When do pharmaceutical patent holders have an incentive to delay the launch of follow-on drugs?*

Kurt R. Brekke[†] Dag Morten Dalen[‡] Odd Rune Straume[§]

January 2025

Abstract

Pharmaceutical companies often delay the introduction of new versions of a drug, allegedly extending exclusivity periods of the original innovation. This practice is viewed as a way to circumvent market cannibalisation, since introducing a new drug early would result in decreased sales of the original drug. However, our study reveals that market cannibalisation as such does not provide sufficient incentives for delaying the introduction of follow-on drugs. We find that, under free pricing, launching a follow-on drug while the original version still enjoys patent protection is optimal, since it also allows for increases in the price of the original drug. Pricing and cost-containment constraints imposed by health plans and regulators, rather than the patent system itself, are shown to incentivise delays of valuable follow-on drugs.

Keywords: Pharmaceutical markets; Follow-on drugs; Evergreening; Therapeutic competition;

JEL Classification: I11; I18; L13; L65.

*Straume acknowledges financial support from National Funds of the FCT – Portuguese Foundation for Science and Technology within the project UIDB/03182/2020.

[†]Norwegian School of Economics (NHH), Department of Economics, Helleveien 30, 5045 Bergen, Norway. E-mail: kurt.brekke@nhh.no

[‡]Corresponding author. BI Norwegian Business School, NO-0442 Oslo, Norway. E-mail: dag.m.dalen@bi.no

[§]Department of Economics/NIPE, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. E-mail: o.r.straume@eeg.uminho.pt

1 Introduction

Patents play a pivotal role in rewarding the development of new drugs. By granting exclusive rights to the inventor, patents offer incentives for pharmaceutical companies to invest in costly research and development processes to create novel drugs that address unmet medical needs and improve patients' health. Pharmaceutical companies need to protect the intellectual property rights to promising new compounds long before they are approved and get marketing authorisation. On average, it takes around 8 years of clinical testing before a new drug is approved (Lakdawalla, 2018). In the early development phase, specific choices need to be made regarding the exact molecule structure, dosage and formulation, and in some cases also combinations with other active ingredients, to be tested in clinical trials (Fowler, 2017). After approval, both clinical experience and medical development can reveal possible adjustments of the original drug that improve treatment outcomes, at least for some groups of patients.

If new versions of a drug offer improved treatment outcomes or better adherence, they should be introduced in the market as soon as they have been recognised. Instead, it appears that follow-on drugs in many cases tend to be introduced shortly before patent protection of original versions expires (Huckfeldt and Knittel, 2011; Shapiro, 2016; Fowler, 2019). This is seen as evidence of life cycle management practices that strategically delay new versions to extend exclusivity periods beyond what is justified by the original innovation (Price II, 2020). The argument is that early introduction of a follow-on drug would cannibalise the sales of the original version (Shapiro, 2016; Fowler, 2019). By delaying introduction, the new version will instead be competing with generic versions of the original drug, but then with some degree of market power since follow-on drugs often benefit from fixed exclusivity periods separate from the original drug.

Despite the existence of some empirical support for the strategic delay argument, economic research has not yet provided a sound theoretical explanation of the phenomenon. Our paper develops a theoretical model to investigate pharmaceutical companies' incentives to delay follow-on drugs until patent expiration of the original drug. We adopt a spatial framework in which a patented original drug constitutes the only treatment option until a follow-on drug is discovered. Once discovered, the producer can start the approval process for immediate launch or wait until patent expiration of the original drug. We analyse this choice within a fairly flexible framework in which the follow-on drug can be both horizontally and vertically differentiated from the

original version.

If the follow-on drug is introduced before patent expiration of the original drug, some patients will benefit from switching to the new version. If the market is fully covered, this will cause an equal loss of sales for the original version. Our analysis reveals that even such an extreme form of market cannibalisation is not sufficient to delay the introduction of follow-on drugs. Lower sales of the original drug are more than compensated for by the fact that the introduction of the follow-on drug allows the pharmaceutical company to adjust the price of the original drug upwards. The profitability of an early launch, therefore, stems not only from the additional surplus extracted through the sales of the follow-on drug, but also from being able to extract more surplus from selling the original drug. If the company delays the launch until patent expiration, this latter effect is absent since the price of the original drug will be forced down to marginal cost due to generic competition. Thus, it is more profitable to launch the follow-on drug when the company still has market power in the price setting of the original drug. In an extension to our main analysis we show that this holds even if we introduce patients that are loyal to the company's original brand when faced with generic versions.

However, this does not mean that delaying the introduction of follow-on drugs cannot be a viable strategy. Instead, we identify under which circumstances such strategies might be chosen by the producer. It turns out that pricing and cost-containment constraints imposed by regulators and health plans, more than the patent system as such, can provide incentives to delay valuable follow-on drugs. This result has support from Kyle (2007) and Cockburn et al. (2016) who find that pharmaceutical companies strategically delay launches in countries with price controls, and that the use of price controls has a statistically and quantitatively important effect on the extent and the timing of the launch of new drugs. In many real-world pharmaceutical markets, companies face various constraints on their price setting, which distorts the crucial link between the launch of follow-on drugs and the price of patented original drugs. We explore two such pricing restrictions.

We first assume that the pharmaceutical company is free to set the prices of its different drug versions, but that each version will be potentially available to patients only if they are accepted by a health plan that maximises total health benefits net of drug purchasing costs. This means that a new drug is included in the health plan only if the additional health gains of the drug are at least as high as the increase in total purchasing costs for the health plan. Although this restricts the pharmaceutical company from fully appropriating the therapeutic

value of follow-on drugs, we show that the cannibalisation effect may still be too weak to delay entry until patent expiration of the original version.

