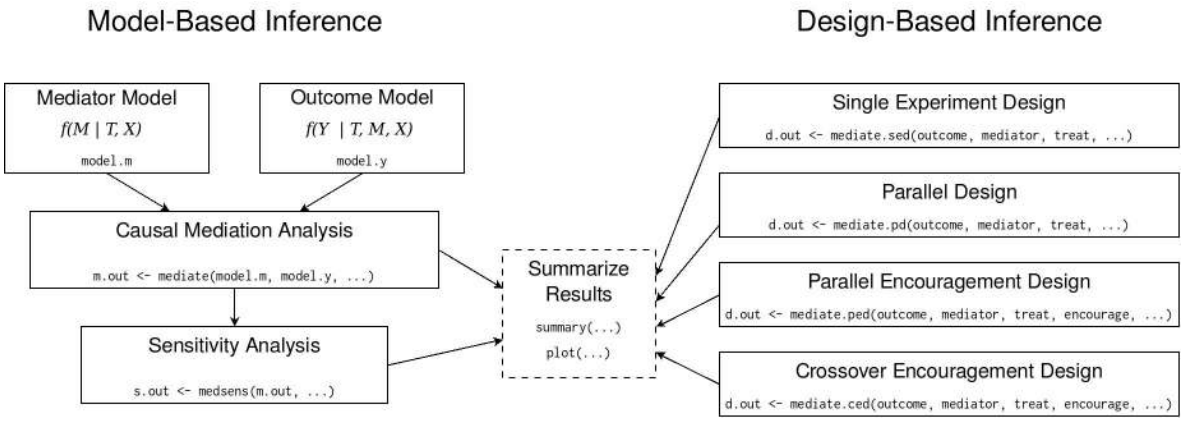

Figure 1: Core structure of the **mediation** package as of version 4.0.

to define these quantities. Let $M_i(t)$ denote the potential value of a mediator of interest for unit i under the treatment status $T_i = t$. Let $Y_i(t, m)$ denote the potential outcome that would result if the treatment and mediating variables equal t and m, respectively. Consider a standard experimental design where only the treatment variable is randomized. We observe only one of the potential outcomes, and the observed outcome, Y_i, equals $Y_i(T_i, M_i(T_i))$ where $M_i(T_i)$ represents the observed value of the mediator M_i. With this notation, the total unit treatment effect can be written as,

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)). \tag{1}$$

We can decompose this total effect into the two components. First, the *causal mediation effects* are represented by (Robins and Greenland 1992; Pearl 2001),

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \tag{2}$$

for each treatment status $t = 0, 1$. All other causal mechanisms can be represented by the *direct effects* of the treatment as,

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \tag{3}$$

for each unit i and each treatment status $t = 0, 1$. Together, we see that they sum up to the total effect,

$$\tau_i = \delta_i(t) + \zeta_i(1 - t) \tag{4}$$

for $t = 0, 1$. The case of multiple candidate mediating variables requires additional notation and is discussed in Section 6. The *average causal mediation effects* (ACME) $\bar{\delta}(t)$ and the average direct effects (ADE) $\bar{\zeta}(t)$, represent the population averages of these causal mediation and direct effects.

Identification of the ACME requires an additional assumption beyond the strong ignorability of the treatment, which is sufficient for identifying the average total effect of the treatment. Let X_i be a vector of the observed pre-treatment confounders for unit i. The key identifying assumption is called sequential ignorability and can be written as,