Paying It Backward and Forward: Expanding Access to Convalescent Plasma Through Market Design[*]

Scott Duke Kominers   Parag A. Pathak   Tayfun Sönmez   M. Utku Ünver[†]

October 2020

Abstract

Convalescent plasma is a blood product produced by recovered patients with several valuable uses, especially during public health emergencies. We develop a model of plasma donation and distribution and consider two incentive schemes to increase plasma supply based on "paying it backward" and "paying it forward" principles. Under the former, donors obtain credits that can be transferred to patients of their choosing. Under the latter, patients obtain priority for plasma-derived products in exchange for a future donation pledge. We show that both incentives generally increase overall treatment rates for all patients—not just those with credits or who have pledged. Finally, we examine the implications of pooling blood types on the efficiency and equity of plasma distribution. Our formal results are of independent interest for egalitarian divisible goods rationing programs with compatibility constraints.

Keywords: COVID-19, convalescent plasma, blood markets, vouchers
JEL codes: D47, C78

[*] We are grateful to the convalescent plasma expanded access group at the Mayo clinic for feedback, including Michael Joyner and Gianrico Farrugia, Mordecai Serle and the Covid Plasma Initiative, and Bill Block and others from the Blood Bank of America. Mohammad Akbarpour, Walter Gilbert, Andrew Kahn, Louis Kaplow, Ellen Kominers, Alan Lam, Peter Lee, Jimmy Lin, Elizabeth Sibert, and especially Govind Persad also provided useful comments. We also thank Deniz Aktürk Erdem for bringing social media appeals for COVID-19 convalescent plasma to our attention.

[†] Kominers: Harvard Business School and Department of Economics, Harvard University; kominers@fas.harvard.edu. Pathak: MIT Department of Economics and NBER; ppathak@mit.edu. Sönmez: Boston College, Department of Economics; sonmezt@bc.edu. Ünver: Boston College, Department of Economics; unver@bc.edu. Kominers gratefully acknowledges the support of National Science Foundation grant SES-1459912, as well as the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications (CMSA). Sönmez gratefully acknowledges the research support of Goldman Sachs Gives via Dalinc Ariburnu - Goldman Sachs Faculty Research Fund.

1 Introduction

Blood plasma is the liquid part of blood that holds blood cells and dissolved proteins. Convalescent plasma is from a patient who has recovered from a disease. Because it contains proteins produced while battling illness, convalescent plasma has several valuable medical uses. One use is convalescent plasma therapy, in which plasma is injected into a sick patient who is blood-type compatible in order to boost that patient's immune response. A second use is in formulating medical treatments like hyperimmune globulin, monoclonal antibodies, and other related prophylatics. Both uses are common during the outbreak of novel diseases, when no other treatments are available.[1]

The Covid-19 pandemic has attracted new attention to the procurement and distribution of convalescent plasma. As the pandemic has evolved, there has been on-going demand for convalescent plasma for the development of new therapies, as well as periods of acute shortage for plasma therapy as the disease spread throughout the world.[2] By and large, the procurement of convalescent plasma from patients is decentralized: local public health authorities, hospital staff, and physicians encourage recovered patients to donate. There have only been a handful of coordinated efforts for donation, typically from hard-hit communities.[3] Several blood donation centers, including the American Red Cross and the Blood Centers of America, established procedures to collect Covid-19 convalescent plasma. These centers sell donated plasma to hospitals and pharmaceutical companies. Some donation centers have even experimented with forms of directed donation. For example, OneBlood, a Florida blood center, allows for referred donation, in which the center attempts to match donated plasma to an intended recipient, who may be a friend or family member (OneBlood, 2020). The New York Blood Center also initially allowed donations from patients from specific hospitals to be returned to be used by other patients at the same hospital (White Plains White Plains Hospital, 2020).[4]

This paper introduces and analyzes a market design approach to collecting and distributing convalescent plasma. We develop a model that jointly incorporates donation and allocation of plasma and explore two incentive schemes to increase the supply of plasma based on *pay-it-backward* and *pay-it-forward* principles. Through the pay-it-backward principle, the system "pays back" a plasma donor for her potentially life-saving donation by giving her a number of credits that can be used to obtain priority for plasma therapies of her loved ones should the need arise. Through the pay-it-forward principle, a patient receives priority access for plasma therapy in exchange for a pledge to return the favor by donating her own plasma in the near future, assuming she recovers and becomes eligible for plasma donation.[5] These features embed and formalize practices that are already informally embraced

[1] Convalescent plasma therapy was used during the 2003 SARS-CoV-1 epidemic, 2009-2010 H1N1 influenza virus pandemic, and 2012-13 MERS-CoV epidemic (EBA, 2020; Rubin, 2020). During the 1918 Spanish flu, fatality rates were cut in half for patients treated with blood plasma (see Luke et al., 2006 and Roos, 2020). Convalescent plasma has also been used to treat measles, influenza, and other infectious diseases. In fact, the first Nobel Prize in Physiology or Medicine was the 1901 prize for serum therapy (serum is the liquid left after coagulant elements are removed from plasma).

[2] Joyner et al., (2020) reports on studies of the effectiveness of convalescent plasma for Covid-19.

[3] Stack, (2020) notes that the Orthodox Jewish community in New York City, initially hard-hit by Covid-19, have likely provided more than half of the plasma in Mayo Clinic's expanded access program as of May 2020.

[4] As of April 15, 2020, this no longer occurs.

[5] A similar feature exists in non-directed donor (NDD) chains in kidney exchange, where a patient receives a living-donor kidney before her incompatible donor donates a kidney to a patient in another incompatible patient-donor pair.

by some doctors in their attempt to increase the recruitment of plasma donors (see, e.g., Rubin, 2020).

In our steady-state model of plasma donation, plasma donors may be given credits that can be used to give treatment priority to family members and other close associates; priority is also given to participants in clinical trials. The steady-state availability of plasma therapy is a function of the number of patients who have recovered (both through plasma therapy and by other means). We find that so long as the plasma replenishment rate is large enough to support the clinical trial, it is possible to treat all prioritized patients in equilibrium. The rate of treatment for non-prioritized patients becomes higher as a result of the priority scheme, as well. We characterize when it is possible to treat all patients—even those who are not ex ante prioritized—and show that so long as recovered patients are more willing to donate if they receive credits, introducing a credit system strictly benefits non-prioritized patients. Overall treatment availability expands further if we prioritize patients who pledge to pay it forward by donating plasma once they have recovered: if patients who pledge to donate have an aggregate plasma replenishment rate that is more than one-for-one, prioritizing those patients increases the treatment rate for non-prioritized patients, irrespective of how many patients pledge to donate ex ante.

Most of our analysis works with a single blood type for ease of illustration. But we show how to combine that analysis with ideas from graph theory to identify the optimal cross-blood-type plasma-pooling strategy to maximize an egalitarian treatment objective. We show that our pooling procedure (1) maximizes the smallest service rate and (2) minimizes the difference between the largest and smallest service rate across blood types. Since our results do not depend on the specific structure of plasma donation compatibility, the results we develop in this section are of independent interest for egalitarian divisible good rationing problems with compatibility constraints.

The remainder of this paper is structured as follows. Section 2 reviews some design considerations that might be relevant for practical implementation of our idea. Section 3 describes our model of plasma donation and distribution under a blood-type identical allocation policy or in allocation of plasma products, such as hyperimmune globulin, where blood-type compatibility is not needed; Section 4 extends the framework to blood-type compatible allocation; we then review related literature in Section 5. Section 6 concludes.

2 Market Design Considerations for Plasma Donation and Distribution

We envision a mechanism in which only a portion of the convalescent plasma supply can be allocated through the two types of incentive schemes we introduce. We refer to that portion as the *incentivized plasma reserve*. The remaining portion is reserved for participants in clinical trials, as well as for any other patient group the central planner selects for special treatment; for simplicity, we refer to that portion as the *clinical trial plasma reserve*. The clinical trial plasma reserve is effectively exogenous—at any point in time, the clinical trial plasma reserve will be allocated to its beneficiaries.

The incentivized plasma reserve, meanwhile, is endogenous: it depends on two different types of

Such an NDD chain becomes possible with the undirected initial donation of a Good Samaritan donor; the longest single-center paired kidney exchange of this form involved 101 donors and recipients (Pope, 2018).

incentives. The first incentive we consider is the provision of a fixed number of credits to plasma donors, which can be later redeemed by patients of the donors' choosing; we refer to this as a *pay-it-backward* incentive. These credits are of potential value to donors because patients who arrive the system with a credit have *first-tier priority* access for units in the incentivized plasma reserve.

The second type of incentive—which we call a *pay-it-forward* incentive—exploits the unusual feature of convalescent plasma that any patient who recovers becomes a potential plasma donor. This provides an opportunity to expand access to plasma: if we can use plasma to increase the patient recovery rate, and those recovered patients go on to donate plasma, then we can grow the plasma supply more than one-for-one. Thus, we propose to provide *second-tier priority* access to units in the incentivized plasma reserve for patients who do not have a credit but who pledge to donate plasma in the near future, in the event that they recover. Any patient who is able to fulfill her pledge through a plasma donation may also receive a number of credits, although fewer than those provided to donors under pay-it-backward incentives.

The priority tiers for access to the plasma product through the incentivized plasma reserve are then:

1. *First-tier priority*: Patients who arrive with a credit.

2. *Second-tier priority*: Patients who arrive without a credit but who pledge to donate plasma upon recovery, subject to eligibility requirements.

3. *Third-tier priority*: Any other patient who is in need of a plasma product.

Within each tier, ties are broken in a systematic way determined by the central planner. The system can be utilized to allocate plasma therapies or other plasma-derived products like hyperimmune globulin.

Meanwhile, the allocation process in the clinical trial plasma reserve is fully regulated by the central planner.

2.1 Pay-it-Backward Incentives

Some donors are purely altruistic and need no incentive to donate. But potential donors may at least in part wish to be able to donate to their loved ones.[6] For these donors, the pay-it-backward incentive can be expected to be valuable because the credit provides a medium of exchange that eases three frictions associated with donation. For example, consider a potential donor who wants to donate to a family member. She may not be able to donate to her intended recipient if any of the following three difficulties arise:

1. The donor and intended recipient are *time-incompatible*: when the beneficiary needs plasma therapy, the donor is medically unable to donate.

2. The donor and intended recipient are *plasma-incompatible*: the beneficiary has antibodies for antigens in the donor's blood that makes the donation medically impossible.

[6] This consideration appears in donor FAQs, such as OneBlood, (2020).

3. The donor and intended recipient are *location-incompatible*: the donation is either difficult or impossible due to travel limitations.

By functioning as an in-kind medium of exchange, a credit surmounts each friction; this should naturally result in greater overall donation. And because plasma donors can donate multiple units of plasma, the resulting increase in plasma supply benefits the overall patient pool—not just credit recipients.

There are two important precedents for the credit system we envision. The first is blood assurance programs used for whole blood donation. In a blood assurance program, a donor obtains credits for a donation. These credits can be used to obtain discounts, refunds or waived feeds if a credit holder is ever to receive blood. For example, in the Cape Fear Valley system, each blood donation equals one blood credit that may be kept by the donor or transferred to a family member or friend in need. Blood credits are used to replace blood charges for patients in the health system (Cape Fear Valley, 2020). Starr, (2002, p. 190) describes the important role that blood assurance programs played in the development of US blood markets in the 1960s, though they are currently less common.