Assuming that the follow-on drug and the original version are only horizontally differentiated, we need to impose a more severe regulatory constraint in the form of a binding price cap in order to create incentives to delay the launch of the follow-on drug. The effect of a binding price cap on the incentives for launch delay is most easily understood in the special case in which the treatment value of the original drug is sufficiently high, such that the market is fully covered even before the follow-on drug is launched. In this case, since the introduction of the follow-on drug does not increase total demand, and since a binding price cap does not allow the company to increase drug prices, there is no additional profit to be gained from early introduction and it is clearly optimal to delay the launch of the follow-on drug until the patent of the original drug expires in order to avoid the cannibalisation effect.

Somewhat surprisingly, therefore, it appears that patent protection and market power in itself is not likely to cause inefficiencies in the timing of follow-on drug launches. Our results reveal instead that attempts to control market power by price cap regulation is what might trigger a delay strategy for follow-on drugs. Discussions of how price regulation affects welfare are usually related to the well-known trade-off between dynamic and static efficiency, where the typical concern is that (excessively strict) price regulation might harm dynamic efficiency by providing insufficient incentives for developing new drugs. We identify another potential inefficiency caused by price regulation; namely delayed launch of drugs that have already been developed.

The rest of the paper is organised as follows. In Section 2 we discuss related literature. In Section 3 we present the model. In Section 4 we analyse the optimal launch time decision under free drug pricing. In Section 5 we investigate the optimal launch time under constrained drug pricing. In Section 6 we explore efficiency and welfare properties of the launch decisions, including the effect of price regulation. In Section 7 we explore an extension with brand-loyal patients, before drawing some policy implications and offering some concluding remarks in Section 8.

2 Related literature

This paper adds to the literature in health economics that study inter-brand competition in pharmaceutical markets. Although novel drugs enjoy patent-protection for an extensive period, their market power is often constrained by the entry of therapeutic substitutes. These are drugs with different active substances that target the same groups of patients and indications. When they are sold by different pharmaceutical companies, they can introduce price competition, depending on their therapeutic overlap (e.g., Ellison et al., 1997; Branstetter et al., 2016, Danzon and Chao, 2020; Dickson et al., 2023). Theoretical studies have explored therapeutic competition and price regulations by adopting spatial frameworks that incorporate both horizontal and vertical differentiation among drugs (e.g., Miraldo, 2009; Bala and Bhardwaj, 2010; Bardey et al., 2010; Bardey et al., 2016; González et al., 2016; Brekke et al., 2022, 2023, 2024). Our model is closely related to this strand of the literature, but distinguishes itself by assuming that therapeutic substitutes are introduced by the same company.

When new drugs and formulations are produced by the same company, a new set of concerns arises, often associated with so-called product-hopping (Carrier, 2010; Shadowen et al., 2009, 2016; Carlton et al., 2016). Carrier and Shadowen (2016) reserve the term 'product-hopping' for instances in which the brand producer (i) reformulates the product with the purpose of making generic drugs non-substitutable to the original drug and (ii) encourages doctors to write prescriptions for the reformulated product rather than the original. According to this reasoning, the producer introduces a new version of the drug shortly before patent expiration that does not provide any significant additional therapeutic benefits over the original drug.[1] Despite this, producers are claimed to be able to persuade health providers to prescribe the new version, and harm consumers by making them pay higher prices than they would have paid for the generic versions of the older alternative.

Carlton et al. (2016) criticise the underlying assumption that generic producers cannot convince potential buyers to buy generic substitutes instead of the new product, if the new product is in fact not offering any therapeutic improvements. Since new drug versions are granted market exclusivity following successful completion of clinical trials, Shapiro (2016) argues that it is not unreasonable to assume that a new version of a drug might be therapeutically superior,

[1] There is a related literature on patentholders' ability to delay generic competition by increasing the cost of challenging patents that protects the brand name drug (e.g., Engelberg et al., 2009; Hemphill and Sampat, 2011, 2012; Lipatov, 2020).

at least for some patients.[2] For example, Huskamp et al. (2009) found evidence of therapeutic benefits of new reformulations of SSRI anti-depressants for subgroups of users. In the present paper we adopt a flexible approach in which the therapeutic value-added of the follow-on drug might range from 'minimal' to 'substantial', for some or all patients. Thus, our modelling framework allows us to explore how the degree of substitutability between the original version and reformulations affects market outcomes.

If a new version of a drug offers improved treatment effects, the first-best outcome (from a utilitarian welfare perspective) is that the new version is introduced as soon as possible.[3] Using data for the sleep-aid drugs, Shapiro (2016) derives estimates of the welfare costs of strategically delaying the entry of a new version until just before patent expiration of the original drug. Yin (2023) extends this line of research by evaluating policies for exclusivity protection of line extensions. She uses individual-level prescription data for SSRI anti-depressant drugs and finds that the consumer surplus losses due to extended market exclusivity exceed the consumer surplus benefits from innovating line extensions. Phadke (2024) finds similar results when using data for dementia drugs. Furthermore, Fowler (2019) shows that the producer's incentive for delaying introduction increases with the share of line extension sales that would cannibalise sales of the original drug. Using a novel dataset of over 700 pharmaceuticals approved in the United States from 1985-2016, linked to all subsequent line extensions in that period, she finds that an original product is almost twice as likely to have a line extension approved in the period leading up to expected generic entry than three or more years before expected generic entry.

A related study to ours is Kyle (2007) who study the relationship between the use of price control schemes and the launch decision by pharmaceutical companies across countries and therapeutic markets. Using a novel dataset on all drugs developed between 1980 and 2000, she obtains information on drug launches in therapeutic markets for the 28 largest pharmaceutical markets globally.[4] Combining this with information about price control schemes, she finds that price controls has a significant and sizeable effect on the extent and timing of the launch of new drugs in therapeutic markets across countries. Similar findings are reported by Cockburn et al. (2016) who study the timing of launches of 642 new drugs in 76 countries during 1983

[2] In their definition of product-hopping, Carrier and Shadowen (2016) wanted to avoid targeting brand reformulations as anti-competitive if they are designed to improve the product by competing with other brands or growing the market.

[3] See Oi (2007) for a general discussion of why new inventions are not always immediately introduced in the market.

