When Delay Becomes a Property of the Treatment

A framework for dynamic therapeutic optimization

For interventions whose production kinetics rival the kinetics of the disease, the treatment with the best average efficacy is not necessarily the one that maximizes expected utility for a given patient.

Utility depends on the patient's anticipated state at the moment when that treatment can actually be administered.

ZUMA-7, TRANSFORM and BELINDA offer a textbook case. Their divergences in efficacy do not stem from a single source (CAR biology, CD28 vs 4-1BB domain, population). They also reflect how each strategy delivers the treatment to the patient, and in what clinical state that patient is at that moment.

1. The three trials and the shared puzzle

Shared context

The three trials address an identical strategic question in second-line treatment: in patients with LBCL (aggressive diffuse large B-cell lymphoma) that is primary refractory or in early relapse (≤12 months), substituting an anti-CD19 CAR-T for the classical pathway (salvage chemotherapy followed by autologous transplant in responders).

The stake: this population has demonstrated its chemo-resistance. The classical pathway requires a response first. CAR-T bypasses that barrier.

Observed clinical results

Trial	CAR-T	N total	N CAR-T arm	CR (%)	Primary measure	Result
ZUMA-7	Axi-cel	359	180	NR	2-year EFS 41% vs 16% (HR 0.40)	Positive
TRANSFORM	Liso-cel	184	92	74% vs 43%	Median EFS 29.5 mo vs 2.4 mo (HR 0.375)	Positive
BELINDA	Tisa-cel	322	173	46% vs 43%	Median EFS 3.0 mo vs 3.0 mo (HR 1.07, p=NS)	Negative

The puzzle

Liso-cel and tisa-cel share the same costimulation domain (4-1BB). One succeeds (TRANSFORM). The other fails (BELINDA).

If CAR biology were the single determinant, TRANSFORM would be hard to explain. Yet TRANSFORM succeeds.

Where to look? Not a single source. All three contribute: product properties, population characteristics, and the architecture of the system that delivers the treatment to the patient.

2. Three central objects: theoretical architecture

Object 1: The patient's dynamic clinical state

$S_i(t)$ = multidimensional state of patient i at time t

This state includes: tumor burden, growth rate, ECOG, LDH, systemic inflammation, organ function, immune fitness.

The state is not binary. It is a continuous trajectory that degrades (or improves) according to the disease, the bridging, the comorbidities.

Object 2: Delivery delay as a distribution

$\tau_{ij} \sim F_{ij}(\tau)$ = distribution of the delay between decision and effective administration

Not a single delay. A distribution:

$E[\tau_{ij}]$, $Var(\tau_{ij})$, $P90(\tau_{ij})$

This distribution captures:

- **Manufacturing**: CAR designs, vein-to-vein time
- **Patient**: permitted bridging, comorbidities inducing delays
- **System**: center capacity, apheresis slots, logistics, qualifications
- **Random**: infections, administrative complications, unforeseeable events

Variance and tail latency matter as much as the median.

Object 3: The treatment's response function

$R_j(S)$ = expected response/utility of treatment j in clinical state S

Each treatment has a continuous performance function that depends on the patient's state.

Examples:

- **Autologous CAR-T**: maximal response if ECOG ≤1, moderate burden, LDH <500. Response decreases progressively if ECOG rises or the state degrades.
- **Allogeneic**: response conditioned by tolerance of rejection, immunosuppression capacity.
- **Bispecifics**: response less dependent on the initial state but subject to other constraints (pharmacokinetic persistence, microenvironment).

No sharp boundary (in/out). A gradual degradation.

One can define a utility threshold:

$\Omega_j(u^*) = \{S : R_j(S) \geq u^*\}$

But this region depends on the chosen threshold and is not an absolute biological property of the treatment.

3. Expected value and decision

Combining the three objects:

$V_{ij} = E_{\tau,S}[R_j(S_i(t0 + \tau))]$

Translation: the expected utility of treatment j for patient i when one accounts for:

- the delay distribution $\tau \sim F_{ij}$
- the uncertain trajectory $S_i(t)$
- the response function $R_j(S)$ of the treatment in that state

The decision becomes:

$j^* = \arg\max_j V_{ij}$

The best treatment for this patient at this instant is the one that maximizes expected utility, integrating timing and anticipated state.

4. Conceptual implications

The therapeutic window depends on the (patient, intervention) pair

There is no universal "patient T_W." Rather: each treatment imposes a different relation between delay and state.

