction to Dynamical Systems*. Oxford University Press.

Cole, S. R., & Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, *75*(1), 45–49.

Cole, T. J. (1988). Fitting Smoothed Centile Curves to Reference Data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *151*(3), 385–406.

Cole, T. J. (1994). Do growth chart centiles need a face lift? *BMJ*, *308*(6929), 641–642.

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, *11*(10), 1305–1319.

Colquhoun, D. (1969). A Comparison of Estimators for a Two-Parameter Hyperbola. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *18*(2), 130–140.

Cornish-Bowden, A. (2014). Analysis and interpretation of enzyme kinetic data. *Perspectives in Science*, *1*(1), 121–125.

Cornish-Bowden, A. (2015). One hundred years of Michaelis–Menten kinetics. *Perspectives in Science*, *4*, 3–9.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Daniel, R., Zhang, J., & Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557.

Daniel, R. M., Farewell, D. M., & Huitfeldt, A. (2024). 'Does God toss logistic coins?' and other questions that motivate regression by composition. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *187*(3), 636–655.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, *25*(4), 631–650.

Darbyshire, J. H. (1996). Delta: A randomised double-blind controlled trial comparing combinations of zidovudine plus didanosine or zalcitabine with zidovudine alone in hiv-infected individuals. *Lancet*, *348*(9023), 283–91.

Dawid, A. P. (2001). Some variations on variation independence. *International Workshop on Artificial Intelligence and Statistics*, 83–86.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600.

Didelez, V., & Stensrud, M. J. (2022). On the logic of collapsibility for causal effect measures. *Biometrical Journal*, *64*(2), 235–242.

Dieudonné, J. (1960). *Foundations of Modern Analysis*. Academic Press.

Diggle, P., Farewell, D., & Henderson, R. (2007). Analysis of longitudinal data with drop-out: Objectives, assumptions and a proposal. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *56*(5), 499–550.

Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models* (4th ed.). CRC Press.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*, *1932*(3), 175–195.

Farewell, D. M., Huang, C., & Didelez, V. (2017). Ignorability for general longitudinal data. *Biometrika*, *104*(2), 317–326.

Farewell, V. T., & Prentice, R. L. (1975). Interpreting the Structural Model. *Statistische Hefte*, *16*(2), 115–122.

Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *Annals of Eugenics*, *6*(4), 391–398.

Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, *71*(3), 431–444.

Ghodrati, L., & Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, *109*(4), 957–974.

Greenland, S., & Pearl, J. (2011). Adjustments and their Consequences—Collapsibility Analysis using Graphical Models. *International Statistical Review*, *79*(3), 401–426.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, *14*(1), 29–46.

Halmos, P. R. (1974). *Finite-Dimensional Vector Spaces* (S. Axler, F. W. Gehring, & K. A. Ribet, Eds.). Springer.

Halmos, P. R. (1956). *Lectures on Ergodic Theory*. American Mathematical Soc.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081–1090.

Hanley, J. A. (2025). Studies in the history of probability and statistics, LI: The first conditional logistic regression. *Biometrika*, *112*(1), asae038.

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, *43*(1), 59–74.

Hernán, M. A. (2010). The Hazards of Hazard Ratios. *Epidemiology*, *21*(1), 13–15.

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, *40*(3), 780–785.

Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, *183*(8), 758–764.

Hernán, M. A., & Robins, J. M. (2023, December 31). *Causal Inference: What If*. CRC Press.

Hothorn, T., Kneib, T., & Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *76*(1), 3–27.

Huitfeldt, A., Fox, M. P., Murray, E. J., Hróbjartsson, A., & Daniel, R. M. (2022). *Shall we count the living or the dead?* arXiv: 2106.06316 [stat].

Hutton, G. (1999). A tutorial on the universality and expressiveness of fold. *Journal of Functional Programming*, *9*(4), 355–372.

Johnson, K. A., & Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry*, *50*(39), 8264–8269.

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, *96*(4), 761–780.

Kalbfleisch, J. D., & Prentice, R. L. (1981). Estimation of the Average Hazard Ratio. *Biometrika*, *68*(1), 105–112.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Keating, K. A., & Quinn, J. F. (1998). Estimating Species Richness: The Michaelis-Menten Model Revisited. *Oikos*, *81*(2), 411–416.

Kenah, E. (2024). Rothman diagrams: The geometry of confounding and standardization. *International Journal of Epidemiology*, *53*(6), dyae139.