[4] Stremersch and Lemmens (2009) examine the relationship between various regulation and sales of new drug therapies.

to 2002. Using information about price control schemes in these countries, they find that price regulation delays launch, while longer and more extensive patent rights accelerates it. While these studies do not focus specifically on the launch decision of follow-on drugs, their findings align with our results in the present paper.

Finally, our paper is also related to the broader literature on innovation and strategic entry in pharmaceutical markets. For instance, Ellison and Ellison (2009) analyse incentives for an incumbent brand-name drug producer to deter entry from generics producers by means of advertising. Another example is Narasimhan and Zhang (2000) who develop a game-theoretical model to analyse firms' decisions about entering untested markets, revealing that both pioneering advantages and laggard disadvantages can drive firms to enter markets quickly. Additionally, they discuss how cannibalisation can either deter or motivate incumbents to become a market pioneer. None of these papers are concerned with the launch decision of follow-on drugs.

3 Model

Consider a recently patented prescription drug for which there is currently no viable therapeutic alternative. There is a unit mass of heterogeneous patients who might benefit, to different degrees, from being treated with this drug. More specifically, suppose that the therapeutic benefit of one drug unit is given by $v - \tau x$, where x is randomly and independently drawn from a uniform distribution on $[0, 1]$ for each patient, and where the parameter $\tau > 0$ therefore measures the degree of patient heterogeneity in therapeutic benefits. We will henceforth refer to this drug as the *original drug*.

As a by-product of the R&D investments undertaken to develop the original drug, the producer also has the possibility of introducing (at a much lower cost) a modified version of this drug that might give additional therapeutic benefit to some patients. More specifically, suppose that the therapeutic benefit of one unit of the new drug version is given by $w - \tau(1 - x)$, which implies that the therapeutic benefits of the two drugs are negatively correlated across the patient mass. Thus, the two drug versions are horizontally differentiated (as long as $\tau > 0$) and also potentially vertically differentiated (if $w \neq v$). If introduced, the new drug version will be protected by either a separate patent or a fixed exclusivity period, depending on the similarity with the original drug. We assume that both drug versions can be produced at marginal cost equal to c, and we will henceforth refer to the modified version of the original drug as the

follow-on drug. We will also intermittently refer to the parameters v and w as the *quality* of the original and the follow-on drug, respectively.[5]

In this context, a key decision the producer has to make is when to launch the follow-on drug. In order to analyse this problem, we consider a two-period model in which the length of each period corresponds to the length of the exclusivity period of the follow-on drug, and where the patent protection of the original drug ends after the first period.[6] In this setting, there are two potentially optimal launch decisions for a profit-maximising producer:

1. The follow-on drug is launched at the start of the first period, and the producer sells both drug versions, under patent protection, in the first period. In the second period, generic competition drives prices down to marginal cost for both drug versions and the producer earns zero profits in this period.

2. The launch of the follow-on drug is delayed until the start of the second period. In this case, the producer sells only the original drug, under patent protection, in the first period. In the second period, generic competition drives the price of the original drug down to marginal cost. However, the producer can still make profits from selling the follow-on drug, which is protected from direct generic competition in the second period, but faces competition from the (horizontally and vertically differentiated) original drug that is sold at price equal to marginal cost.

The analysis is based on the basic observation that profits from prescription drug sales are predominantly made during the patent or exclusivity period, which we take to the extreme by assuming that direct generic competition will imply drug prices equal to marginal cost. With this simplifying assumption, the launch decision boils down to whether it is more profitable for the producer to sell a patent-protected follow-on drug (i) in competition with its own patent-protected original drug (in the first period), or (ii) in competition with generic versions of the off-patent original drug (in the second period).

Regarding drug demand, we assume that all patients are covered by a health plan, and that each patient will be prescribed at most one unit of drug treatment in each period. The treatment choices are made by a prescribing physician who considers both treatment costs and

[5] Notice also that the degree of horizontal differentiation between the original drug and the follow-on drug is measured by the magnitude of τ *relative to* v and w. Thus, higher values of v and w are equivalent to a lower value of τ; in both cases, the two drugs are less horizontally differentiated.

[6] Thus, if the patent period of the original drug ends at time T and the length of the exclusivity period of the follow-on drug is t, the first period starts at time $T - t$, while the second period ends at time $T + t$.

therapeutic benefits for each patient, which we assume are observable to the physician. More specifically, we assume that the physician has three treatment choices for each patient, and that the utility assigned to each of these choices by the physician is given by

$$U = \begin{cases} v - \tau x - \beta p_o & \text{if the original drug is prescribed} \\ w - \tau (1 - x) - \beta p_f & \text{if the follow-on drug is prescribed} \\ 0 & \text{if no drug treatment is prescribed} \end{cases}, \tag{1}$$

where p_o and p_f are the unit prices of the original and follow-on drugs, respectively. The parameter $\beta \in (0, 1]$ measures the price sensitivity of the physician's treatment choice. We can think of the physician as being an agent for both the patient and for the health plan that purchases the drugs. In the special case of $\beta = 1$, the physician takes drug prices fully into account and acts as a perfect agent for a health plan that maximises total health benefits net of purchasing costs. However, in the more general case of $\beta < 1$, the physician is more concerned about treatment benefits than treatment costs.[7] For the sake of terminological brevity, we will henceforth refer to βc as the *perceived treatment cost* of drug treatment; i.e., the cost of one unit of drug treatment as 'perceived' by the prescribing physician. The total per-period demand for each drug in the market results then from maximisation of (1) for each patient.

Finally, our analysis relies on the underlying assumption that the cost of launching the follow-on drug is always low enough to make a launch profitable. Since these costs have to be paid regardless of whether the drug is launched in the first or in the second period, they do not affect the optimal launch time. For simplicity, we also abstract from any discounting of profits between the two periods. All else equal, the only effect of a positive discount rate is to make launch delay less profitable. Thus, the assumption of no discounting means that we maximise the scope for launch delay to be the profit-maximising decision.