The second precedent for a credit system is from kidney exchange: A *voucher for a chronologically incompatible* pair (Veale et al., 2017) involves giving a (typically young) patient priority for a future kidney transplant in exchange for a kidney donation from an older donor today; this mechanism is used when the donor is expected to be too old to donate when the patient will need a transplant. A relatively modest number of these intertemporal exchanges have been organized by the National Kidney Registry, which arranges kidney chains initiated by good-samaritan donors.[7] We anticipate a potentially more substantial role for credits in plasma donation, because the risk and potential negative consequences to the donor are much lower for plasma donation than for kidney donation.

2.2 Pay-it-Forward Incentives

The pay-it-backward principle just discussed rewards plasma donation ex post. The pay-it-forward principle, by contrast, gives an ex ante reward for a pledge to donate in the future conditional on recovery and eligibility; as we show in the next section, this too can be expected to increase the overall plasma supply, so long as a large enough fraction of the pledged donations are actually carried through.

It is thus essential to think about how many pledged donations will actually materialize. Some patients who benefit from pay-it-forward incentives may be unable to donate for medical eligibility reasons.

It is also possible that a patient may simply decide not to honor her pledge. This is an important practical issue, but non-directed donor chains in kidney exchange show that it is surmountable. In a non-directed donor kidney exchange chain, a patient receives a kidney based on the pledge that their donor will donate a kidney to another patient in the future. It is possible that after a patient receives a kidney, their donor may renege; however, in practice this rarely occurs. Cowan et al., (2017) report that only six donors reneged over the course of 1,700 transplants. And the incentive to renege on upfront pledges may be stronger for kidneys than for plasma, since kidneys do not regenerate and require a much more invasive procedure for donation.

[7] These chains were introduced by Roth et al., (2006), and the proof of concept was documented by Rees et al., (2009).

In our model, we allow for the possibility that a patient who pledges to donate in the future ends up not donating (for whatever reason); in the steady-state of our model, what we need is for the fulfilled plasma donation pledges to cover the flow of units used by the patients who pledge (both those who do and do not end up donating in the future).

Since pay-it-forward incentives have not been used in plasma donation before, it is difficult to estimate what fraction of patients will end up fulfilling their pledges. But in any event, the plasma replenishment rate under pay-it-forward incentives depends on (i) the rate of pledge fulfillment, (ii) how many units of plasma each patient who does fulfill a pledge donates each time she does so, and finally (iii) how many times those patients donate; of these parameters, the only one recovered patients can control is (iii).

3 Model of Plasma Donation and Demand

To formalize our conceptual intuitions about the interaction between plasma donation and treatment, we develop a simple steady-state model of plasma donation and demand. In this section, we assume that each patent receives plasma from a donor of the same blood type or the allocation is for a plasma product, such as hyperimmune globulin, for which blood-type compatibility is not needed. We extend our analysis to blood-type compatible allocation of plasma in the next section.

3.1 Paying it Backward through Priority Credit

We consider a plasma rationing system that sets aside some units of plasma for clinical trial patients through a *clinical trial plasma reserve*; the rest of the plasma supply is available to be distributed through our incentive schemes through the *incentivized plasma reserve.*

We first consider a pay-it-backward incentive scheme: We suppose that each individual who donates plasma receives $v > 0$ *priority credits* that can be used to give treatment priority to family members or other close associates.[8]

The novel feature of this incentivized plasma reserve is that while the clinical trial plasma reserve capacity is set as an exogenous parameter, the incentivized plasma reserve capacity will be endogenously determined at steady-state as a function of certain population parameters as well as the priority credit scheme in place. In particular, the incentivized plasma reserve will prioritize patient groups in the following order:

1. patients who have credits (we refer to these patients as *credit-prioritized*); then

2. patients who do not have a credit (*non-prioritized*).

Within each group priority group, plasma therapy is allocated based on a well-defined rule such as a point system or a lottery.

We contrast this system with one in which no credits are provided—i.e., $v = 0$—in which, there is a set-aside reserve for clinical-trial patients and the rest of the plasma supply is rationed among the remaining patients, with all plasma being supplied through purely altruistic donation.

[8] We introduce the pay-it-forward incentive scheme in the next section.

We consider a continuum flow model over (continuous) time and analyze the system at a steady-state. Flow rates are defined as one-dimensional Lebesgue measures of sets of individuals that become available at each time.[9]

We suppose that there is a separate market for each blood type or the donated plasma is purified and pooled to produce hyperimmune globulin shots, which does not require blood-type compatibility for its administration.

Let τ be the flow clinical trial plasma reserve size. We assume that there is overdemand for the trial, so that a flow rate of $\pi^t = \tau$ of patients participate.

At steady-state we assume that there are patients who arrive to the medical system with the credit-prioritized status; we denote the steady-state flow arrival rate of these patients by π^v. Each of these patients hold a credit given to her by a plasma donor.

The remaining patients are non-prioritized; we denote their steady-state flow rate by $\pi^n \geq 0$.[10]

Some of the patients recover without any plasma therapy; we denote the flow arrival rate of these recovering patients by ω.

The plasma therapy has steady-state arrival flow rate γ. We assume for simplicity that each patient who is treated recovers.[11]

We denote the service rates for clinical-trial patients, credit-prioritized patients, and non-prioritized patients by s^t, s^v, and s^n respectively; these are the proportions of the respective populations that are treated with plasma. The flow rates of recovery for each type of patient are then $s^t\pi^t$, $s^v\pi^v$, and $s^n\pi^n$.

Plasma can only be supplied by recovered patients. The flow rate of patients who can potentially provide plasma thus has four components: $s^t\pi^t$, $s^v\pi^v$, and $s^n\pi^n$—all described in the previous paragraph—as well as patients who have recovered without plasma therapy, with flow rate ω. We assume that recovering clinical-trial patients, recovering non-prioritized patients, and recovering patients using alternative treatment models donate plasma at the same rate p.[12] We also make a simplifying worst-case scenario assumption regarding credit-prioritized patients: we assume that credit-prioritized patients who recover do not donate plasma.[13]

Thus, the steady-state plasma therapy supply flow rate is endogenously determined by

$$\gamma = p(s^t\pi^t + s^n\pi^n + \omega)k, \tag{1}$$

where p is the probability that a given recovered patient donates and k is the number of units of

[9] We denote measures of sets, i.e., flow rates, with Greek letters, while we use Latin letters for numbers and proportions.

[10] We treat $\pi^t = \tau$ as an exogenous parameter and π^n as a steady-state rate so that π^v is endogenously determined as a function of these and other population and credit system parameters at the steady-state.

[11] Our qualitative results are the same if only a proportion of treated patients recover and only a proportion of non-treated patients die.

[12] We make this assumption for simplicity; all our results are robust to relaxing it. In particular, if recovered patients who received plasma donate at a different rate than those who did not, our analysis here provides a lower bound on the total plasma stock if we take p to equal the minimum of the two donation probabilities.

[13] If we instead assumed that recovering credit-prioritized patients donate at the same rate as the other patient groups, our Propositions 1, 2, 4 and 5 would all still hold, as donation by recovered credit-prioritized patients increases the net plasma supply, and all four results provide sufficient conditions for a priority system to function under a minimum plasma supply. The qualitative conclusion of Proposition 3 that a credit system is better than an altruistic donation scheme under Assumption 1, as well as the given sufficient condition, would also continue to hold.

plasma that patient can donate.[14]

As mentioned before, each individual who donates plasma receives $v \geq 0$ priority "credits" that can be used to give treatment priority to a family member or other close associate. Patients become credit-prioritized if, and only if, some donor allocates one of her v priority credits to them; thus, we must have

$$\pi^v = p(s^t\pi^t + s^n\pi^n + \omega)qv, \tag{2}$$

where q is the proportion of credits actually redeemed. We will use $r = qv$ to denote the average number of redeemed credits used per donor, which we call the *credit redemption rate.* We refer to

$$p(k - r)$$

as the *replenishment rate* of the plasma therapy; this is the average amount of net plasma donated to the general pool per recovered patient.

Our first result states conditions that guarantee all prioritized groups have service rate 1, i.e., $s^t = 1$ and $s^v = 1$:

Proposition 1. *So long as the plasma replenishment rate is large enough to support the clinical trial, i.e.,*

$$p(k - r) \geq \frac{\tau}{\tau + \omega}, \tag{3}$$

it is possible to ensure that all clinical-trial and credit-prioritized patients receive plasma therapy, so that

$$s^t = 1 \qquad \textit{and} \qquad s^v = 1. \tag{4}$$

Proof. The total flow rate of patients who are prioritized is given as $\pi^t + \pi^v$. To serve all of them, we need (4), i.e., that

$$\gamma \geq \pi^t + \pi^v \tag{5}$$

Substituting in (1) and (2), we see that (5) is equivalent to

$$p(\pi^t + s^n\pi^n + \omega)(k - r) \geq \pi^t \iff k - \frac{\pi^t}{p(\pi^t + s^n\pi^n + \omega)} \geq r.$$

In the worst-case scenario, the service rate for non-prioritized patients would be $s^n = 0$, yielding

$$k - \frac{\pi^t}{p(\pi^t + \omega)} \geq r$$

as a sufficient condition for (5); this is precisely (3) since $\pi^t = \tau$ is the reserve size. □

[14] In the model we think of each donor as donating just once; however, the analysis is unchanged if donors can donate repeatedly and we take k to be the average total donations per-individual.

We next turn our attention to the plasma therapy service rate s^n for non-prioritized patients, which takes the form

$$s^n = \frac{\gamma - s^t\pi^t - s^v\pi^v}{\pi^n}. \tag{6}$$

Assuming that (3) holds (i.e., $s^t = 1$ and $s^v = 1$) we substitute (1), (2), and the reserve size $\pi^t = \tau$ into (6) to find:

$$s^n = \begin{cases} \frac{\omega p(k-r) - \tau\left(1-p(k-r)\right)}{\pi^n\left(1-p(k-r)\right)} & \text{if } p(k-r) < 1 \\ +\infty & \text{if } p(k-r) \geq 1. \end{cases} \tag{7}$$

There is positive feedback: raising the number of patients who recover without plasma therapy, ω, increases the steady-state service rate—and this effect is greater the larger the probability that recovering patients donate, and the more units they contribute to the system. Naturally, the service rate is also increasing in the replenishment rate.

We see from (7) that if the plasma replenishment rate is greater than 1, we will have an arbitrarily large amount of plasma available at steady-state, so that all patients will be able to be treated. On the other hand, even if the replenishment rate is less than 1, we may still be able treat everybody and end up with finite but excess supply of plasma; this is characterized by (finite) $s^n \geq 1$.

We note in particular that so long as (3) holds, we have

$$s^n \geq 0,$$

which leads to the following corollary:

Corollary 1. *So long as the plasma replenishment rate is large enough to support the clinical trial (i.e., (3) holds), the flow recovery rate of non-prioritized patients, $s^n\pi^n + \omega$, is weakly higher than the rate that would arise absent plasma donation, ω, even when all plasma-clinical-trial patients and credit-prioritized patients are treated ahead of non-prioritized patients.*

From (7), we compute that $s^n \geq 1$ whenever

$$p \geq \frac{\tau + \pi^n}{(\tau + \pi^n + \omega)(k - r)}. \tag{8}$$

We thus find:

Proposition 2. *Whenever (8) holds, it is possible to treat all patients—prioritized and non-prioritized—at steady-state. In particular, it is possible to treat all patients when replenishment rate is above replacement; that is, when*

$$p(k - r) \geq \frac{\tau + \pi^n}{\tau + \pi^n + \omega}.$$

3.1.1 Altruistic Donation vs. Incentivized Backward Donation

Additionally, we can think of p in terms of a supply curve $p(\,\cdot\,)$ that is strictly increasing and differentiable as a function of the credit redemption rate, r. Thus, $p(0)$ refers to the altruistic donation probability (which is what would arise without any incentive scheme involving prioritization through credits).