Kernighan, B. W., & Pike, R. (1984). *The UNIX Programming Environment*. Prentice-Hall.

Kreyszig, E. (1978). *Introductory Functional Analysis with Applications*. John Wiley & Sons.

L'Abbé, K. A., Detsky, A. S., & O'rourke, K. (1987). Meta-Analysis in Clinical Research. *Annals of Internal Medicine*, *107*(2), 224–233.

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, *38*(4), 963–974.

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, *15*(3), 209–225.

Ley, C., Sanchez, M. d. l. L., Mathur, A., Yang, S., Sundaram, V., & Parsonnet, J. (2016). Stanford's Outcomes Research in Kids (STORK): A prospective study of healthy pregnant women and their babies in Northern California. *BMJ Open*, *6*(4), e010810.

Lineweaver, H., & Burk, D. (1934). The Determination of Enzyme Dissociation Constants. *Journal of the American Chemical Society*, *56*(3), 658–666.

Luo, X., & Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika*, *99*(1), 211–222.

Mac Lane, S. (1971). *Categories for the Working Mathematician*. Springer Science & Business Media.

Martinussen, T., Vansteelandt, S., & Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, *26*(4), 833–855.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall/CRC.

McCullagh, P. (1996). Möbius transformation and Cauchy parameter estimation. *The Annals of Statistics*, *24*(2), 787–808.

McCullagh, P. (1999). *The algebraic structure of generalized linear models* (Technical Report No. 489). University of Chicago.

McCullagh, P. (2002). What Is a Statistical Model? *The Annals of Statistics*, *30*(5), 1225–1267.

McCullagh, P. (2022). *Ten Projects in Applied Statistics*. Springer International Publishing.

Mehta, S., Moore, R. D., & Graham, N. M. H. (1997). Potential factors affecting adherence with HIV therapy. *AIDS*, *11*(14), 1665–1670.

Michaelis, L., & Menten, M. L. (1913). Die kinetik der invertinwirkung. *Biochem Z*, *49*, 333–369.

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika*, *41*(3–4), 544–548.

Nelder, J. A. (1965). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proceedings of the*

Royal Society of London. Series A. Mathematical and Physical Sciences, *283*(1393), 147–162.

Nelder, J. A. (1974). Log Linear Models for Contingency Tables: A Generalization of Classical Least Squares. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *23*(3), 323–329.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Neuhaus, J. M., & Jewell, N. P. (1993). A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models. *Biometrika*, *80*(4), 807–815.

Panaretos, V. M., & Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, *22*(57), 1–64.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Piccininni, M., & Stensrud, M. J. (2025). Immune-selection stability is a neglected property of the causal risk ratio. *American Journal of Epidemiology*, kwaf086.

Pinheiro, J. C., & Bates, D. M. (2009). *Mixed-effects models in S and S-PLUS*. Springer Verlag.

Pollard, D. (2002). *A user's guide to measure theoretic probability* (Vol. 8). Cambridge University Press.

Polya, G. (1954). *Mathematics and Plausible Reasoning, Volume 1: Induction and Analogy in Mathematics*. Princeton University Press.

Raaijmakers, J. G. W. (1987). Statistical Analysis of the Michaelis-Menten Equation. *Biometrics*, *43*(4), 793–803.

Reeve, R., & Turner, J. R. (2013). Pharmacodynamic Models: Parameterizing the Hill Equation, Michaelis-Menten, the Logistic Curve, and Relationships Among These Models. *Journal of Biopharmaceutical Statistics*, *23*(3), 648–661.

Richardson, T. (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality — Center for Statistics and the Social Sciences. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*(30).

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*(3), 507–554.

Robins, J. M., & Richardson, T. S. (2011, February 3). Alternative Graphical Causal Models and the Identification of Direct Effects. In P. Shrout, K. Keyes, & K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures* (p. 0). Oxford University Press.

Robinson, D. J. S. (1996). *A Course in the Theory of Groups* (Vol. 80). Springer.

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, *104*(6), 587–92.

Rothman, K. J. (1975). A pictorial representation of confounding in epidemiologic studies. *Journal of Chronic Diseases*, *28*(2), 101–108.

Saravolatz, L. D., Winslow, D. L., Collins, G., Hodges, J. S., Pettinelli, C., Stein, D. S., Markowitz, N., Reves, R., Loveless, M. O., Crane, L., Thompson, M., Abrams, D., & Investigators for the Terry Beirn Community Programs for Clinical Research on AIDS. (1996). Zidovudine alone or in combination with didanosine or zalcitabine in hiv-infected patients with the acquired immunodeficiency syndrome or fewer than 200 cd4 cells per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1099–1106.