A patient may be:

- Outside the useful domain of autologous CAR-T (delay too long, state too degraded)
- But still within the domain of the bispecific (immediate availability)
- Or of allogeneic CAR-T (moderate delay, different tolerance domain)

The optimal choice exploits these offsets.

The stages of a time-constrained intervention

Stage 1: the patient trajectory $S_i(t)$ begins to degrade from relapse / "refractoriness."

Stage 2: each treatment j introduces a delay τ_{ij} before it can be administered.

Stage 3: at the moment $t0 + \tau_{ij}$, the patient is in a state $S_i(t0 + \tau_{ij})$.

Stage 4: this treatment j has a utility $R_j(S_i(t0 + \tau_{ij}))$.

Stage 5: integrating the distribution of τ gives the expected utility V_{ij}.

5. Decomposition of the determinants of delay

The observed delay is not a simple sum. It is a complex function:

$\tau_{observed,ij} = g(\tau_{strategy}, \tau_{patient}, \tau_{system}, \tau_{stochastic})$

where:

- **$\tau_{strategy}$**: delay intrinsic to the product/protocol (design, manufacturing, release)
- **$\tau_{patient}$**: delay induced by disease aggressiveness, fitness, endogenous bridging
- **τ_{system}**: delay imposed by center capacity, slots, logistics
- **$\tau_{stochastic}$**: random events (infections, administrative blockages)

The components are simultaneous and interacting. An infection can occur during manufacturing without adding linearly. Bridging can be caused by a manufacturing delay.

Implication for the estimand

- For a **treatment-policy estimand**: the observed delay is part of the effect of the assigned strategy, whatever its determinant. One does not "adjust" it away by neutralizing it statistically.
- For a **causal analysis comparing architectures**: this decomposition becomes indispensable, because a fraction of the delay reflects patient characteristics, not therapeutic design.
- The right doctrine: do not naively adjust a post-randomization variable **without first precisely defining the target estimand and the causal DAG.**

6. Reinterpretation of the three trials

ZUMA-7: positive result

Axi-cel (CD28):

- Manufacturing time: tolerable
- Response function R_axi(S): broad domain of useful state (tolerates ECOG 2, moderate burden)
- Delay distribution: median acceptable for the majority of patients
- Conclusion: the majority of enrolled patients had a state S compatible with R_axi at the moment of infusion. Observed efficacy reflects temporal success + biological response.

TRANSFORM: positive result

Liso-cel (4-1BB, controlled CD4/CD8 architecture):

- Manufacturing time: short
- Bridging: 63% of patients
- Response function R_liso(S): domain comparable to axi-cel
- Conclusion: the positive result validates that 4-1BB is not intrinsically inferior. Short manufacturing architecture + favorable bridging selection = the majority of patients still in the useful domain.

Important: it does not refute that CD28 vs 4-1BB is biologically equivalent. Simply: TRANSFORM demonstrates that if the biology is different, the system's temporal architecture can compensate.

BELINDA: negative result

Tisa-cel (4-1BB, without CD4 control):

- Manufacturing time: **very long (52 days)**
- Permissive bridging: 83-97% of patients
- Apheresis population: very aggressively selected (last-resort salvage)
- Response function R_tisa(S): domain biologically comparable to liso-cel
- Critical problem: very long delay + ultra-sick population = when the patient reaches infusion (day 52+), the state S_i(t0 + 52d) is often **outside the domain where R_tisa retains utility.**

The result EFS = 3.0 mo vs 3.0 mo SOC does **not** invalidate 4-1BB biology.

It demonstrates that a long delay combined with a very aggressively fast population can make **the patient leave the useful performance domain** before administration, independently of the CD28 vs 4-1BB design.

7. Core vocabulary (four terms only)

Potency: a property of the product. Intrinsic capacity to recognize antigen, activate, proliferate.

Efficacy: effect observed under trial conditions. What is measured in the clinical protocol.

Effectiveness: effect expected in the real patient-system context. Integrates delay, the patient's real state, real system constraints.

Utility: effectiveness weighted by toxicity, quality of life, preferences. This is what V_ij represents.

8. Architectural alternatives: comparisons

Autologous

Advantages: perfect immunological compatibility, possible persistence.

Costs:

- τ_strategy long (manufacturing)
- The patient's S(t) degrades during this delay
- If the delay is very long: the patient may leave the domain where R_autologous is useful

Allogeneic

Advantages: standardized input, manufacturable, τ_strategy potentially shorter.