We make the following assumption:

Assumption 1. The replenishment rate $p(r)\cdot(k-r)$ is strictly increasing at $r=0$ (i.e., $p'(0)k > p(0)$).

Assumption 1 is fairly mild; it is satisfied if a sufficiently small percentage of recovering patients donate altruistically without any credit scheme in place. Under Assumption 1, assuming an interior maximum $s^* < 1$ (i.e., $s=1$ cannot be achieved no matter what r is), our expression (7) for s^n implicitly defines the optimal r through the necessary first-order condition:

$$0 = \frac{ds^n}{dr} = \frac{d}{dr}\left[\frac{\omega\cdot p(r)\cdot(k-r) - \tau\big(1-p(r)\cdot(k-r)\big)}{\pi^n\big(1-p(r)\cdot(k-r)\big)}\right],$$

so that we have

$$\frac{p'(r^*)}{p(r^*)} = \frac{1}{k-r^*}. \tag{9}$$

Observe that the r^* in (9) is also the value that maximizes the replenishment rate $p(r)\cdot(k-r)$.[15] Such an interior maximum exists for the service rate because the service rate is increasing in the replenishment rate and the replenishment rate is increasing at $r=0$ by Assumption 1 (and hence is positive at a small $r\approx 0$); moreover the service rate falls back to 0 when r satisfies (3) with equality.

We summarize our findings with the following proposition:

Proposition 3. *Under Assumption 1, so long as the plasma replenishment rate is large enough to support the clinical trial, (i.e., (3) holds) the credit redemption rate that maximizes the plasma service rate for non-prioritized patients satisfies $r^* > 0$—that is, using a credit scheme strictly improves outcomes for non-prioritized patients.*

Moreover, the service rate for non-prioritized patients s^n is strictly increasing in the plasma replenishment rate $p(r)\cdot(k-r)$ and is maximized either

- *at $s^{n*}=1$ by all credit redemption rates r that satisfy (8), or*
- *at some $s^{n*}<1$ (if there is no r such that we can have $s^n=1$) by a credit redemption rate $r^*>0$ satisfying (9).*

3.2 Paying it Forward through a Pledge of Future Donation

We now suppose that there is also a pathway some patients can use to gain priority for treatment, which is to pledge upfront to donate plasma upon recovery. We suppose that in addition to upfront treatment, we give such a patient $v^f \geq 0$ credits after (and if) she donates plasma.[16]

As before, we set aside a reserve for clinical-trial patients with the flow capacity τ. The rest of the plasma therapy is allocated within the incentivized plasma reserve, which now has three priority classes ordered as follows:

[15] If there are multiple such values, we pick the one among them that achieves the highest service rate s^n.

[16] We may also count the treatment of the pledged patient herself as the upfront redemption of a credit, in which case we would think of this patient as receiving credits to treat as many as v^f+1 patients, including herself.

1. patients who have credit (whom we refer to as *credit-prioritized*, as before);
2. patients who do not have credits but pledge to donate after they recover (*pledged* patients); and
3. patients not in any of the other categories (*non-prioritized* patients).

We denote the steady-state flow rate of patients participating in clinical trial by $\hat{\pi}^t = \tau$; the flow rate of credit-prioritized patients by $\hat{\pi}^v$; the flow rate of pledged patients by $\hat{\pi}^f$; and the flow rate of non-prioritized patients by

$$\hat{\pi}^n = \pi^n - \hat{\pi}^f \leq \pi^n.$$

We refer to the different types of patients' respective plasma therapy service rates as $\hat{s}^t$, $\hat{s}^v$, $\hat{s}^f$, and $\hat{s}^n$.

Then the total flow rate of recovering patients has four components:

- patients who participate in clinical trials, with a flow rate $\hat{s}^t\hat{\pi}^t$;
- patients who are credit-prioritized, with a flow rate $\hat{s}^v\hat{\pi}^v$;[17]
- patients who have pledged to donate ex ante, with a flow rate $\hat{s}^f\hat{\pi}^f$; and
- patients who are not part of clinical trials, do not have credits, and have not pledged to donate, with a flow rate of $\hat{s}^n\hat{\pi}^n + \omega$.

The total steady-state flow of plasma therapy is

$$\hat{\gamma} = \big(p(\hat{s}^t\hat{\pi}^t + \hat{s}^n\hat{\pi}^n + \omega) + p^f\hat{s}^f\hat{\pi}^f\big)k, \tag{10}$$

where p is the population probability to donate in return for credits (as in the prior sections) and p^f is the probability with which pledged patients donate upon recovery. We allow the possibility that some patients who pledge may not end up donating—perhaps due to medical ineligibility—so that p^f is expected to be less than 1. We only assume that pledging increases one's probability of donation, so that $p^f \geq p$.

We assume that patients who decide to donate ex post each receive v priority credits to be used by their loved ones, as before. On the other hand, pledged patients possibly also receive a number of vouchers upon recovery and donation—if they they donate k units of plasma, they receive v^f credits. The v^f credits are only given after the pledged recovering patient "pays it forward" by donating plasma, which occurs with probability p^f.

Thus, the flow rate of credit-prioritized patients $\hat{\pi}^v$ satisfies

$$\hat{\pi}^v = p(\hat{s}^t\hat{\pi}^t + \hat{s}^n\hat{\pi}^n + \omega)qv + p^f\hat{s}^f\hat{\pi}^f qv^f. \tag{11}$$

[17] As before, we conduct a worst-case analysis under the assumption that patients who have credits do not become plasma donors upon recovery. Propositions 4 and 5 continue to hold if we assume credit-prioritized patients also donate with probability p upon recovery.

As before, we will work with the credit redemption rates

$$r = qv$$

for the patients who have not pledged ex ante but decide to donate upon recovery. Similarly, for pledged patients, we write:

$$r^f = qv^f.$$

The following proposition gives conditions under which we can fully serve all prioritized patient groups (i.e., so that $\hat{s}^t = 1$, $\hat{s}^v = 1$, and $\hat{s}^f = 1$):

Proposition 4. *Regardless of the pledged patient arrival rate $\hat{\pi}^f$, so long as we have*

$$p(k - r) \geq \frac{\tau}{\tau + \omega} \qquad \textit{and} \qquad p^f(k - r^f) \geq 1, \tag{12}$$

it is possible to ensure that all clinical-trial patients, credit-prioritized patients, and pledged patients receive plasma therapy, so that

$$\hat{s}^n = 1, \qquad \hat{s}^v = 1, \qquad \textit{and} \qquad \hat{s}^f = 1.$$

Proof. Clinical-trial patients, credit-prioritized patients, and pledged patients are prioritized over non-pledged patients. Thus, by setting $\hat{s}^t = \hat{s}^v = \hat{s}^f = 1$ and using (10) and (11), we see that all prioritized patient groups can all be treated by plasma if

$$\hat{\gamma} \geq \hat{\pi}^t + \hat{\pi}^v + \hat{\pi}^f \iff p(\hat{\pi}^t + \hat{s}^n\hat{\pi}^n + \omega)(k - r) + p^f\hat{\pi}^f(k - r^f) \geq \hat{\pi}^t + \hat{\pi}^f. \tag{13}$$

To capture the minimum amount of plasma needed to treat all pledged patients, we consider the worst-case scenario in which no non-prioritized patients are treated, i.e., $\hat{s}^n = 0$. Then necessary and sufficient conditions for (13) to be satisfied regardless of $\hat{\pi}^f$ are

$$p(k - r) \geq \frac{\hat{\pi}^t}{p(\hat{\pi}^t + \omega)} \qquad \text{and} \qquad p^f(k - r^f) \geq 1. \tag{14}$$

Replacing $\hat{\pi}^t$ with τ in (14), we obtain (12). □

The first condition in (12) is the same condition as (3): The replenishment rate of the plasma obtained from initially non-pledged patients should be at least as large as is needed to support the clinical trial plasma reserve. The second condition in (12) requires that the replenishment rate of plasma obtained from pledged patients should at least cover those patients' own initial treatment in steady-state.

We now examine the plasma service rate for non-prioritized patients when (12) holds:

$$\hat{s}^n = \frac{\hat{\gamma} - \hat{s}^t\hat{\pi}^t - \hat{s}^v\hat{\pi}^v - \hat{s}^f\hat{\pi}^f}{\hat{\pi}^n}. \tag{15}$$

Expanding (15) assuming $\hat{s}^v = 1$, we find that

$$\hat{s}^n = \frac{\big(p(\hat{s}^t\hat{\pi}^t + \hat{s}^n\hat{\pi}^n + \omega)(k-r)\big) - \hat{s}^t\hat{\pi}^t + \big(p^f\hat{s}^f\hat{\pi}^f(k-r^f)\big) - \hat{s}^f\hat{\pi}^f}{\hat{\pi}^n}. \tag{16}$$

Solving (16) for $\hat{s}^n$ (replacing $\hat{\pi}^t = \tau$ and $\hat{s}^t = 1$), we see that, assuming the pay-it-backward credit replenishment rate does not on its own lead to infinite excess supply of plasma (i.e., $p(1-r) < 1$),

$$\hat{s}^n = \frac{\omega p(k-r) - \tau\big(1 - p(k-r)\big) + p^f\hat{s}^f\hat{\pi}^f\big(k - r^f - \frac{1}{p^f}\big)}{\hat{\pi}^n\big(1-p(k-r)\big)}. \tag{17}$$

Comparing (17) to (7), we see that non-prioritized patients are served at a weakly higher rate than they would be under a system that does not prioritize pledged patients whenever

$$\frac{\omega p(k-r) - \tau\big(1 - p(k-r)\big) + p^f\hat{s}^f\hat{\pi}^f\big(k - r^f - \frac{1}{p^f}\big)}{\hat{\pi}^n\big(1-p(k-r)\big)}$$
$$= \hat{s}^n \geq \frac{\pi^n}{\hat{\pi}^n}s^n = \frac{\pi^n}{\hat{\pi}^n}\left(\frac{\omega p(k-r) - \tau\big(1-p(k-r)\big)}{\pi^n\big(1-p(k-r)\big)}\right).$$

Thus, we find that $\hat{s}^n \geq s^n$ when (12) holds, and conclude:

Proposition 5. *So long as (12) holds, besides treating every clinical-trial patient and credit-prioritized patient ($\hat{s}^t = \hat{s}^v = 1$), it is possible to treat every patient who pledges to donate plasma upfront ($\hat{s}^f = 1$), while still raising the service rate for non-prioritized patients who have not pledged to donate.*

4 ABO Blood-type Compatible Plasma Donation

We now build on the analysis from the preceding section to allow patients receive donation from plasma-compatible donors who do not necessarily have identical blood type.

There are four blood types O, A, B, and AB. Type AB plasma can be used to treat patients of all blood types; blood-type A plasma can be used to treat patients of blood types O and A; blood-type B plasma can be used to treat patients of blood types O and B; and blood-type O plasma can only be used to treat patients of blood type O. (Since convalescent plasma is a type of plasma, those same compatibility requirements are needed for plasma transfusion.) We let $\mathcal{B} = \{O, A, B, AB\}$ be the set of blood types.

Under an ABO-identical treatment policy, non-prioritized patients of different blood types may be served in unequal service rates because the parameters $\hat{\pi}^f_X/\pi^n_X$, ω_X/π^n_X, p_X, and p^f_X may vary based on blood-type $X \in \mathcal{B}$ even if the voucher redemption rates r_X and r^f_X are chosen to take these differences into account. The main reason behind this variation is due to the location of outbreaks and differences in people's ability to socially distance, COVID-19 has so far affected some national and ethnic groups more than others. Moreover, some blood types may have excess supply of plasma while the others do not; for example, (8) may hold for some blood types while it does not for others.