Sarvet, A. L., & Stensrud, M. J. (2022). Without Commitment to an Ontology, There Could Be No Causal Inference. *Epidemiology*, *33*(3), 372.

Sheps, M. C. (1958). Shall We Count the Living or the Dead? *New England Journal of Medicine*, *259*(25), 1210–1214.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238–241.

van der Laan, M. J., Hubbard, A., & Jewell, N. P. (2007). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 463–482.

Vansteelandt, S., & Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(3), 657–685.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (J. Chambers, W. Eddy, W. Härdle, S. Sheather, & L. Tierney, Eds.). Springer.

Walters, W. A., Ley, C., Hastie, T., Ley, R. E., & Parsonnet, J. (2024). A modified Michaelis-Menten equation estimates growth from birth to 3 years in healthy babies in the USA. *BMC Medical Research Methodology*, *24*(1), 27.

Watkins, W. J., Farewell, D., Banerjee, S., Nasef, H., James, A., & Chakraborty, M. (2020). Modelling predictive gender- and gestation-specific weight reference centiles for preterm infants using a population-based cohort study. *Scientific Reports*, *10*(1), 4032.

Weber, S. C., Pallas, J., Olson, G., Love, D., Malunjkar, S., Boosi, S., Loh, E., Datta, S., & Ferris, T. A. (2024). *Compliant Self Service Access to Secondary Use Clinical Data at Stanford Medicine*. arXiv: 2412.04248 [cs].

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, *11*(14–15), 1871–1879.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *22*(3), 392–399.

A. Notational choices

Since there is such limited supply, we reduce, reuse and recycle our letters and labels. An example is our decision to represent measurable spaces (formally consisting of a set and a sigma algebra) simply by the symbol for the set (Ω, say). This is consistent with standard practice in abstract algebra, where a set G can also denote the same set enriched with a group operation. When distinguishing specific sigma algebras on measurable spaces, we name them. Two other attempts at notational minimalism: we write sequences and tuples as $(a_1, a_2, a_3, \ldots)$ without assigning a symbol to their cardinalities, which could be zero, any natural number, or indeed countably infinite. And symbols for measures ($p, P, \mathbb{P}$, etc.) mean both the measure and integration (expectation) with respect to that measure (de Finetti notation; see Pollard, 2002, pp. 7–11).

B. Inferential and inductive setting

The ambitious inductive goal of statistical reasoning is to turn findings about *specific* observational units into more *general* scientific statements. Using the language of category theory, McCullagh (2002) describes a very rich class of internally-consistent observational schemas; these enable us to consider the transport of findings from one setting to decisions or predictions in another. Although our notional aim is to carry out conventional inference about the conditional law $\mathbb{P}(Y \mid \mathcal{F})$ of some variate of interest Y corresponding to an idealized individual sampled in a *similar* way to the units actually observed, we might equally hope to characterize the experiences of an imagined participant in a hypothetical randomized *target trial* (Hernán & Robins, 2016), retaining some conditional distributions from the observed distributions while replacing others (Pearl, 2009, pp. 72 sqq.). This alternative type of induction underlies much of statistical causal inference. It follows that conditional distributions (and hence regression models) are central to ordinary prediction *and* causal inference, and our paper has both use cases in mind.

C. Affine spaces

We require a space of laws $\mathcal{P}$ with sufficient algebraic structure that it permits study of collapsible contrasts of law-valued random variables (Greenland & Pearl, 2011, and see also Section 4). Minimally, an affine structure allows us to talk about convexity (where a law is meaningfully 'between' other laws), characterizes affine transformations of laws (which preserve straight lines, and the relative distances between laws upon them), and admits recognizable versions of integration (for marginalization) and differentiation (for optimization).