4 Launch time decision under free drug pricing

We start out by assuming that the patent-holding producer can freely set the prices of both the original drug and the follow-on drug. Below we present the profit-maximising pricing decisions, and the corresponding profits, in each possible market structure, before deriving the optimal

[7] Notice that our assumption of drug prices equal to marginal cost after patent expiry arises as a Nash equilibrium outcome as long as β is strictly positive and as long as the prescribing physician considers generic copy drugs to be perfect substitutes to the off-patent drug (original or follow-on) for all patients.

launch time of the follow-on drug. In order to make the analysis as complete as possible, we do not impose any *ex ante* restrictions on the parameters of the model.

4.1 Equilibrium drug prices, demand and profits

Depending on the launch time decision, there are three different cases in which the producer is able to earn positive profits. In the first period, the producer can earn positive profits from selling either one or two drugs, depending on whether the follow-on drug is launched in this period or not. Furthermore, if the launch is delayed, the producer can also earn positive profits from selling the follow-on drug under exclusivity in the second period. We will consider each of these three cases in turn. Below we present and discuss the profit-maximising outcomes, in terms of prices and profits, while further details about the derivation of these outcomes are relegated to Appendix A.

4.1.1 First-period outcome with launch delay (market structure M)

Suppose that the launch of the follow-on drug is delayed until the second period. In this case, the producer is a single-product monopolist in the first period, which we henceforth refer to as market structure M. The profit-maximising price of the original drug in the first period, denoted by p_o^M, is then given by

$$p_o^M = \begin{cases} \frac{v+\beta c}{2\beta} & if \quad v \leq v^M \\ \frac{v-\tau}{\beta} & if \quad v > v^M \end{cases}, \tag{2}$$

where

$$v^M := 2\tau + \beta c. \tag{3}$$

When setting the profit-maximising price, the producer faces a basic trade-off between surplus extraction at the intensive versus the extensive margin. As long as the quality of the original drug is below a certain threshold level, given by v^M, this trade-off results in a market that is not fully covered. In other words, the optimal price is such that some patients, those with relatively low therapeutic benefit, are not prescribed the drug. Although the producer does not capture all potential demand, this is more than compensated by higher surplus extraction from inframarginal patients who derive a higher therapeutic benefit from drug treatment.

However, for a sufficiently high drug quality, $v > v^M$, the profit-maximising price is such

that all patients are prescribed the drug, resulting in a fully covered market. For $v > v^M$, the price of the original drug is set such that the prescribing physician is indifferent between prescribing or not the drug to the patient who derives the lowest therapeutic benefit from drug treatment (i.e., the patient with $x = 1$). Intuitively, the profit-maximising price is inversely related to the price-sensitivity (β) of the prescribing physician.

The resulting first-period profits, denoted by π^M, are given by

$$\pi^M = \begin{cases} \frac{(v-\beta c)^2}{4\beta\tau} & if \quad v \leq v^M \\ \frac{v-\tau}{\beta} - c & if \quad v > v^M \end{cases}, \tag{4}$$

and are monotonically increasing in the quality of the drug. Notice that this solution requires that $v > \beta c$, i.e., that the maximum therapeutic benefit of drug treatment is higher than the perceived treatment cost. Otherwise, if $v < \beta c$, no patient will be prescribed the drug even if the price is set at marginal cost, and the drug will therefore not survive in the market.

4.1.2 First-period outcome without launch delay (market structure MM)

Suppose that the producer instead launches the follow-on drug immediately, at the start of the first period. In this case, the producer is a multi-product monopolist in the first period, which we henceforth refer to as market structure MM. In this case, the profit-maximising prices for the original and the follow-on drug, denoted by p_o^{MM} and p_f^{MM}, respectively, are given by

$$p_o^{MM} = \begin{cases} \frac{v+\beta c}{2\beta} & if \quad v \leq v^{MM} \\ \frac{3v+w-2\tau}{4\beta} & if \quad v > v^{MM} \end{cases}, \tag{5}$$

$$p_f^{MM} = \begin{cases} \frac{w+\beta c}{2\beta} & if \quad v \leq v^{MM} \\ \frac{3w+v-2\tau}{4\beta} & if \quad v > v^{MM} \end{cases}, \tag{6}$$

where

$$v^{MM} := 2(\tau + \beta c) - w. \tag{7}$$

Also in this case, the market might be either partially or fully covered in equilibrium. If $v+w < 2(\tau + \beta c)$, which implies that $v < v^{MM}$, the 'aggregate drug quality' in the market is so low that some patients will not be prescribed either of the drugs at the profit-maximising prices. Since the therapeutic benefits of the two drugs are negatively correlated across the patient mass,

this implies that there is no *de facto* substitutability between the two drugs, which are then priced as if they were monopoly products in two separate markets.

However, if $v+w > 2(\tau + \beta c)$, which implies that $v > v^{MM}$, the drug qualities are sufficiently high for all patients to be prescribed either the original or the follow-on drug at the profit-maximising prices. In this case, the profit-maximising prices are set such that the maximum surplus is extracted from the market while still making sure that all patients are prescribed one of the drugs. In other words, the prices are such that there exists a patient for whom the physician is indifferent between all three treatment choices: prescribing the original drug, prescribing the follow-on drug, or not prescribing any drug treatment. The exact characteristics of this patient are determined by the relative quality of the two drug versions. More specifically, this patient is characterised by $x > (<) 1/2$ if $v > (<) w$, which implies that the higher-quality drug will be prescribed to the majority of the patients. It is also easily verified from (5)-(6) that the higher-quality drug has a higher price, i.e., that $p_o^{MM} > (<) p_f^{MM}$ is $v > (<) w$.

Furthermore, it is worth noticing that, when the market is fully covered, the price of each drug is increasing not only in its own quality, but also in the quality of the therapeutically substitutable drug. This occurs because the two drug versions are sold by the same producer. Suppose that the quality of the follow-on drug increases. All else equal, this shifts some demand towards the follow-on drug and also implies that the new marginal patient (for whom the physician is indifferent between prescribing the original and the follow-on drug) obtains a therapeutic benefit of drug treatment that is strictly higher than the perceived treatment cost. The optimal way for the producer to extract some of the additional surplus created by this quality increase is to increase the prices of *both* drugs, though with a larger price increase for the follow-on drug.