Costs:

- Major host-versus-graft rejection
- Immunosuppression required, massive infrastructure
- Different performance domain Ω (not a superset of autologous)

Bispecifics

Advantages: τ_strategy very short (generally available immediately).

Costs:

- Drug persistence (classical pharmacokinetics) ≠ response durability (prolonged clinical effect)
- Microenvironment not bypassed (Treg, hypoxia)
- Antigen escape identical to CAR-T

Competitive timing: when the CAR-T delay becomes penalizing (very fast disease, long τ_strategy), their relative value rises in the V_ij calculation. They can "compete."

Anti-CD19/CD3 or anti-CD20/CD3 bispecifics

Major advantage: τ_strategy very short (0-7 days), often immediately available.

But: the profile of R_bisp(S) is dramatically different from CAR-T.

The drug persistence vs response durability distinction

Central confusion: "Bispecifics persist for a shorter time than CAR-T, therefore less useful."

Technically: true for pharmacokinetics (T1/2 days-weeks vs cellular expansion).

Clinically: false for durable response. An antibody can disappear in 4 weeks and induce a clinical response that lasts months to years.

Mechanism: the bispecific does not persist, but it induces the reconstitution of T cell immunity, elimination of persistent antigen, tissue remodeling, and then these effects persist.

Therefore:

R_CAR(S) = f(cellular auto-expansion + long persistence)

R_bisp(S) = f(exposure-induced response + post-exposure durability)

These functions are not ordered, they are different.

Empirical data (2024-2025):

- **Glofitamab (anti-CD20/CD3)**: CR 93%, 2-year relapse-free ~75%

- **Epcoritamab (anti-CD20/CD3)**: CR 76%, 12-month RFS ~80%
- **Vs CAR-T**: CR 74-87%, 2-year RFS ~75-80%

Bispecifics numerically comparable on initial response and 2-year durability.

Critical limitation: patient T cell fitness

Autologous CAR-T: infuses its own genetically modified cells. One can force it even if the patient's T lymphocytes are exhausted.

Bispecific: recruits the patient's endogenous T cells. If the patient has:

- CD3+ count <200
- CD8 <50
- Elevated exhaustion markers (PD-1+/TIM-3+ >50%)

then R_bisp(S) collapses drastically vs R_auto(S), which tolerates it.

This parameter is not routinely measured. So at the moment of decision, one is blind.

Microenvironment not bypassed

CAR-T can (partially) bypass obstacles (Treg, TGF-β, hypoxia).

The bispecific depends on a permissive microenvironment for T cell recruitment.

A severe state S with Treg dominance, very high TGF-β, hostile microenvironment → R_bisp(S) drops.

The role of Θ for bispecifics

Θ_bisp = τ_R,bisp / τ_clinical evolution = (1-7 days) / doubling time

Example 1: Ultra-fast disease (doubling time 7 days, HGBCL Ki67 95%)

Θ_bisp = 4 days / 7 days = 0.57 → competitive Θ_auto = 28 days / 7 days = 4.0 → the patient leaves the useful state

Selection: j* = bisp (if T cell fitness is acceptable)

Example 2: Slow disease (doubling time 60 days)

Θ_bisp = 4 days / 60 days = 0.07 → timing not decisive Θ_auto = 28 days / 60 days = 0.47 → timing acceptable

Selection: j* = auto (better R_j on durability + T cell tolerance)

Realistic clinical context for selecting j*

Scenario A: HGBCL, Ki67 95%, doubling ~7d, patient aged 55, ECOG 1, Lym 600, fatigued T cells

V_auto(S): low (state degrades by D28, R_auto reduced) V_bisp(S): high if T cells are vital, moderate otherwise V_allo(S): low (too slow for this kinetics)

Decision: j* = bisp if T cell fitness is acceptable, otherwise autologous CAR-T (risky but the only option).

Scenario B: DLBCL indolent-transformation, doubling ~60d, patient aged 65, comorbidities, ECOG 1, Lym 400, very low CD8

V_auto(S): high (stable state, can be forced despite bad T cells) V_bisp(S): low (depends on T cells, CD8 very low) V_allo(S): moderate (acceptable delay, but additional risks)

Decision: j* = auto, clearly. Bispecific competitive only if the patient refuses genetic modification.