Throughout this paper we work with the set $\mathcal{P}$ of signed laws p on the measurable space Υ with total measure $p(\Upsilon) = 1$. To endow $\mathcal{P}$ with an affine structure, we associate to this set the vector space $\mathcal{Q}$ of signed laws q having total measure $q(\Upsilon) = 0$, which describe *translations* of laws. Addition and scalar multiplication of translations in $\mathcal{Q}$ mirror

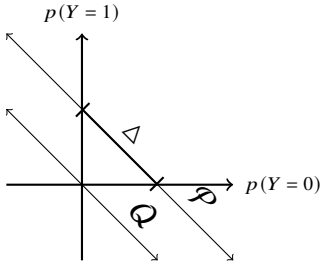

Fig. 7: For a binary variate $Y \in \Upsilon = \{0, 1\}$, diagrammatic representation of the affine space $\mathcal{P}$ of signed laws p with total measure $p(\Upsilon) = 1$. The affine space $\mathcal{P}$ contains the simplex $\triangle$ of probability distributions and is modelled on the Banach space $\mathcal{Q}$ of laws with total measure $p(\Upsilon) = 0$.

the corresponding addition and scalar multiplication of the measures they assign to sets: that is, addition of laws $q, q' \in \mathcal{Q}$ and scalar multiplication by reals $a, b \in \mathbb{R}$ is defined via $(aq + bq')(A) = aq(A) + bq'(A)$, for all measurable $A \subseteq \Upsilon$. Equipping $\mathcal{Q}$ with the variation norm (Bogachev, 2007, p. 177) makes it a Banach space (Kreyszig, 1978, p. 58), and the affine space $\mathcal{P}$ gains a metric (but not a norm), an induced topology, and Borel sets, so becomes suitable for defining law-valued random variables $P : \Omega \to \mathcal{P}$.

Affine spaces are unbounded, so while the signed laws in $\mathcal{P}$ are like probability distributions in that they have unit sum over admissible values of Y, they differ in that events are permitted to have measure outside $[0, 1]$. The probability laws $\triangle$ are a strict subset of $\mathcal{P}$, as Figure 7 illustrates.

In this section we show briefly how infinitesimal calculus is possible in affine spaces that are associated with Banach spaces, allowing us to compute conditional means $\overline{P}$ of law-valued random variables P, and derivatives f' of transformations of laws $f : \mathcal{P} \to \mathcal{P}$.

Means as centres of mass in affine spaces

Integration of functions taking values in general vector (Banach) spaces is described by Bochner (1933). This theory permits us to define the mean $\mathbb{P}(X)$ under the measure $\mathbb{P}$ of a random variable $X : \Omega \to B$ taking values in any Banach space B.

We make use of a straightforward generalization to affine spaces that retains the defining characteristic of the mean as the *centre of mass*. Let A be an affine space with associated vector (Banach) space B and binary addition $\oplus : A \times B \to A$ and subtraction $\ominus : A \times A \to B$ operators. By way of definition, we say that a random variable $X : \Omega \to A$ taking values in the affine space A has mean $\mu \in A$ under the measure $\mathbb{P}$ if

$$\mathbb{P}(X \ominus \mu) = 0,$$

where 0 denotes the identity element in B. The integral $\mathbb{P}(X \ominus \mu)$ is well-defined because the random variable $X \ominus \mu$ takes values in the Banach space B. Thus defined, the mean

μ is unique: if also $\mathbb{P}(X \ominus \mu') = 0$ for some $\mu' \in A$, then $\mu \ominus \mu' = \mathbb{P}(\mu \ominus \mu') = \mathbb{P}((\mu \ominus X) + (X \ominus \mu')) = 0$. The first equality holds because $\mu \ominus \mu'$ is a constant, the second uses one of Weyl's axioms, and the third follows from linearity of integration and from the defining centre of mass property shared by μ and μ'.

Differentiation as best local linear approximation in affine spaces

The Fréchet derivative is an abstraction of differentiation to normed vector spaces (see, for example, Dieudonné, 1960, pp. 143 sqq.). Here we restrict attention to Banach spaces: *complete* normed vector spaces. When it exists, the Fréchet derivative of a function f between Banach spaces B and B' is the linear map-valued function $f' : B \to \mathcal{L}(B, B')$ satisfying

$$\lim_{\|h\| \to 0} \frac{\|f(x+h) - f(x) - f'(x)(h)\|}{\|h\|} = 0.$$

Generalization to affine spaces is again natural: for a function f between affine spaces A and A' with associated vector (Banach) spaces B and B', respectively, the Fréchet derivative is a function $f' : A \to \mathcal{L}(B, B')$ satisfying

$$\lim_{\|h\| \to 0} \frac{\|f(x \oplus h) \ominus f(x) - f'(x)(h)\|}{\|h\|} = 0$$

if such a function f' exists. Here $f(x \oplus h), f(x) \in A'$, but both $f(x \oplus h) \ominus f(x)$ and $f'(x)(h)$ are elements of B', so $f(x \oplus h) \ominus f(x) - f'(x)(h)$ is an ordinary subtraction operation in B'.