The resulting first-period profits, denoted by π^{MM}, are given by

$$\pi^{MM} = \begin{cases} \frac{v^2+w^2}{4\beta\tau} - \frac{(v+w-\beta c)c}{2\tau} & if \quad v \leq v^{MM} \\ \frac{w+v-\tau}{2\beta} + \frac{(w-v)^2}{8\beta\tau} - c & if \quad v > v^{MM} \end{cases}. \tag{8}$$

As expected, these profits are increasing in the quality of either drug version, and decreasing in the price sensitivity of the physician's prescription choices.

As before, positive drug sales require that the maximum therapeutic benefit is higher than the perceived treatment cost; i.e., $\min\{v, w\} > \beta c$. Notice, however, that the above presented solution also requires that the quality difference between the two drug versions is not too

large. More specifically, the prices given by (5)-(6) ensure positive sales of both drugs only if $|w - v| < 2\tau$. If this condition does not hold, only one of the two drug versions will survive in the market. There are therefore two remaining possible market outcomes, which can be characterised by defining two threshold levels of v, given by

$$\underline{v}^{MM} := w - 2\tau \tag{9}$$

and

$$\overline{v}^{MM} := w + 2\tau. \tag{10}$$

If $v < \underline{v}^{MM}$, the follow-on drug replaces the original drug for all patients and is priced such that the market is fully covered, with price and profits given by

$$p_f^{MM} = \frac{w - \tau}{\beta} - c \text{ and } \pi^{MM} = \frac{w - \tau}{\beta} - c. \tag{11}$$

On the other hand, if $v > \overline{v}^{MM}$, the follow-on drug is never introduced and the original drug is priced such that the market is fully covered, with price and profits given by

$$p_o^{MM} = \frac{v - \tau}{\beta} - c \text{ and } \pi^{MM} = \frac{v - \tau}{\beta} - c. \tag{12}$$

4.1.3 Second-period outcome with launch delay (market structure GM)

If the launch of the follow-on drug is delayed until the second period, the producer can sell this drug under exclusivity in this period, facing only competition from (generic versions of) the original drug that is sold at a price equal to c, which we henceforth refer to as market structure GM. In this case, the profit-maximising price of the follow-on drug, denoted by p_f^{GM}, is given by

$$p_f^{GM} = \begin{cases} \frac{w+\beta c}{2\beta} & if \quad v \leq v_1^{GM} \\ \frac{v+w-\tau}{\beta} - c & if \quad v_1^{GM} < v \leq v_2^{GM} \\ c + \frac{w-v+\tau}{2\beta} & if \quad v > v_2^{GM} \end{cases}, \tag{13}$$

where

$$v_1^{GM} := \tau + \frac{3\beta c}{2} - \frac{w}{2} \tag{14}$$

and

$$v_2^{GM} := \tau + \frac{4\beta c}{3} - \frac{w}{3}. \tag{15}$$

If the quality of the original drug is sufficiently low, $v \leq v_1^{GM}$, the profit-maximising price of the follow-on drug is such that the market is not fully covered. This means that the original drug and its generic copies are not *de facto* competitors to the follow-on drug even if they are priced at marginal cost. In this case, the producer sets the price of the follow-on drug as if it was a monopoly product in a separate market.

On the contrary, if $v > v_1^{GM}$, it is optimal for the producer to price the follow-on drug such that the market is fully covered and all patients are prescribed one of the drug versions. In this case, there is *de facto* competition between the two drug versions, and the profit-maximising price of the follow-on drug is set as a best-response to the price c of the original drug (and its generic copies). This price is a piecewise linear function of v that increases (decreases) in v for $v < (>)\, v_2^{GM}$. In the interval $v_1^{GM} < v \leq v_2^{GM}$, the producer's profits are maximised by setting a price of the follow-on drug such that the marginal patient (for whom the physician is indifferent between prescribing the original and the follow-on drug) has a therapeutic benefit that is just equal to the perceived treatment cost. Within this interval of v, a higher quality of the original drug therefore allows the producer to *increase* the price of the follow-on drug in order to keep the physician indifferent between all three treatment choices for the marginal patient. However, for sufficiently high values of v, more specifically $v > v_2^{GM}$, it is better for the producer to set a price that leaves a strictly positive surplus (i.e., therapeutic benefit higher than perceived treatment cost) for the marginal patient, in order to boost the demand for the follow-on drug. Thus, as long as $v > v_2^{GM}$, a higher quality of the original drug induces the producer to *reduce* the price of the follow-on drug in order to dampen the resulting demand loss.

The above described pricing strategy generates the following second-period profits, denoted by π^{GM}, in case of launch delay:

$$\pi^{GM} = \begin{cases} \frac{(w-\beta c)^2}{4\beta\tau} & if \quad v \leq v_1^{GM} \\ \frac{(v+w-2\beta c-\tau)[\tau-(v-\beta c)]}{\beta\tau} & if \quad v_1^{GM} < v \leq v_2^{GM} \\ \frac{(w-v+\tau)^2}{8\beta\tau} & if \quad v > v_2^{GM} \end{cases} . \tag{16}$$

As expected, these profits are increasing in the quality of the follow-on drug, decreasing in the quality of the original drug (and its generic copies), and, as before, decreasing in the price sensitivity of the physician's prescription choices.

However, the above derived solution is an equilibrium outcome only if the quality difference

between the two drug versions is not too large. More specifically, the solution exists for a parameter set given by $\underline{v}^{GM} < v < \overline{v}^{GM}$, where

$$\underline{v}^{GM} := w - 3\tau \tag{17}$$

and

$$\overline{v}^{GM} := w + \tau \tag{18}$$

If $v \notin \left(\underline{v}^{GM}, \overline{v}^{GM}\right)$, there are two possible remaining outcomes. If $v > \overline{v}^{GM}$, the follow-on drug cannot profitably survive in the market even under exclusivity and the producer earns zero profits in the second period.