We aim for an egalitarian service policy for plasma therapy with multiple blood types—thus we seek to make the non-prioritized patient service rates of different blood types as equal as possible without affecting efficiency.

We need to account for voucher holders possibly having different blood types from their original donors; we assume that their blood types are independently distributed from their donors'. Suppose b_X is the probability that a given patient is of blood type X. Let

$$r_X = b_X \sum_{Y \in \mathcal{B}} q_Y v_Y$$

be the voucher redemption rate for backward donation, and let

$$r_X^f = b_X \sum_{Y \in \mathcal{B}} q_Y v_Y^f$$

be the voucher redemption rate for forward donation.

We refer to the service rates for non-prioritized patients for each blood type X given in (17) as the ABO-identical service rate, and rephrase it here once more assuming all clinical-trial patients, voucher-prioritized patients and pledged patients are served, i.e., $\hat{s}_X^t = 1$, $\hat{s}_X^v = 1$, and $\hat{s}_X^f = 1$. Define

$$\sigma_X := \omega_X p_X(k - r_X) - \tau_X\big(1 - p_X(k - r_X)\big) + p_X^f \hat{\pi}_X^f \left(k - r_X^f - \frac{1}{p_X^f}\right) \tag{18}$$

$$\delta_X := \hat{\pi}_X^n\big(1 - p_X(k - r_X)\big) \tag{19}$$

for each blood type X. Here, σ_X is the *steady-state net supply* of blood-type X plasma to be rationed to non-prioritized patients while δ_X is the *steady-state net demand* for plasma by non-prioritized blood-type X patients.

4.1 Pooling for Plasma Treatment

Whenever, $\delta_X < 0$, which happens when the plasma replenishment rate for X is greater than 1, the blood-type X non-prioritized patients are self-sufficient, and we can distribute the remaining plasma to other compatible blood types to serve all of them.[18] Thus, assume that replenishment rate $p_X(k - r_X) < 1$ for at least one blood type $X \in \mathcal{B}$, as otherwise all blood types will be self-sufficient and non-prioritized patients who survive donate enough plasma on net to supply future generations of patients.

Moreover, assuming $p_X(k - r_X) < 1$, we observe that σ_X is the numerator and δ_X is the denominator of $\hat{s}_X^n$ in (17)

$$\hat{s}_X^n = \frac{\sigma_X}{\delta_X}. \tag{20}$$

Another way the excess plasma of one blood type can be used for other blood types is that if $\delta_X > 0$ and still $\sigma_X > \delta_X$. Suppose as an example, for $\delta_O, \delta_A > 0$ we have,

Since blood-type O patients can receive blood-type A plasma, for an egalitarian plasma allocation,

[18] Relative to our model as presented in the previous section, this is the case in which we obtain infinite supply of blood-type X plasma in the steady-state.

we can give some of the blood-type A plasma to blood-type O patients and increase the service rate for O patients and decrease the service rate for A patients. Let $\sigma_{A\to O}$ be the resulting net transfer flow of blood-type A plasma to blood-type O patients.

Then, the new service rates of both types will be

$$s_O = \frac{\sigma_O + \sigma_{A\to O}}{\delta_O} \le s_A = \frac{\sigma_A - \sigma_{A\to O}}{\delta_A}. \tag{21}$$

We can continue increasing the net transfer $\sigma_{A\to O}$ until both service rates become equal, to sustain an egalitarian service rate among the two blood types. Either we will eventually have both service rates exceeding 1, and hence all of these patients are served, or we will end up with an equal service rate for A and O less than 1. Observe that the amount of plasma transfer from A to O that makes (21) hold with equality is

$$\sigma_{A\to O} = \frac{\sigma_A\delta_O - \sigma_O\delta_A}{\delta_O + \delta_A}, \tag{22}$$

which is strictly greater than 0 (by (18) and $\delta_O, \delta_A > 0$) and strictly smaller than σ_A (as $\delta_O, \delta_A > 0$).

This resulting service rate, what we call the *pooling service rate* for A and O is then

$$\hat{s}^n_{\{O,A\}} := \frac{\sigma_O + \sigma_A}{\delta_O + \delta_A} = s_O = s_A. \tag{23}$$

Observe that (23) treats patients as if A and O together form a "composite blood type" and yet the subsidy of plasma is one way: some blood-type A plasma is used to treat blood-type O patients, but blood-type O plasma is never used on blood-type A patients (as it would not be compatible).

As $\sigma_A, \sigma_O, \delta_A, \delta_O > 0$, we have

$$\hat{s}^n_O = \frac{\sigma_O}{\delta_O} < \hat{s}^n_{\{O,A\}} < \hat{s}^n_A = \frac{\sigma_A}{\delta_A}.$$

Additionally, if the service rate for B, $\hat{s}^n_B$ is larger than the pooled rate in (22) but lower than $\hat{s}^n_A$, we can further subsidize blood-type O patients and return some of the blood-type A plasma that was earmarked for O patients in (21) back to blood-type A patients.[19] Eventually, we would end up with a pooled service rate for the blood types $\{O, A, B\}$; as long as $\delta_B > 0$, we would have

$$\hat{s}^n_{\{O,A\}} < \hat{s}^n_{\{O,A,B\}} = \frac{\sigma_O + \sigma_A + \sigma_B}{\delta_O + \delta_A + \delta_B} < \hat{s}^n_B.$$

4.2 Optimal Pooling

We now introduce a formal iterative pooling procedure to determine the service rates of non-prioritized patients when there are four blood types.[20]

[19] If $\hat{s}^n_B > \hat{s}^n_A$, then we would start with blood-type B plasma to subsidize blood-type O patients and then check later for further blood-type A plasma subsidy opportunities.

[20] The procedure discussed here subsumes the procedure that was introduced in an earlier working paper, Sönmez and Ünver, (2015).

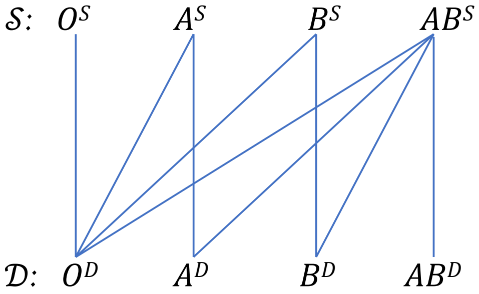

Figure 1: The plasma compatibility graph $(\mathcal{S}, \mathcal{D}, \mathbf{C})$.

Consider the following bipartite graph (see Figure 1), with sides labeled $\mathcal{S}$ and $\mathcal{D}$. Either side consists of the 4 nodes, one for each blood type:

$$\mathcal{S} = \{O^S, A^S, B^S, AB^S\},$$
$$\mathcal{D} = \{O^D, A^D, B^D, AB^D\}.$$

Each node in $\mathcal{S}$ represents the plasma *supply* of the corresponding blood type; each node in $\mathcal{D}$ represents *demand*, i.e., the patients of the corresponding blood type. The nodes in both sides are connected with undirected *blood-type compatibility* edges, when possible. Each edge is represented as $\{X^D, Y^S\} \in \mathcal{S} \cup \mathcal{D}$, which means that blood-type Y plasma can be transfused to blood-type X patients; we let $\mathbf{C}$ be the set of edges. We refer to $(\mathcal{S}, \mathcal{D}, \mathbf{C})$ as the *plasma compatibility graph.*

For any $\mathcal{D}' \subseteq \mathcal{D}$ and $\mathcal{S}' \subseteq \mathcal{S}$, define

$$\mathcal{C}_{\mathcal{D}'}(\mathcal{S}') := \big\{Y^S \in \mathcal{S}' \ : \ \{Y^S, X^D\} \in \mathbf{C} \text{ for some } X^D \in \mathcal{D}'\big\};$$

set $\mathcal{C}_{\mathcal{D}'}(\mathcal{S}')$ is the set of plasma supply nodes in $\mathcal{S}'$ that are compatible with the patients of blood types in $\mathcal{D}'$. We also define

$$s_{\mathcal{D}'}(\mathcal{S}') := \frac{\sum_{Y^S \in \mathcal{C}_{\mathcal{D}'}(\mathcal{S}')} \sigma_Y}{\sum_{X^D \in \mathcal{D}'} \delta_X}, \tag{24}$$

which is the supply-to-demand ratio for demand nodes in $\mathcal{D}'$ when the supply nodes in $\mathcal{S}'$ are exclusively available for them. The ratio $s_{\mathcal{D}'}(\mathcal{S}')$ is our generalization of the service rates for non-prioritized patients of a certain blood-type X under ABO-identical plasma rationing as defined in (20).

The steps of the pooling construction are as follows:

Plasma Pooling Procedure – We iteratively construct partitions $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{\bar{\ell}}$ of $\mathcal{S}$ and $\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_{\bar{\ell}}$ of $\mathcal{D}$ as follows:

Step 0. Define

$$\mathcal{D}_0 := \big\{X^D \in \mathcal{D} : \delta_Y < 0 \text{ for some } Y^S \in \mathcal{C}_{\{X^D\}}(\mathcal{S})\big\}, \tag{25}$$
$$\mathcal{S}_0 := \big\{X^S \in \mathcal{S} : \delta_Y < 0 \text{ for some } Y^S \in \mathcal{C}_{\{X^D\}}(\mathcal{S})\big\}. \tag{26}$$

If $\delta_Y < 0$, then the plasma replenishment rate for blood-type Y is greater than 1, so there is arbitrarily large steady-state supply—and the patients of the demand nodes in $\mathcal{D}_0$ can be fully served with that plasma; meanwhile, the plasma of any supply node $X^S \in \mathcal{S}_0$ will not be required in the pooling procedure because any blood type that blood-type X plasma can serve is also served by blood-type Y plasma. For each $X^D \in \mathcal{D}_0$, set the service rate of non-prioritized blood-type X patients to

$$\hat{s}^n_X := 1.$$

$\vdots$

Step $\boldsymbol{\ell \geq 1}$: Given sets $\mathcal{S}_0, \ldots, \mathcal{S}_{\ell-1}$ and $\mathcal{D}_0, \ldots, \mathcal{D}_{\ell-1}$, we find the

$$\mathcal{D}' \subseteq \mathcal{D} \setminus \cup_{m=0}^{\ell-1} \mathcal{D}_m,$$

that minimizes $s_{\mathcal{D}'}(\mathcal{S} \setminus \cup_{m=0}^{\ell-1} \mathcal{S}_m)$:[21]

$$\mathcal{D}_\ell := \underset{\mathcal{D}' \subseteq \mathcal{D} \setminus \cup_{m=0}^{\ell-1} \mathcal{D}_m}{\arg\min} \left\{ s_{\mathcal{D}_\ell}\Big(\mathcal{S} \setminus \cup_{m=0}^{\ell-1} \mathcal{S}_m\Big) \right\}, \tag{27}$$

$$\mathcal{S}_\ell := \mathcal{C}_{\mathcal{D}_\ell}\Big(\mathcal{S} \setminus \cup_{m=0}^{\ell-1} \mathcal{S}_m\Big). \tag{28}$$

We allocate all plasma of supply nodes in $\mathcal{S}_\ell$ to patients of demand nodes in $\mathcal{D}_\ell$ and for each $X^D \in \mathcal{D}_\ell$, we set the service rate of non-prioritized patients of blood type X to be

$$\hat{s}^n_X := \min\Big\{1,\ s_{\mathcal{D}_\ell}(\mathcal{S}_\ell)\Big\}. \tag{29}$$

If $\mathcal{D} \setminus \cup_{m=0}^{\ell} \mathcal{D}_m \neq \emptyset$, then we continue with Step $\ell + 1$.