D. Marginal laws

As we saw in Appendix C, assigning an affine geometry to a set $\mathcal{P}$ suffices to give meaning to the expectation $\mathbb{P}(P)$ of an $\mathcal{F}$-measurable $\mathcal{P}$-valued random variable P under the measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ or, given $\mathcal{G} \subseteq \mathcal{F}$, a conditional expectation $\mathbb{P}(P \mid \mathcal{G})$, itself a $\mathcal{G}$-measurable law-valued random variable that we will denote $\overline{P}$.

Motivating our particular choice of affine structure is the fact that any $\mathcal{P}$-valued random variable P having an interpretation as a conditional law retains a conditional law interpretation under marginalization. To see this, let $\mathbb{P}$ be any measure on Ω satisfying $\mathbb{P}(Y \in A \mid \mathcal{F}) = P(A)$ for all measurable $A \subseteq \Upsilon$, $\mathbb{P}$-almost surely, so that the law-valued random variable P may justifiably be identified with the conditional law of Y under $\mathbb{P}$, given $\mathcal{F}$. Then for any measurable $A \subseteq \Upsilon$, we have by the tower law that

$$\mathbb{P}(Y \in A \mid \mathcal{G}) = \mathbb{P}(\mathbb{P}(Y \in A \mid \mathcal{F}) \mid \mathcal{G}) = \mathbb{P}(P(A) \mid \mathcal{G}) = \mathbb{P}(P \mid \mathcal{G})(A),$$

$\mathbb{P}$-almost surely, showing that $\mathbb{P}(P \mid \mathcal{G})$ may indeed be identified with the conditional law of Y under $\mathbb{P}$, given $\mathcal{G}$. The final equality in this proof results from the fact that, in the Banach space $\mathcal{Q}$ on which $\mathcal{P}$ is modelled, measures add and scalar multiply according to their probabilities.

E. Covariate contrasts

For so-called *treatment contrasts* of categorical covariates, the reference covariate value implied by the identity contrast $X = 0 \in \mathbb{V}$ will be one actually taken by the covariate (often a control or untreated level), while for *sum-to-zero contrasts* the reference point will be a fictional one located within the convex hull of the covariate contrasts, but not actually assumed by any subject. As an example, for a three-category covariate, the sum-to-zero contrast $X \in \{(1,0),(0,1),(-1,-1)\}$ has notional reference value $(0,0)$, the identity element in $\mathbb{U} = (\mathbb{R}^2, +)$.

F. Isomorphisms of the affine space $\mathcal{P}$

We often work with equivalent but (in some contextually-convenient sense) 'simpler' expressions of laws $p \in \mathcal{P}$, such as cumulative distribution functions F, probability-generating functions G, characteristic functions ϕ, and so on (Polya, 1954, p. 101). By *equivalent*, we mean that knowing one implies knowledge of the other. Given a characteristic function, it is possible in principle to compute the law, and vice versa: there is a bijection between characteristic functions and laws. Given such a bijection, we can define a group action on laws by defining a group action on characteristic functions, and conversely.

For the purposes of definition, and for studying invariance or other algebraic properties of group actions on laws, the affine structure of $\mathcal{P}$ is not relevant. However, collapsibility relates to properties of marginalized laws $\mathbb{P}(P \mid \mathcal{G})$, where geometric considerations *do* matter. Therefore, for considerations of collapsibility and affine transformations of laws p via some other expression of them (like characteristic functions ϕ), we insist that the relevant bijection (such as $p \mapsto \phi$) is an *isomorphism of affine spaces*: a bijection $f : \mathcal{P} \to \mathcal{P}'$ for which

(a) there exists an isomorphism $g : \mathcal{Q} \to \mathcal{Q}'$ between the vector spaces associated with $\mathcal{P}$ and $\mathcal{P}'$, and
(b) the function f is *equivariant* in the sense that $f(p \oplus q) = f(p) \oplus g(q)$ for all $p \in \mathcal{P}$ and $q \in \mathcal{Q}$.

Equipped with such an isomorphism, expectations taken in the two affine spaces will also be equivalent.