On the other hand, if $v < \underline{v}^{GM}$, the quality difference is so large in favour of the follow-on drug that the original drug and its generic copies are driven out of the market. In the resulting equilibrium, the follow-on drug serves all patients in a fully covered market with a price that is just low enough to make the physician indifferent between prescribing this drug and the original drug (or one of its generic copies) for the patient with the lowest therapeutic benefit of the follow-on drug (i.e., the patient with $x = 0$). Thus, if $v < \underline{v}^{GM}$, the pricing of the follow-on drug is constrained by *potential competition* from generic versions of the original drug. The resulting price and profits are then given by, respectively,

$$p_f^{GM} = c + \frac{w - v - \tau}{\beta} \tag{19}$$

and

$$\pi^{GM} = \frac{w - v - \tau}{\beta}. \tag{20}$$

4.2 Price effects of immediate launch

In order to facilitate the understanding of the producer's optimal launch decision, it is instructive first to look at how different launch decisions affect drug prices. In particular, we are interested in investigating how immediate launch of the follow-on drug affects the price of the original drug in the first period. The answer is given by the following Lemma:[8]

Lemma 1 *As long as a first-period launch of the follow-on drug leads to some substitution between the two drug versions, the price of the original drug increases as a result of the launch.*

[8] See Appendix B for a formal proof.

Thus, a first-period introduction of the follow-on drug might not only increase the producer's total demand (in case the market is not fully covered with only the original drug), but it will also induce the producer to charge a higher price for the pre-existing original drug. The reason is that offering two horizontally differentiated drugs (instead of one) enables the producer to extract more surplus from inframarginal patients by charging higher prices. The easiest way to explain the intuition for this is to consider a specific example. Suppose that the producer launches the follow-on drug with the same price as the original drug. If this leads to some substitution (i.e., some patients switch from the original to the follow-on drug), then the market will be fully covered with a therapeutic benefit for the marginal patient that is strictly higher than the perceived treatment cost. This implies in turn that the producer can increase the price of the original drug without any loss in total demand. The price increase will only induce the physician to prescribe the follow-on drug instead of the original drug to some patients, and if the follow-on drug is sold at the same price as the original drug before the launch, such a price increase is clearly profitable for the producer.

4.3 The optimal launch time of the follow-on drug

Should the producer launch the follow-on drug in the first period, or delay the launch until the second period? In the former case, the total profits of the producer is given by π^{MM}, whereas, in the second case, the total profits are given by $\pi^{M} + \pi^{GM}$. Thus, the profit gain of launch delay, denoted by $\Delta\pi$, is given by

$$\Delta\pi := \pi^{M} + \pi^{GM} - \pi^{MM}. \tag{21}$$

The following Proposition gives, perhaps surprisingly, an unambiguous answer to the producer's launch time decision problem:[9]

Proposition 1 *Under free pricing, the patent holder has never any incentive to delay the launch of the follow-on drug.*

Thus, for the entire set of parameter values for which the follow-on drug can profitably survive in the market, the producer maximises its profits by launching the drug immediately. This might appear far from obvious, since immediate launch has a cannibalising effect on the producer's own sale of the original drug, while launch delay implies that the follow-on drug

[9] See Appendix B for a formal proof.

instead captures market shares from generic competitors to the original drug. Thus, delaying the launch of the follow-on drug in order to avoid this cannibalisation effect might seem like a profitable solution. However, this intuition turns out to be flawed when the producer can freely set the prices of both drugs.

A key to understanding the launch time incentives under free drug pricing is the result in Lemma 1. Although immediate launch of the follow-on drug has a cannibalising effect on the sales of the original drug, this effect is more than compensated by the fact that the introduction of the follow-on drug allows the producer to adjust the price of the original drug upwards. The profitability of launching the follow-on drug in the first period stems not only from the additional surplus extracted through the sales of this drug, but also from the fact that selling a horizontally differentiated product allows the producer to extract more surplus from inframarginal patients when selling the original drug, as explained in Section 4.2. The latter effect is not present in the case of launch delay, since the second-period price of the original drug is forced down to marginal cost due to generic competition. Thus, it is more profitable to launch the follow-on drug in a period where the producer is also able to control the price of the competing drug, which in this context implies that launch delay is never optimal.

Notice that the result in Proposition 1 holds for all possible parameter configurations in which the follow-on drug can profitably survive in the market, including the case in which the quality of the follow-on drug is so high that it will drive the generic copies of the original drug out of the market if launched in the second period. In this case, which occurs if $v < \underline{v}^{GM}$, the follow-on drug will also fully replace the original drug if launched in the first period. Launch delay therefore implies that the producer is a monopolist (selling the follow-on drug) in two periods, while it is a monopolist only in one period if the follow-on drug is launched immediately. Still, launch delay is not profitable! The reason is that *potential competition* from generic versions of the original drug significantly reduces the amount of surplus that the producer is able to extract from selling the follow-on drug if the launch of this drug is delayed until the second period. In this case, the foregone profits in the first period more than outweigh the additional profit gain of being a monopolist in two periods instead of one.

Finally, notice also that launch delay is never profitable under free pricing even in the absence of discounting between the two periods. If instead second-period profits are discounted when the launch time decision is made, delaying the launch of the follow-on drug becomes even more unprofitable.

5 Launch time decision under constrained drug pricing

The previous analysis is based on the key assumption that pharmaceutical patent holders are able to freely set the prices of their drugs. However, in many real-world pharmaceutical markets, producers face various constraints on their price setting, which implies that the result in Proposition 1 is unlikely to be generally applicable. In this section we will consider two such price constraints. First, we assume that a new drug will only be included in the health plan if the additional health gains are at least as high as the increase in drug purchasing costs, and we consider a case in which this assumption imposes a binding constraint on the producer's price setting when launching the follow-on drug in the first period. Second, we assume that drug prices are directly regulated and consider cases in which the producer faces a binding price cap in at least one of the three possible market structures. Under both assumptions, we investigate how such pricing constraints might affect the incentives for launch delay.