The pooling procedure just described first identifies patient types that can be completely served using plasma that is in excess supply. Then, we iteratively match plasma to patients by finding the patients who are hardest to serve with the remaining supply and assigning them as much plasma as possible.

We first demonstrate that the procedure we have just described is *feasible*, in the sense that it never over-allocates plasma supply:

Proposition 6. *The plasma pooling procedure generates a feasible outcome service-rate vector.*

In the Appendix, we prove Proposition 6 using a general method based on maximum-flow theory. The argument we give does not depend on the structure of plasma donation compatibility, and thus may be of independent interest for other divisible goods rationing problems with compatibility requirements.

As we might naturally expect given that our pooling procedure serves the hardest-to-serve patients in each step, the service-rate vector we obtain has an intuitive fairness property. Given a service rate

[21] If there is more than one such set, then we take the largest of them, which is uniquely defined.

vector s, *the largest service rate under* s is

$$\max_{X \in \mathcal{B}} \{s_X\}$$

and *the smallest service rate under* s is

$$\min_{X \in \mathcal{B}} \{s_X\}.$$

Proposition 7. *The service rate vector obtained through our plasma pooling procedure*

- *maximizes the smallest service rate and*
- *minimizes the difference between the largest and smallest service rates*

among service rate vectors that maximize the measure of patients served (under given pay-it-backward and pay-it-forward voucher redemption rates $(r_X, r^f_X)_{X \in \mathcal{B}}$ *and clinical-trial reserves* $(\tau_X)_{X \in \mathcal{B}}$*).*

The proof of Proposition 7 is also in the Appendix.

5 Related Literature

To our knowledge, this paper is the first to propose a market design approach to plasma donation. That said, several of the key insights and tools in our proposed mechanisms for increasing plasma donation in have parallels in the kidney exchange literature (see, e.g., Roth, Sönmez, and Ünver, 2004, 2005a,b). Within that literature, our model is most closely related to that of Sönmez, Ünver, and Yenmez, (2020), who introduced a dynamic continuum matching model to study the effects of incentivizing compatible kidney donor-patient pairs to participate in exchange by providing increased priority in the deceased-donor queue. The most important difference is that patients and donors are distinct in Sönmez, Ünver, and Yenmez, (2020), whereas in our model they are the same population. The incentive schemes we propose exploit the fact that patients can go on to become donors unlike kidney exchange settings.

There are parallels between the pay-it-forward and pay-it-backward idea in kidney exchange. *Non-directed donor chains* involve paying it forward (Roth et al., 2006; Rees et al., 2009). In such a chain, each participating incompatible patient-donor pair first receives a kidney donation for their patient and at a later date their donor returns the favor by donating a kidney to another pair. These chains start with the gift of an altruistic donor, and can lead to quite long sequences of donations. Intertemporal incentives in kidney exchange also relate to the paying it backward concept. In a patient-donor pair where the patient is not ready for a transplant yet, the donor will no longer be eligible for donation when the patient is expected to need a transplant in the future (perhaps due to donor age). Veale et al., (2017) report on a kidney voucher system where an older living donor of a young patient starts a chain of kidney exchanges through donation to an incompatible pair. Since the younger patient will likely need a kidney in the future, the patient receives priority for a kidney at the end of a similar future chain if her kidney fails. Since the donor is old, the window for donation is short and the scheme helps other pairs receive transplants through chain exchanges in the present and in some sense "insures" the initial patient paired with the donor. Similarly, Akbarpour et al., (2019) study unpaired kidney exchange, where a patient i can receive a kidney from patient j and the system will remember that patient j has the right to receive a kidney in the future.

Since plasma is part of blood, our work is also related to research on the design of blood markets. Slonim, Wang, and Garbarino, (2014) provide a recent summary, and show that providing donors some form of non-monetary incentive, such as a medal or trinket increases donation. Lacetera, Macis, and Slonim, (2013) report that 18 of 19 different incentives in observational or field experimental studies increase blood donation. The responsiveness of blood donation to incentives suggests that a voucher may increase convalescent plasma donation rates. Heger et al., (forthcoming) have proposed introducing a registry for prospective blood donors. There is also precedent for the formation of a centralized plasma bank during a pandemic. Delamou et al., (2016), for example, report on the Guinean National Blood Transfusion Center, which involved donor mobilization and plasma collection for Ebola therapy in 2015.

Our paper is also related to schemes used to incentivize donation of solid organs in other countries by using pledges to donate. Singapore has a presumed consent/opt-out policy for donation of cadaveric kidneys, livers, hearts, and corneas. If someone does not want to donate a particular organ, they would receive lower priority for receiving that particular organ (Singapore, 2012). In Israel, a patient who holds a donor card or is a first-degree relative with a donor card obtains priority over patients who do not. To obtain a donor card, the individual has to opt-in to donation (Lavee et al., 2010). In Chile, an individual who does not wish to donate their organ would receive lower priority for organ transplantation than a registered person if there is equal need and compatibility (Zúniga-Fajuri, 2015).

Last, we note that our continuum model is related to a growing literature in matching theory that considers large-market models. Large-market models oriented towards market-design applications include those of Kojima and Pathak, (2009), Che and Kojima, (2010), Abdulkadiroğlu, Che, and Yasuda, (2011), Azevedo and Leshno, (2016), Azevedo and Hatfield, (2018), Azevedo and Budish, (2019), Che, Kim, and Kojima, (2019), and Che and Tercieux, (2019). Our steady-state analysis is also related to recent models of dynamic matching markets, such as the work of Ünver, (2010), Anderson et al., (2017), Baccara, Lee, and Yariv, (2018), and Akbarpour, Li, and Gharan, (2020).

6 Conclusion

In this paper, we propose a market design approach to convalescent plasma donation and distribution. Plasma donors may be given credits that can be used to give treatment priority to their loved ones; priority is also given to participants in clinical trials. Our model illustrates important possibilities: if the plasma replenishment rate is large enough to support the patients in a clinical trial, it is possible to treat all prioritized patients in equilibrium. There is also a positive spillover for non-prioritized patients. Moreover, if recovered patients are more willing to donate if they receive credits, introducing a credit system strictly benefits non-prioritized patients. Overall treatment availability expands further if we prioritize patients who pledge to “pay it forward” by donating plasma once they have recovered.

In terms of the model, there are several directions for future work. Our analysis has focused on the steady-state for analytical convenience. Since both convalescent plasma supply and demand are evolving rapidly, it will be important to understand transition dynamics leading to a steady-state.

Second, we did not consider the possibility of other allocation systems, including those based on price. As far as we know, there is no current market where infected patients can buy convalescent plasma or where recovered patients can sell their plasma. However, as the market matures, these institutions may develop, and it is worth understanding how they relate to our model.

Our model has been motivated by convalescent plasma, but our ideas can apply more generally to increase supply for other human products. The key property necessary for our pay it forward incentive to work is that a patient who receives treatment can go on to become a donor in the future. There are several other products for which this property holds, including other blood components.

After we circulated the first version of this paper, we were approached by the leadership at the US Covid plasma initiative (http://covidplasmasavealife.com), a leading network of patients, donors, and hospitals which has supplied more than half of the plasma to the Mayo Clinic's expanded access program (Stack, 2020). The leadership inquired about a variant of our credit system for increasing plasma donation for their hyperimmune globulin product under development. Pending clinical and regulatory approval of their product, they intend to use our credit system to collect plasma in their program (see https://www.covidplasmasavealife.com/hig).

Appendix A Proof of Propositions 6 and 7

Instead of a brute-force approach tailored for only 4 blood types and the particular plasma compatibility digraph, we give a proof that is general and can be used even if there were many other "blood types" and the compatibility digraph were constructed (arbitrarily) using those blood types. A more plausible application of this general version of the result is constructing an egalitarian rationing scheme for a divisible good with compatibility constraints.

A.1 Preliminaries

We use the concept of *flow networks* developed in the combinatorial optimization and graph theory literature (see, e.g., Korte and Vygen, (2012) for a survey). We construct flow networks that are isomorphic to our ABO-compatible plasma rationing problem; the flows on the network will correspond to feasible plasma allocation policies. We then use our flow networks to show that for each $\ell \in \{1, ..., \bar{\ell}\}$, for each demand node $X^D \in \mathcal{D}_\ell$ (as defined in (27)), we can feasibly serve non-prioritized patients of blood type X at the rate defined in (29) using plasma from the corresponding supply nodes in $\mathcal{S}_\ell$.

A.1.1 Flow Networks Isomorphic to the Rationing Problem and the Maximum Flow–Minimum Cut Theorem

An *ABO-compatible rationing flow network* is defined through an acyclic digraph with *nodes* $\mathcal{N} = \{E, K\} \cup (\mathcal{D} \setminus \mathcal{D}_0) \cup (\mathcal{S} \setminus \mathcal{S}_0)$ such that node E is referred to as the *source*; node K is referred to as the *sink*; set $\mathcal{D}_0$ is defined in (25); and set $\mathcal{S}_0$ is defined in (26).[22] A *directed edge* of the flow network originating from node i and pointing at node j is denoted by (i, j). We define the set $\mathbf{E}$ of edges as follows:

- $(E, X^D) \in \mathbf{E}$ for each $X^D \in \mathcal{D} \setminus \mathcal{D}_0$,
- $(Y^S, K) \in \mathbf{E}$ for each $Y^S \in \mathcal{S} \setminus \mathcal{S}_0$,
- $(X^D, Y^S) \in \mathbf{E} \iff \{X^D, Y^S\} \in \mathbf{C}$ for each $X^D \in \mathcal{D} \setminus \mathcal{D}_0$ and $Y^S \in \mathcal{S} \setminus \mathcal{S}_0$.

A flow network of the acyclic digraph $(\mathcal{N}, \mathbf{E})$ carries *flows*, which we will formally define below as a function, from source E through the edges of the graph to the sink K. Each edge $(i, j) \in \mathbf{E}$ has a *capacity* $\kappa(i, j) > 0$ denoting the maximum flow it can carry. Let $\kappa = (\kappa(i, j))_{(i,j)\in\mathbf{E}}$ denote the capacity vector for all the edges.

Formally, a *flow network* of the digraph $(\mathcal{N}, \mathbf{E})$ is denoted by the pair $(\mathcal{N}, \kappa)$. A *flow function* $\varphi : \mathbf{E} \to \mathbb{R}^+$ is a mapping such that we have

(i) if $(i, j) \in \mathbf{E}$ then $\varphi(i, j) \leq \kappa(i, j)$;

(ii) if $i \in \mathcal{N} \setminus \{E, K\}$, then $\sum_{h\in\mathcal{N}:(h,i)\in\mathbf{E}} \varphi(h, i) = \sum_{h\in\mathcal{N}:(i,h)\in\mathbf{E}} \varphi(i, h)$.

[22] Note that we can retain $\mathcal{S}_0$ in the set of nodes and use $\mathcal{S}$ as the set of supply nodes included in the flow network. In this case, all of our results will go through. For more general compatibility graphs than plasma donation this should be done.

Property (i) says that no edge can carry flow higher than its capacity. Property (ii) says that for any node other than the source and the sink, the total flow entering must be equal to the total flow exiting. We refer to $\varphi(i,j)$ as the *flow from node i to j under φ*. Let $\Phi(\mathcal{N},\kappa)$ be the set of flow functions defined for the flow network $(\mathcal{N},\kappa)$.