In many cases, the required isomorphisms of affine spaces are conspicuous: for example, translations of laws $p \in \mathcal{P}$ by laws $q \in \mathcal{Q}$ correspond to exactly similar translations of characteristic functions. Defining the bijection $f : p \mapsto (t \mapsto p(e^{itY}))$ and the vector space isomorphism $g : q \mapsto (t \mapsto q(e^{itY}))$, we have $f(p \oplus q) = t \mapsto (p \oplus q)(e^{itY}) = t \mapsto (p(e^{itY}) + q(e^{itY})) = (t \mapsto p(e^{itY})) \oplus (t \mapsto q(e^{itY})) = f(p) \oplus g(q)$ and hence an isomorphism between the affine space $\mathcal{P}$ of laws and an affine space of characteristic functions, as required. Similarly direct isomorphisms exist for probabilities of a binary variate ($p \mapsto p(Y = 1)$, say), probability density functions, cumulative distribution

functions, survivor functions, probability-generating functions, and moment-generating functions.

These isomorphisms are 'natural' in the sense that the implied affine structure in $\mathcal{P}'$ is precisely the one we might assign to it even without reference to $\mathcal{P}$; speaking loosely, the bijection f is 'naturally' affine in the input law p. Alas, not all isomorphisms of $\mathcal{P}$ are natural in this way. However, given *any* bijection $f : \mathcal{P} \to \mathcal{P}'$ and an isomorphism $g : \mathcal{Q} \to \mathcal{Q}'$ (the identity map, say, although other choices are possible), an isomorphism of affine spaces can always be constructed by insisting that the translation of $p' \in \mathcal{P}'$ by $q' \in \mathcal{Q}'$ is $p' \oplus q' = f(f^{-1}(p') \oplus g^{-1}(q'))$, addition on the right-hand-side being a translation of $f^{-1}(p') \in \mathcal{P}$ by $g^{-1}(q') \in \mathcal{Q}$, subsequently mapped by f back into $\mathcal{P}'$. Writing $p = f^{-1}(p')$ and $q = g^{-1}(q')$, it is immediate that $f(p) \oplus g(q) = f(p \oplus q)$ so, as required, f is indeed equivariant. Defined in this way, the affine structures in the two sets are necessarily identical.

At least to us, probability-respecting affine spaces of hazard functions, cumulative hazard functions, and cumulant-generating functions are harder to visualize, and caution is warranted when working geometrically with them. For a binary variate Y, the relationship between the space $\mathcal{P}$ of laws and its equivalent in terms of odds is another example of this less immediate kind of isomorphism. The punctured, wraparound domain $(-1, \infty) \cup \{\infty\} \cup (-\infty, -1)$ of the odds $p \mapsto p(Y = 1)/p(Y = 0)$ offers a clue that an unusual affine structure could be needed. Indeed, the *decimal* or *European* odds $p \mapsto p(Y = 1)/p(Y = 0) + 1$ have a similar punctured domain $(0, \infty) \cup \{\infty\} \cup (-\infty, 0)$ and a natural *reciprocal* geometric structure, because $p(Y = 1)/p(Y = 0) + 1 = 1/p(Y = 0)$. The expectation associated with the reciprocal affine space of the decimal odds is the harmonic mean. The curious geometry of the usual odds therefore has an increment-reciprocal flavour in which, in particular, the appropriate associated expectation is a quasi-arithmentic or generalized f-mean with $f(\omega) = 1/(\omega + 1)$.

G. Algebraic structures for collapsibility

Weak collapsibility requires bounding the mean of a two-argument function f in terms of its values while holding the first argument at its mean. The function f could describe a group action—and it is in this form that we use the following result—but it need not do so. Jensen's inequality for convex g is a fairly direct consequence of this result, taking $f(x, v) = g(x) + v$ for $x, v \in \mathbb{R}$.

Lemma G.1 (Jensen bounds). *Let $f : A \times \mathbb{V} \to A$ be family of transformations of an affine space A indexed by the vector space $\mathbb{V}$. Let $R \subseteq A$ and $H \subseteq \mathbb{V}$ be convex regions, and suppose the random variables X, V satisfy $X \in R$ and $V \in H$ almost surely. If the set*

$$\{(x, f(x, v)) : x \in R, v \in H\}$$

is a convex subset of A^2, then $\overline{f(X, V)} \in f(\overline{X}, H)$, almost surely.

Proof. Almost surely, $(X, f(X,V)) \in \{(x, f(x,v)) : x \in R, v \in H\}$. Since by assumption this set is convex, we have $(X, f(X,V) \in \{(x, f(x,v)) : p \in R, v \in H\}$. But $(X, f(X,V) = (\overline{X}, \overline{f(X,V)})$, so $\overline{f(X,V)} = f(\overline{X}, v)$ for some $v \in H$, almost surely. □