The four truths about bispecifics

1. **No absolute advantage.** Better in specific contexts (high Θ, good T cell fitness), worse or equal elsewhere.
2. **Very low Θ → zero timing advantage in slow disease.** The only stake becomes R_j, where autologous CAR-T wins (persistence).
3. **T cell fitness is THE determinant.** CD3+, CD8, exhaustion markers (not routinely measured at the moment of decision).
4. **Long-term durability TBD.** Bispecifics have 2-3 years of follow-up. CAR-T have 5+. The gap is narrowing but not yet closed.

Synthesis

Bispecifics do not "dominate." They compete according to the dynamic V_ij.

The framework makes it possible to formalize when:

V_bisp > V_auto, V_allo ⇒ j* = bisp

vs when CAR-T wins.

This is intellectually far more honest than "bispecifics = the future."

The truth: bispecifics = an additional option in a multi-arm decision, not a universal superiority, and that is already very useful.

9. Important blind spots

Therapeutic reliability: beyond the median

A treatment at a median of 25 days / P90 = 60 days can be **less useful** than one at a median of 30 days / P90 = 35 days.

Therapeutic reliability = f(E[τ], Var(τ), P90(τ))

In a capacity-constrained system operated near saturation, the **right tail of the τ distribution explodes.**

Reserve capacity as clinical capacity

Naive max-utilization optimization ignores the clinical cost of tail latency.

Total Value(C) = Economic Efficiency(C) − Clinical Loss from Tail(C) dClinicalLoss/dC increases dramatically near saturation

So C* < 100% can be optimal **both economically and clinically.**

Dimensionless ratio of temporal competition (heuristic)

Θ_ij = E[τ_R,ij] / τ_clinical evolution

An intervention belongs to "time-constrained therapeutics" if Θ is not much less than 1 (that is, Θ approaches or exceeds 1).

Important: Θ is a **dimensional heuristic, not a validated biomarker.** The denominator remains to be defined precisely (doubling time? time to ECOG 3? to a 50% drop in utility?). Without calibration, use it qualitatively ("major temporal competition") rather than as numerical thresholds.

Class concerned: CAR-T, TIL, neoantigen vaccines, adaptive radiotherapy, transplantation, any situation where Θ approaches or exceeds 1.

10. Operational implications

Consequence 1: Measure decision-to-treatment

Current KPIs report vein-to-vein or apheresis-to-infusion. They ignore decision-to-apheresis.

The real stake: patients can lose the window **even before apheresis.**

Candidate KPI: decision point → effective administration.

Consequence 2: Measure the distribution, not the median alone

Report E[τ], variance, P90 tail latency.

P90 is clinically more relevant than the median for predicting the % of patients with temporal failure.

Consequence 3: Integrate anticipated τ_ij into selection

The optimal decision must explicitly consider the anticipated τ_R,ij and its impact on S_i(t0 + τ_ij).

Illustration: decision-support pathways based on [anticipated Θ, available temporal alternatives], with outcomes measured: decision-to-treatment, observed toxicities, survival-since-diagnosis.

11. Generalization: time-constrained interventions

Any intervention where Θ_ij is not much less than 1 obeys the same principles:

1. The patient's state at the moment of administration modifies the relative value of treatments
2. System reserve capacity = clinical capacity
3. The therapeutic window depends on the (patient, intervention) pair
4. The optimal decision = dynamic optimization j* = arg max_j V_ij

A unified efficacy metric

Rather than comparing "raw efficacy" across trials, use the V_ij framework integrating timing.

This reconciles:

- ZUMA-7 (efficacy visibly superior to SOC)
- TRANSFORM (efficacy visibly superior to SOC)
- BELINDA (efficacy = SOC)

Not by saying "4-1BB is inferior." But: "temporal architecture + ultra-fast population = temporal failure independent of the costimulation domain."

12. Theoretical and practical conclusions

Rigorous theoretical conclusion

For interventions whose production kinetics rival those of the disease, the treatment with the best average efficacy is not necessarily the one that maximizes expected utility for a given patient.

The anticipated utility V_ij depends on the state S_i at the moment when treatment j can be administered, not only on the treatment itself.

Therefore:

The relevant treatment is not necessarily the treatment with highest average efficacy. It is the treatment with the highest expected value given the patient's anticipated state when that treatment can actually be delivered.

Conclusion for clinical practice

The relevant treatment is not the best treatment. It is the best reachable treatment, the one that can reach the patient while the patient remains in its performance domain.