A *cut* of the network is a subset of nodes $\mathcal{V}$ such that $\{E\} \subseteq \mathcal{V} \subseteq \mathcal{N} \setminus \{K\}$. The *total capacity of a cut* $\mathcal{V}$ is

$$\kappa(\mathcal{V}) := \sum_{i\in\mathcal{V},\, j\in\mathcal{N}\setminus\mathcal{V}\,:\,(i,j)\in\mathbf{E}} \kappa(i,j);$$

$\kappa(\mathcal{V})$ is the sum of the capacities of edges originating from a node in $\mathcal{V}$ and ending at a outside $\mathcal{V}$. A *minimum cut* is a cut with the minimum possible total capacity, i.e., a cut $\mathcal{V}$ such that

$$\kappa(\mathcal{V}) = \min_{\{E\}\subseteq\mathcal{V}'\subseteq\mathcal{N}\setminus\{K\}} \{\kappa(\mathcal{V}')\}.$$

Given a flow function $\varphi \in \Phi(\mathcal{N},\kappa)$, the *flow from cut $\mathcal{V}$ (to $\mathcal{N}\setminus\mathcal{V}$) under φ* is denoted by

$$\varphi(\mathcal{V}) := \sum_{i\in\mathcal{V},\, j\in\mathcal{N}\setminus\mathcal{V}\,:\,(i,j)\in\mathbf{E}} \varphi(i,j);$$

this is the total flow carried by the directed edges that start at a node in $\mathcal{V}$ and end at a node in $\mathcal{N}\setminus\mathcal{V}$.

The *value of* φ is the flow under φ from cut $\mathcal{N}\setminus\{K\}$ (to $\{K\}$), which is also equal to the flow from cut $\{E\}$ (to $\mathcal{N}\setminus\{E\}$). The *maximum value* of a flow function over the flow network $(\mathcal{N},\kappa)$ is defined as

$$\max_{\varphi\in\Phi(N,\kappa)} \varphi(\mathcal{N}\setminus\{K\}) = \max_{\varphi\in\Phi(\mathcal{N},\kappa)} \{\varphi(\{E\})\}.$$

We refer to any flow function $\varphi^* \in \arg\max_{\varphi\in\Phi(\mathcal{N},\kappa)}\{\varphi(\{E\})\}$ as a *maximum flow function.*

The following fundamental theorem relates the capacities of the edges to the maximum flow that can be carried over a network:

Theorem 1 (Maximum Flow–Minimum Cut Theorem (Ford and Fulkerson, 1956))**.** *The value of a maximum flow function over a flow network is equal to the total capacity of one of the network's minimum cuts.*

A.1.2 Egalitarian Rationing and x-Parametrized Flow Networks

We now consider a continuum of flow networks $(\mathcal{N},\kappa^x)$ on the digraph $(\mathcal{N},\mathbf{E})$ where the edge capacities from the source are parametrized by a vector $x = (x_X)_{X^D\in\mathcal{D}\setminus\mathcal{D}_0}$ such that we have $x_X \in [0,1]$ for each X (see Figure 2 for an illustration):

- for any edge $(X^D,Y^S)\in\mathbf{E}$, $\kappa^x(X^D,Y^S) := +\infty$ (and hence, (X^D,Y^S) can carry any flow);
- for edge $(X^S,K)\in\mathbf{E}$ (for any $X^S\in\mathcal{S}\setminus\mathcal{S}_0$), $\kappa^x(X^S,K) := \sigma_X$ is constant and equal to the steady-state supply of blood-type X plasma defined in (18); and
- for edge $(E,X^D)\in\mathbf{E}$ (for any $X^D\in\mathcal{D}\setminus\mathcal{D}_0$), $\kappa^x(E,X^D) := x_X\cdot\delta_X$, where δ_X is the steady-state net demand of blood-type X patients defined in (19).

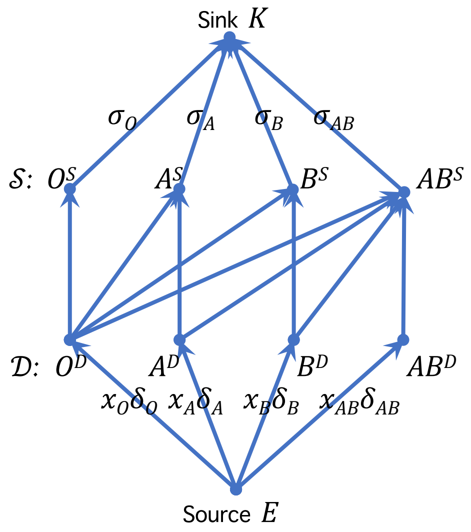

Figure 2: The x–parametric flow network when $\mathcal{D}_0 = \emptyset$ and $\mathcal{S}_0 = \emptyset$ for the proof of Propositions 6 and 7, which uses the plasma compatibility graph in Figure 1.

We refer to flow network $(\mathcal{N}, \kappa^x)$ as the *x–parametric flow network*.

As an intuitive metaphor, we may think of the demand nodes in our flow networks as representing queues of patients of each blood type arriving from the source; flows across the network correspond to directing patients in those queues to specific plasma supply nodes for treatment. Our x-parametric networks in some sense represent gating processes that limit the fraction of patients of each blood type that are allowed to proceed through the network to x. Our pooling algorithm corresponds to slowly opening gates, while keeping track of how much supply has been allocated at each step.

Given any $x \in (0,1]^{|\mathcal{D}\setminus\mathcal{D}_0|}$, we have the following feasibility constraints by construction:

Lemma 1. *For any flow function $\varphi \in \Phi(\mathcal{N}, \kappa^x)$, if $\varphi(X^D, Y^S) > 0$ for some $X^D \in \mathcal{D} \setminus \mathcal{D}_0$ and $Y^S \in \mathcal{S} \setminus \mathcal{S}_0$, then*

- *blood-type Y plasma is compatible with blood-type X patients, and $\varphi(X^D, Y^S)$ units of such blood-type Y plasma is given to blood-type X patients,*
- *in total, $\varphi(E, X^D) \le x_X \cdot \delta_X \le \delta_X$ units of blood-type X patients receive plasma, and*
- *in total, $\varphi(Y^S, K) \le \sigma_Y$ units of blood-type plasma Y is allocated.*

Conversely, for any feasible service rate vector s, there is a flow function $\varphi^s \in \Phi(\mathcal{N}, \kappa^{\mathbf{1}})$ where $\mathbf{1} = (1, \ldots, 1)$ such that for all $X^D \in \mathcal{D} \setminus \mathcal{D}_0$, we have $\varphi^s(E, X^D) = s_X \delta_X$.

A.2 Proof of Proposition 6

Now, we consider the family of x-parametric networks and set $x_X = c$ for all $X^D \in \mathcal{D} \setminus \mathcal{D}_0$ and increase c continuously from 0 to 1. When c is close to 0, the value of the network's maximum flow is equal to the total capacity of edges from the source, $\{(E, X^D)\}_{X^D \in \mathcal{D}\setminus\mathcal{D}_0}$ because when $c \approx 0$, any such flow can be carried over the network as long as $\sigma_X > 0$ for all $X \in \mathcal{B}$. Hence, when $c \approx 0$, $\{E\}$ is a minimum cut. We increase x continuously until either

- a break-point occurs at some $c = c_1 < 1$ such that $\{E\}$ is no longer a minimum cut of $(\mathcal{N}, \kappa^{c_1 \cdot \mathbf{1}})$, or
- $c = c_1 = 1$ such that cut $\{E\}$ stays as a minimum cut of $(\mathcal{N}, \kappa^{c \cdot \mathbf{1}})$ until c reaches 1.[23]

If $c_1 = 1$, then all patients can be served and we have $\hat{s}^n_X = x_1 = 1$ for all $X \in \mathcal{B}$ by construction of the network.

Thus we suppose that $c_1 < 1$. Then all minimum cuts are strictly larger than cut $\{E\}$ at $c = c_1$, meaning that we will not be able to send a flow with value equal to the the sum of capacities of all edges from the source anymore if c exceeds c_1. Let $\mathcal{N}_1 \supsetneq \{E\}$ be a minimum cut of the c_1-parametric network; if there are multiple such minimum cuts, let $\mathcal{N}_1$ be the largest of them.[24]

Now, suppose that $X^D \in \mathcal{N}_1 \cap \mathcal{D}$. Then for all supply nodes $Y^S \in \mathcal{S} \setminus \mathcal{S}_0$ such that $\{X^D, Y^S\} \in \mathbf{C}$ we must have $Y^S \in \mathcal{N}_1$, as otherwise the edge (X^D, Y^S) with capacity $\kappa^{c_1 \cdot \mathbf{1}}(X^D, Y^S) = +\infty$ would make the total capacity of the minimum cut of the whole network equal to $+\infty$.[25] (See Figure 3 for an example of a possible minimum cut at some c_1.)

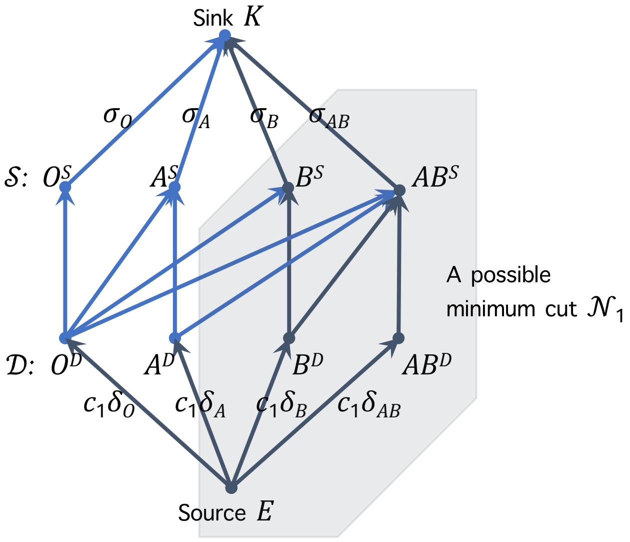

Figure 3: Example of a possible minimum cut $\mathcal{N}_1$ at some $c = c_1$ (assuming $\mathcal{D}_0 = \mathcal{S}_0 = \emptyset$) with $\mathcal{N}_1 \cap \mathcal{D} = \mathcal{D}^*_1 = \{B^D, AB^D\}$. Hence, $\mathcal{N}_1 \cap \mathcal{S} = \mathcal{S}^*_1 = \mathcal{C}_{\mathcal{D}^*_1}(\mathcal{S}) = \{B^S, AB^S\}$. The edges from $\mathcal{N}_1$ to $\mathcal{N} \setminus \mathcal{N}_1$ are denoted by thicker lines exiting the set $\mathcal{N}_1$. This cut's total capacity is $\kappa^{c_1 \cdot \mathbf{1}}(\mathcal{N}_1) = c_1\delta_O + c_1\delta_A + \sigma_B + \sigma_{AB}$.

Therefore, total capacity of $\mathcal{N}_1$ satisfies

$$\begin{aligned}\kappa^{c_1 \cdot \mathbf{1}}(\mathcal{N}_1) &= \sum_{X^D \in \mathcal{D} \setminus (\mathcal{D}^*_1 \cup \mathcal{D}_0)} \kappa^{c_1 \cdot \mathbf{1}}(E, X^D) + \sum_{Y^S \in \mathcal{S}^*_1} \kappa^{c_1 \cdot \mathbf{1}}(Y^S, K) \\ &= c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus (\mathcal{D}^*_1 \cup \mathcal{D}_0)} \delta_X \right) + \left(\sum_{Y^S \in \mathcal{S}^*_1} \sigma_Y \right), \qquad (30)\end{aligned}$$

[23] Recall that $\mathbf{1} = (1, \ldots, 1)$ is the $|\mathcal{D} \setminus \mathcal{D}_0|$ dimensional unit vector.