5.1 Drug pricing constrained by the health plan's inclusion decision

Suppose that the producer is in principle free to set the price(s) of its drug(s), but that the drugs will be prescribed to patients only if they are accepted by a health plan that maximises total health benefits net of drug purchasing costs. We assume that the health plan adopts the following inclusion criteria: (i) the original drug is included in the health plan only if the health gains of the drug are at least as high as the purchasing costs, and (ii) the follow-on drug is included in the health plan only if the increase in total health gains is at least as high as the increase in total purchasing costs.

In Appendix A we show that, under free pricing, criterion (i) is satisfied if the physician's prescription choices are sufficiently price-sensitive (i.e., if β is sufficiently high). However, we also show that criterion (ii) is not always satisfied even if $\beta \to 1$. More specifically, if the degree of horizontal differentiation between the original drug and the follow-on drug is sufficiently low, the health plan will include the follow-on drug under free pricing if it is launched in the second-period but not if it is launched in the first-period. In other words, if the producer wants to launch the follow-on drug immediately (in the first period), the producer's drug pricing is constrained by the health plan's requirement that the increase in total purchasing costs must not outweigh the additional health gains generated by new drug. All else equal, this clearly reduces the profitability of immediate launch.

In order to explore the incentives for launch delay in this case, we restrict attention to the parameter set given by $w = v > v^M$, in which the two drugs are horizontally differentiated only, and with a relatively low degree of differentiation. Within this parameter set, the market is fully covered in all market structures under free pricing, but the health plan will not accept the follow-on drug at these prices if the drug is launched in the first period. The reason for this is once more related to Lemma 1 and the discussion in Section 4.2. Even if the market is fully covered, there are positive health gains from introducing the follow-on drug, because it allows some of the patients to switch to a therapeutically more suitable drug. The health plan is therefore willing to accept a price increase that reflects the value of these additional health benefits. However, selling two horizontally differentiated products to a fully covered market also allows the producer to extract more surplus from inframarginal patients, as previously explained, which under free pricing yields a price increase that exceeds the value of the additional health benefits. Thus, in order to have the follow-on drug included in the health plan when launched in the first period, the producer must reduce the prices of both drugs below the unconstrained profit-maximising level. The producer does not face a similar constraint if the follow-on drug is instead launched in the second period, since the price of the original drug is then forced down to marginal cost due to generic competition.

Although the stage is now seemingly set for the launch of the follow-on to be delayed until the second period, the next Proposition shows that this is nevertheless not a profit-maximising strategy for the producer:[10]

Proposition 2 *Suppose that β is sufficiently close to one and that $w = v > v^M$, which imply that drug pricing is constrained by the health plan's inclusion decision in market structure MM, but not in the other market structures. Nevertheless, the producer has no incentive to delay the launch of the follow-on drug.*

Notice that, although the producer's drug pricing is constrained in market structure MM, immediate launch of the follow-on drug still leads to higher drug prices because of the additional health gain it offers.[11] Thus, the constraint imposed by the health plan only means that the price increase is less than what it would have been under free pricing. And even though the producer is restricted from appropriating the value of its innovations to the fullest extent, it is still more profitable to launch the follow-on drug immediately, and extract the value of the

[10] A formal proof is given in the Appendix.

[11] If $w = v$, the follow-on drug yields higher therapeutic benefit for half of the patients in the market.

additional health gain generated by the follow-on drug when sold alongside the original drug under patent protection, than to delay the launch of the follow-on drug until the patent of the original drug expires.

For the sake of analytical tractability, the above result is derived for the case of $w = v$, which implies that the two drugs are only horizontally differentiated. By continuity, the result clearly also holds for sufficiently low degrees of vertical differentiation. Moreover, it is straightforward to show that immediate launch of the follow-on drug is also the most profitable strategy if the degree of vertical differentiation is sufficiently large. Consider for example the case of $w > v+3\tau$, which implies that the follow-on drug will fully outcompete both the original drug (if launched in the first period) and the generic copies of the original drug (if launched in the second period). In this case, if inclusion of a single drug with quality v yields a positive surplus for the health plan, replacing this drug with another drug of quality $w > v$ will clearly give the health plan an even higher surplus. Thus, the pricing constraint does not bind and the result of Proposition 1 applies.

5.2 Price cap regulation

Suppose instead that the producer faces a potentially more severe constraint on its price setting in the form of a price cap, denoted by $\overline{p}$, which binds in at least one of the three possible market structures. In order to make the analysis tractable, we will once more make some simplifying assumptions regarding the therapeutic characteristics of the two drug versions. We start out by assuming that the maximum therapeutic benefit is the same for the two drugs, i.e., $w = v$, which implies that the drugs are only horizontally differentiated. In this case, a binding price cap yields the following profits in each of the three possible market structures:[12]

$$\pi^{M}(\overline{p}) = \begin{cases} \overline{p} - c & if \quad c < \overline{p} \leq \frac{v-\tau}{\beta} \\ (\overline{p} - c)\left(\frac{v-\beta\overline{p}}{\tau}\right) & if \quad \overline{p} > \max\left\{c, \frac{v-\tau}{\beta}\right\} \end{cases}, \tag{22}$$

$$\pi^{MM}(\overline{p}) = \begin{cases} \overline{p} - c & if \quad c < \overline{p} \leq \frac{2v-\tau}{2\beta} \\ (\overline{p} - c)\left(\frac{2(v-\beta\overline{p})}{\tau}\right) & if \quad \overline{p} > \max\left\{c, \frac{2v-\tau}{2\beta}\right\} \end{cases}, \tag{23}$$

$$\pi^{GM}(\overline{p}) = \begin{cases} (\overline{p} - c)\left(\frac{1}{2} - \frac{\beta}{2\tau}(\overline{p} - c)\right) & if \quad c < \overline{p} \leq \frac{2v-\tau-\beta c}{\beta} \\ (\overline{p} - c)\left(\frac{v-\beta\overline{p}}{\tau}\right) & if \quad \overline{p} > \max\left\{c, \frac{2v-\tau-\beta c}{\beta}\right\} \end{cases}. \tag{24}$$

[12] See Appendix A for further details.