[24] Since the network is finite, it is straightforward to see that at least one minimum cut must exist.

[25] That would contradict $\mathcal{N}_1$ being a minimum cut of the network, as the cut $\{E\}$ of the network has always a finite total capacity.

where

$$\mathcal{D}_1^* := \mathcal{N}_1 \cap \mathcal{D} \qquad \text{and} \qquad \mathcal{S}_1^* := \mathcal{N}_1 \cap \mathcal{S} = \mathcal{C}_{\mathcal{D}_1^*}(\mathcal{S} \setminus \mathcal{S}_0). \tag{31}$$

As $\{E\}$ is a minimum cut of $(\mathcal{N}, \kappa^{c \cdot \mathbf{1}})$ for all $c < c_1$, by the Maximum Flow–Minimum Cut Theorem, a maximum flow function $\varphi^c \in \Phi(\mathcal{N}, \kappa^{c \cdot \mathbf{1}})$ satisfies

$$\varphi^c(\{E\}) = \sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \varphi^c(E, X^D) = \kappa^{c \cdot \mathbf{1}}(\{E\}) = \sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \kappa^{c \cdot \mathbf{1}}(E, X^D) = c \cdot \left(\sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \delta_X \right). \tag{32}$$

The value of the maximum flow of $(\mathcal{N}, \kappa^{c \cdot \mathbf{1}})$ is continuous in c at c_1.[26] Since $\mathcal{N}_1$ is a minimum cut of $(\mathcal{N}, \kappa^{c_1 \cdot \mathbf{1}})$, by the Maximum Flow–Minimum Cut Theorem and (32), we have

$$\kappa^{c_1 \cdot \mathbf{1}}(\mathcal{N}_1) = \varphi^{c_1}(\mathcal{N}_1) = \lim_{c \to c_1^-} \varphi^c(\{E\}) = c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \delta_X \right). \tag{33}$$

From (30) and (33), we then have

$$c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus (\mathcal{D}_1^* \cup \mathcal{D}_0)} \delta_X \right) + \sum_{Y^S \in \mathcal{S}_1^*} \sigma_Y = c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \delta_X \right),$$

leading together with (31) to

$$c_1 = \frac{\sum_{Y^S \in \mathcal{S}_1^*} \sigma_Y}{\sum_{X^D \in \mathcal{D}_1^*} \delta_X} = s_{\mathcal{D}_1^*}(\mathcal{S} \setminus \mathcal{S}_0) = s_{\mathcal{D}_1^*}(\mathcal{S}_1^*), \tag{34}$$

where s is as defined in (24).

The following claim shows that we have constructed precisely the sets $\mathcal{D}_1$ and $\mathcal{S}_1$ from our plasma pooling procedure.

Claim 1. *We have $\mathcal{D}_1^* = \mathcal{D}_1$ and $\mathcal{S}_1^* = \mathcal{S}_1$, where $\mathcal{D}_1$ and $\mathcal{S}_1$ are defined in (27) and (28), respectively (for $\ell = 1$).*

Proof. For any subset $\mathcal{D}' \subseteq \mathcal{D} \setminus \mathcal{D}_0$, we define cut $\mathcal{V} = \mathcal{D}' \cup \mathcal{C}_{\mathcal{D}'}(\mathcal{S} \setminus \mathcal{S}_0) \cup \{E\}$. Since $\mathcal{N}_1$ is a minimum

[26] Clearly, the value of the maximum flow is increasing in c. To see that we must have continuity in c, suppose for the sake of seeking a contradiction that the value ϕ of the maximum flow at c_1 is strictly greater than $\lim_{c \to c_1^-} \varphi^c(\{E\})$, and pick some $\epsilon > 0$ with

$$\epsilon < \frac{\phi - \lim_{c \to c_1^-} \varphi^c(\{E\})}{|\mathbf{E}|}.$$

For some c' sufficiently close to c_1, the network $(\mathcal{N}, \kappa^{c' \cdot \mathbf{1}})$ must be able to support the maximum flow at c_1 with all positive flow components decreased by ϵ because $(\mathcal{N}, \kappa^{c_1 \cdot \mathbf{1}})$ supports that flow and $\kappa^{c' \cdot \mathbf{1}} \geq \kappa^{c_1 \cdot \mathbf{1}} - (c_1 - c') \cdot \mathbf{1}$. But

$$\phi - |\mathbf{E}|\epsilon > \lim_{c \to c_1^-} \varphi^c(\{E\}) \geq \varphi^{c'}(\{E\}),$$

contradicting the assumption that $\varphi^{c'}$ is a maximum flow for $(\mathcal{N}, \kappa^{c' \cdot \mathbf{1}})$.

cut of $(\mathcal{N}, \kappa^{c_1 \cdot \mathbf{1}})$, using (33), we obtain

$$\kappa^{c_1 \cdot \mathbf{1}}(\mathcal{V}) = c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus (\mathcal{D}' \cup \mathcal{D}_0)} \delta_X \right) + \left(\sum_{Y^S \in \mathcal{C}_{\mathcal{D}'}(\mathcal{S} \setminus \mathcal{S}_0)} \sigma_Y \right) \geq \kappa^{c_1 \cdot \mathbf{1}}(\mathcal{N}_1) = c_1 \cdot \left(\sum_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \delta_X \right).$$

and hence,

$$s_{\mathcal{D}'}(\mathcal{S} \setminus \mathcal{S}_0) = \frac{\sum_{Y^S \in \mathcal{C}_{\mathcal{D}'}(\mathcal{S} \setminus \mathcal{S}_0)} \sigma_Y}{\sum_{X^D \in \mathcal{D}'} \delta_X} \geq c_1 = s_{\mathcal{D}_1^*}(\mathcal{S} \setminus \mathcal{S}_0),$$

which proves that $\mathcal{D}_1^* = \mathcal{D}_1$ and $\mathcal{S}_1^* = \mathcal{S}_1$. □

We now fix the capacity of the edge from the source (E, X^D) for each $X^D \in \mathcal{D}_1$ at $x_X = c_1$. We continue to increase the coefficients of other edges from the source (E, X^D) for each $X^D \in \mathcal{D} \setminus (\mathcal{D}_1 \cup \mathcal{D}_0)$ at the same speed $x_X = c$ above $c = c_1$. That is, we allocate all steady-state plasma associated with the supply nodes in $\mathcal{S}_1$ to patients associated with the demand nodes in $\mathcal{D}_1$, and continue to increase the capacities of all other edges from the source (i.e., the ones that do not end at nodes in $\mathcal{D}_1$), so that all new (largest) minimum cuts will include $\mathcal{N}_1$. That is: for any maximum flow function $\varphi \in \Phi(\mathcal{N}, \kappa^x)$, where $x_X = c_1$ for all $X^D \in \mathcal{D}_1$ and $x_X = c > c_1$ for all $X^D \notin \mathcal{D}_1 \cup \mathcal{D}_0$, whenever $X^D \in \mathcal{D}_1$ and $Y^S \in \mathcal{S}_1$,

- $\varphi(X^D, Y^S)$ is the amount of blood-type Y plasma allocated to blood-type X patients at steady-state and
- $\varphi(E, X^D) = c_1 \cdot \delta_X$ is the steady-state net flow of blood-type X patients served.

Then by iterating the argument just described, we determine a sequence of minimum cuts $\mathcal{N}_1, \ldots, \mathcal{N}_{\bar{\ell}}$ of networks $(\mathcal{N}, \kappa^{x^1}), \ldots, (\mathcal{N}, \kappa^{x^{\bar{\ell}}})$, respectively such that for all $\ell = 1, \ldots, \bar{\ell}$,

$$x^\ell = (x_X^\ell)_{X^D \in \mathcal{D} \setminus \mathcal{D}_0} \qquad \text{with} \quad x_X^\ell = \begin{cases} c_m & \text{if } X^D \in \mathcal{D}_m^* \text{ for some } m \leq \ell \\ c_\ell & \text{otherwise.} \end{cases}$$

Then

$$\mathcal{D}_\ell^* := (\mathcal{D} \setminus \cup_{m=1}^{\ell-1} \mathcal{D}_m^*) \cap \mathcal{N}_\ell \qquad \text{and} \qquad \mathcal{S}_\ell^* := (\mathcal{S} \setminus \cup_{m=0}^{\ell-1} \mathcal{S}_m^*) \cap \mathcal{N}_\ell = \mathcal{C}_{\mathcal{D}_\ell^*}\left(\mathcal{S} \setminus \cup_{m=0}^{\ell-1} \mathcal{S}_m^*\right) = \mathcal{C}_{\mathcal{D}_\ell^*}(\mathcal{S}_\ell^*),$$

with breakpoints $c_1 \leq \ldots \leq c_\ell \leq \ldots \leq c_{\bar{\ell}} \leq 1$ such that

$$c_\ell = \frac{\sum_{Y^S \in \mathcal{S}_\ell^*} \sigma_Y}{\sum_{X^D \in \mathcal{D}_\ell^*} \delta_X} = s_{\mathcal{D}_\ell^*}(\mathcal{S}_\ell^*). \tag{35}$$

Moreover, $\mathcal{D}_\ell^* = \mathcal{D}_\ell$ and $\mathcal{S}_\ell^* = \mathcal{S}_\ell$ for each ℓ as defined in (27) and (28).

Because the final flow obtained in this way is feasible, we see that we can feasibly serve plasma of blood types corresponding to nodes in $\mathcal{S}_\ell$ to blood-type X patients at a rate

$$\varphi^{c_\ell}(E, X^D) = c_\ell \cdot \delta_X = s_{\mathcal{D}_\ell}(\mathcal{S}_\ell)\delta_x,$$

where $\varphi^{c_\ell} \in \Phi(\mathcal{N}, \kappa^{x^\ell})$ is a maximum flow function. Hence, the service rate $\hat{s}^n_X = s_{\mathcal{D}_\ell}(\mathcal{S}_\ell)$ is feasible to achieve, completing the proof of Proposition 6.

A.3 Proof of Proposition 7

Now, let s' be a service rate vector that maximizes plasma distribution and let $\hat{s}^n$ be the service rate vector of the plasma pooling procedure. Observe that $\hat{s}^n$ maximizes the possible plasma distribution for demand nodes in $\mathcal{D}_0$, as it serves them with maximum rate of 1; thus, this also must the case for s'. If $\mathcal{D}_0 = \mathcal{D}$ then we are done, as all service rate vectors that maximize total plasma allocation are identical and serve all patients at steady-state.

So suppose $\mathcal{D}_0 \neq \mathcal{D}$. We prove the two parts of the proposition separately.

First part: By the construction of $\mathcal{D}^*_1$ and $\mathcal{S}^*_1$ in (31) in the proof of Proposition 6, the maximum possible flow for the smallest service rate is achieved at $c = c_1$. Since $\mathcal{D}^*_1 = \mathcal{D}_1$ and $\mathcal{S}^*_1 = \mathcal{S}_1$, we see that $\hat{s}^n$ maximizes the smallest service rate among all service rate vectors, including s'.

Second part: We first show that the largest service rate under vector $\hat{s}^n$ is weakly smaller than the largest service rate under vector s'.

To the contrary, suppose $\max_{X\in\mathcal{B}}\{s'_X\} < \max_{X\in\mathcal{B}}\{\hat{s}^n_X\}$, and let $Y \in \arg\max_{X\in\mathcal{B}}\{\hat{s}^n_X\}$.

Since s' maximizes the amount of steady-state plasma distributed, we have $s'_X = 1$ for all $X^D \in \mathcal{D}_0$, and (by Lemma 1) $(s'_X)_{X\in\mathcal{D}\setminus\mathcal{D}_0}$ is represented by a maximum flow function $\varphi^{s'} \in \Phi(\mathcal{N}, \kappa^{\mathbf{1}})$ such that $\varphi^{s'}(E, X^D) = s'_X\delta_X$ for all $X^D \in \mathcal{D} \setminus \mathcal{D}_0$.

As blood type Y has the largest service rate under $\hat{s}^n$, either $Y^D \in \mathcal{D}_0$ or $Y^D \in \mathcal{D}^*_{\bar{\ell}}$ (recall that $\mathcal{D}^*_\ell$ and $\mathcal{S}^*_\ell$ are iteratively defined in (34) for $\ell > 1$). If $Y^D \in \mathcal{D}_0$, then $\hat{s}^n_Y = 1$. Since s' maximizes plasma allocation, we should also have $s'_Y = 1$, leading to $1 = \max_{X\in\mathcal{B}}\{s'_X\} < \max_{X\in\mathcal{B}}\{\hat{s}^n_X\} = 1$—a contradiction.

Therefore, we must have $Y^D \in \mathcal{D}^*_{\bar{\ell}}$, and $\hat{s}^n_Y = c_{\bar{\ell}}$ as defined in (35). Since $s'_Y \leq \max_{X\in\mathcal{B}} s'_X < \hat{s}^n_Y$, some of the plasma of some blood type Z that is allocated to blood-type Y patients under $\hat{s}^n$ is being allocated to some other blood-type X patients under s'. Since blood-type Z plasma is given to blood-type Y patients under $\hat{s}^n$, Z^S must be available in round $\bar{\ell}$, i.e.,

$$Z^S \in \mathcal{S}^*_{\bar{\ell}}. \tag{36}$$

Moreover, in the pooling procedure, at least one such demand node X^D has to be included in $\mathcal{D}^*_\ell$ for some $\ell < \bar{\ell}$—as otherwise, s'_Y would be equal to $\hat{s}^n_Y$.

As we showed in the proof of Proposition 6, the plasma pooling procedure exclusively allocates all possible plasma associated with supply nodes in $\mathcal{C}_{\cup_{m=0}^{\ell}\mathcal{D}^*_\ell}(\mathcal{S}) = \cup_{m=0}^{\ell}\mathcal{S}^*_m$ to patients associated with demand nodes in $\cup_{m=0}^{\ell}\mathcal{D}^*_\ell$. Therefore, we must have $Z^S \in \cup_{m=0}^{\ell}\mathcal{S}^*_m$, contradicting (36) since $\bar{\ell} > \ell$.

Thus, we see that

$$\max_{X'\in\mathcal{B}}\{s'_{X'}\} \geq \max_{X'\in\mathcal{B}}\{\hat{s}^n_{X'}\}$$

which, in turn, together with the first part of the proposition implies that

$$\max_{X'\in\mathcal{B}}\{s'_{X'}\} - \min_{X'\in\mathcal{B}}\{s'_{X'}\} \geq \max_{X'\in\mathcal{B}}\{\hat{s}^{n}_{X'}\} - \min_{X'\in\mathcal{B}}\{\hat{s}^{n}_{X'}\}.$$

References

Abdulkadiroğlu, Atila, Yeon-Koo Che, and Yosuke Yasuda (2011). "Resolving Conflicting Preferences in School Choice: The "Boston Mechanism" Reconsidered." *American Economic Review*, 101(1), 399–410.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan (2020). "Thickness and Information in Dynamic Matching Markets." *Journal of Political Economy*, 128(3), 783–815.

Akbarpour, Mohammad, Julien Combe, Yinghua He, Victor Hiller, Robert Shimer, and Olivier Tercieux (2019). "Unpaired Kidney Exchange: Overcoming Double Coincidence of Wants without Money." August 1, Stanford GSB Working Paper.

Anderson, Ross, Itai Ashlagi, David Garmanik, and Yash Kanoria (2017). "Efficient Dynamic Barter." *Operations Research*, 65(6), 1446–1459.

Azevedo, Eduardo and Eric Budish (2019). "Strategy-proofness in the Large." *Review of Economic Studies*, 86(1), 81–116.

Azevedo, Eduardo and Jacob Leshno (2016). "A Supply and Demand Framework for Two-Sided Matching Markets." *Journal of Political Economy*, 124(5), 1235–1268.

Azevedo, Eduardo M. and John William Hatfield (2018). "Existence of equilibrium in large matching markets with complementarities." University of Pennsylvania Working Paper.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv (2018). "Optimal Dynamic Matching." CEPR Discussion Paper 12986.

Cape Fear Valley (2020). "Blood Donor Center." Available at: http://www.capefearvalley.com/blood/index.html. Last Accessed: August 3, 2020.

Che, Yeon-Koo, Jinwoo Kim, and Fuhito Kojima (2019). "Stable Matching in Large Economies." *Econometrica*, 87, 65–110.

Che, Yeon-Koo and Fuhito Kojima (2010). "Asymptotic Equivalence of Probabilistic Serial and Random Priority Mechanisms." *Econometrica*, 78(5), 1625–1672.

Che, Yeon-Koo and Olivier Tercieux (2019). "Efficiency and Stability in Large Matching Markets." *Journal of Political Economy*, 127, 2301–2342.

Cowan, N., H. A. Gritsch, N. Nassiri, J. Sinacore, and J. Veale (2017). "Broken Chains and Reneging: A Review of 1748 Kidney Paired Donation Transplants." *American Journal of Transplantation*, 1–7.

Delamou, Alexandre, Nyankoye Yves Haba, Almudena Mari-Saez, Pierre Gallian, Maya Ronse, Jan Jacobs, Bienvenu Salim Camara, Kadio Jean-Jacques Olivier Kadio, Achille Guemou, Jean Pe Kolie Maaike De Crop, Patricia Chavarrin, Chantatl Jacquot, Catherine Lazaygues, Anja De Weggheleire, Lutgarde Lynen, and Johan van Griensven (2016). "Organizing the Donation of Convalescent Plasma for a Therapeutic Clinical Trial on Ebola Virus Disease: The Experience in Guinea." *American Journal of Tropical Medical Hygiene*, 95(3), 647–653.

EBA (2020). "Covid-19 and Blood Establishments." April 25, Available at: https://europeanbloodalliance.eu/covid19-and-blood-establishment/. Last accessed: August 3, 2020.

Ford, L. R. and D. R. Fulkerson (1956). "Maximal flow through a network." *Canadian Journal of Mathematics*, 8, 399–404.

Heger, Stephanie A., Robert Slonim, Ellen Garbarino, Carmen Wang, and Daniel Waller (forthcoming). "Redesigning the Market for Volunteers: A Donor Registry." *Management Science*.

Joyner, Michael J et al. (2020). "Evidence favouring the efficacy of convalescent plasma for COVID-19 therapy." *medRxiv*.

Kojima, Fuhito and Parag A. Pathak (2009). "Incentives and Stability in Large Two-Sided Matching Markets." *American Economic Review*, 99, 608–627.

Korte, Bernhard and Jens Vygen (2012). *Combinatorial Optimization: Theory and Algorithms*. Springer.

Lacetera, Nicola, Mario Macis, and Robert Slonim (2013). "Economic Rewards to Motivate Blood Donations." *Science*, 340, 927–928.

Lavee, Jacob, Tamar Ashkenazi, Gabriel Gurman, and David Steinberg (2010). "A new law for allocation of donor organs in Israel." Lancet, 375: 1131-33.

Luke, Thomas C., Edward M. Kilbane, Jeffrey L. Jackson, and Stephen L. Hoffman (2006). "Meta-Analysis: Convalescent Blood Product for Spanish Influenza Pneumonia: A Future H1N1 Treatment?" *Annals of Internal Medicine*, 145, 599–609.

OneBlood (2020). "COVID-19 Convalescent Plasma: Donor FAQs." Available at: https://www.oneblood.org/lp/covid-19-convalescent-plasma.stml. Last accessed: August 3, 2020.

Pope, Adam (2018). "Nation's longest single-site kidney chain reaches 100." *UAB News*. July 30, Available at: https://www.uab.edu/news/health/item/9638-nation-s-longest-single-site-kidney-chain-reaches-100. Last accessed: August 3, 2020.

Rees, Michael A., Jonathan E. Kopke, Ronald P. Pelletier, Dorry L. Segev, Matthew E. Rutter, Alfredo J. Fabrega, Jeffrey Rogers, Oleh G. Pankewycz, Janet Hiller, Alvin E. Roth, Tuomas Sandholm, M. Utku "Unver, and Robert A. Montgomery (2009). "A Nonsimultaneous, Extended, Altruistic-Donor Chain." *New England Journal of Medicine*, 360 (11), 1096–1101.

Roos, Dave (2020). "Before Vaccines, Doctors 'Borrowed' Antibodies from Recovered Patients to Save Lives." *History Stories*. April 1, Available at: https://www.history.com/news/blood-plasma-covid-19-measles-spanish-flu, Last accessed: August 3, 2020.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver (2004). "Kidney Exchange." *Quarterly Journal of Economics*, 119, 457–488.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver (2005a). "A Kidney Exchange Clearinghouse in New England." *American Economic Review: Papers and Proceedings*, 95(2), 376–380.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver (2005b). "Pairwise Kidney Exchange." *Journal of Economic Theory*, 125, 151–188.

Roth, Alvin E., Tayfun Sönmez, M. Utku Ünver, Francis L. Delmonico, and Susan L. Saidman (2006). "Utilizing List Exchange and Undirected Good Samaritan Donation through "Chain" Paired Kidney Exchange." *American Journal of Transplantation*, 6(11), 2694–2705.

Rubin, Rita (2020). "Testing an Old Therapy Against a New Disease: Convalescent Plasma for Covid-19." *Journal of the American Medical Association*. April 30.

Singapore (2012). "Human Organ Transplant Act of 1987, Revised 2012." Available at: https://sso.agc.gov.sg/Act/HOTA1987. Last Accessed: August 3, 2020.

Slonim, Robert, Carmen Wang, and Ellen Garbarino (2014). "The Market for Blood." *Journal of Economic Perspectives*, 28(2), 177–196.

Sönmez, Tayfun and M. Utku Ünver (2015). "Enhancing the Efficiency of and Equity in Transplant Organ Allocation via Incentivized Exchange." Available as Boston College Working Paper 868.

Sönmez, Tayfun, M. Utku Ünver, and M. Bumin Yenmez (2020). "Incentivized Kidney Exchange." *American Economic Review*, 110 (7), 2198–2224.

Stack, Liam (2020). "Hasidic Jews, Hit Hard by the Outbreak, Flock to Donate Plasma." NY Times, May 12.

Starr, Douglas (2002). *Blood: An Epic History of Medicine and Commerce*. Harper Perennial.

Ünver, Utku (2010). "Dynamic Kidney Exchange." *Review of Economic Studies*, 77(1), 372–414.

Veale, Jeffrey L., Alexander M. Capron, Nima Nassiri, Gabriel Danovitch, H. Albin Gritisch, Amy Waterman, Joseph Del Pizzo, Jim C. Hu, Marek Pycia, Suzanne McGuire, Marian Charlton, and Sandip Kapur (2017). "Vouchers for Future Kidney Transplants to Overcome "Chronological Incompatibility" Between Living Donors and Recipients." *Transplantation*, 101(9).

White Plains Hospital (2020). "Convalescent Plasma Program." Available at: https://www.wphospital.org/patients-and-visitors/coronavirus/plasma-donation. Last accessed: August 3, 2020.

Zúniga-Fajuri, Alejandra (2015). "Increasing organ donation by presumed consent and allocation priority: Chile." Bulletin World Health Organization, 93:199-202.