In each market structure, the market is fully covered only if the price cap is sufficiently low. By considering all feasible levels of the price cap, ranging from marginal cost to the level where it (just) binds in only one of the three market structures, we derive the following results regarding the profitability of launch delay:[13]

Proposition 3 *Suppose that $w = v$ and that both drugs are subject to the same price cap $\overline{p}$.*

(i) For every therapeutic value $v > \beta c + \frac{\tau}{2}$, there exists a threshold level of the price cap, denoted by $\widetilde{p} > c$, such that delaying the launch of the follow-on drug is profitable if $\overline{p} \in (c, \widetilde{p})$.

(ii) A necessary but not sufficient condition for profitable launch delay is that the price cap binds in market structure MM. However, launch delay might not be profitable even if the price cap binds in all three market structures.

The above Proposition confirms that sufficiently strict price regulation will always make launch delay profitable as long as the maximum therapeutic value is sufficiently large: $v > \beta c + \frac{\tau}{2}$. On the other hand, if $v < \beta c + \frac{\tau}{2}$, the market is not fully covered in either market structure even at prices equal to marginal cost, so there is no *de facto* substitution between the original drug and the follow-on drug, which in turn makes the launch time decision irrelevant (at least in the absence of discounting).

The reason why price regulation might make launch delay profitable is perhaps easiest seen by considering a special case. Suppose that v is so high that the market is fully covered in all three market structures, and suppose also that $\overline{p}$ is so low that the price cap is always binding. In this case, the follow-on drug will take all its demand from the original drug if it is launched in the first period. This will therefore increase the producer's profits only if it can make an upward adjustment of the drug prices (cf. Lemma 1). However, this is not possible under binding price cap regulation, in which case there is no additional profit gain from introducing the follow-on drug in the first period. Thus, it is clearly optimal to delay the launch of the follow-on drug until the second period in order to avoid this pure cannibalisation effect.

However, launch delay might not always be profitable even if the price cap binds in all market structures. If v is sufficiently low, such that the market is not fully covered in market structure M, introducing the follow-on drug alongside the original drug in the first period might still be a profitable strategy because it leads to additional demand, although the producer is not able to increase prices. Even with a binding price cap, this demand gain might be valuable

[13] A formal proof is given in Appendix B.

enough to make it unprofitable to delay the launch of the follow-on drug.

More generally, launch delay is profitable if the additional profits obtained by immediate launch, given by $\pi^{MM}(\overline{p}) - \pi^{M}(\overline{p})$, is sufficiently low. This requires that the price cap binds in market structure MM, which will reduce this profit differential compared with the free pricing scenario. With a binding price cap in MM, the profit differential $\pi^{MM}(\overline{p}) - \pi^{M}(\overline{p})$ is sufficiently low to make launch delay profitable if (i) the demand increase due to immediate launch is sufficiently low, which requires that v is sufficiently high, or if (ii) the value of the demand increase due to immediate launch is sufficiently low, which requires that $\overline{p}$ is sufficiently low.

The results in Proposition 3 are based on a scenario in which the two drug versions are only horizontally differentiated. By continuity, it is quite obvious that the results hold also for a sufficiently small degree of vertical differentiation. However, what if the therapeutic differentiation between the drugs is predominantly vertical? To answer this question, we go to the other extreme and assume that $w > v + 3\tau$, which means that the follow-on drug will monopolise the market during its exclusivity period, regardless of whether it is launched in the first or in the second period. In this case, the profits in M with a binding price cap is still given by (22), while a binding price cap in MM and GM yields $\pi^{MM}(\overline{p}) = \pi^{GM}(\overline{p}) = \overline{p} - c$. Considering once more all feasible values of $\overline{p}$ such that the price cap binds in at least one market structure, the profitability of launch delay under price regulation is then characterised as follows:[14]

Proposition 4 *Suppose that $w > v + 3\tau$ and that both drugs are subject to the same price cap $\overline{p}$.*

(i) For every therapeutic value $v > \beta c$, there exists a threshold value of the price cap, denoted by $\widetilde{p}' > 0$, such that delaying the launch of the follow-on drug is profitable if $\overline{p} \in (c, \widetilde{p}')$.

(ii) A necessary condition for profitable launch delay is that the price cap binds in market structure MM, and a sufficient condition for profitable launch delay is that the price cap additionally binds in either M or GM.

The result in the first part of Proposition 4 mirrors the first part of Proposition 3. Thus, regardless of whether the two drug versions are horizontally or vertically differentiated, sufficiently strict price regulation will always make it profitable to delay the launch of the follow-on

[14] A formal proof is given in Appendix B.

drug until the second period, as long as there is *de facto* substitutability between the two drugs (which is always the case when $w > v + 3\tau$). Furthermore, the second part of Proposition 4 implies that the scope for profitable launch delay is even larger than in the case of purely horizontally differentiated drugs. If $w > v + 3\tau$, launch delay is always profitable as long as the price cap binds in MM and in at least one other market structure.

6 Efficiency and welfare

What are the efficiency and welfare implications of the producer's launch time incentives? In order to answer this question, a useful starting point is to consider the first-best outcome. Suppose that both launch time decisions and drug prescription choices are made by a utilitarian social planner that maximises total welfare, defined as aggregate health benefits net of drug production costs. Each patient would then be prescribed the drug that yields the highest therapeutic benefits, as long as these benefits are at least as high as the marginal cost of production. This means that the follow-on drug would be prescribed to at least some patients, and thus generate a welfare surplus, as long as $w > c$. If this condition holds, any launch delay would therefore represent a welfare loss. Thus:[15]

Proposition 5 *If $w > c$, immediate launch of the follow-on drug is the first-best outcome.*

However, immediate launch of the follow-on drug might not necessarily be the welfare-maximising outcome in a second-best context, where prescription choices are not made by a social planner but instead depend on the purchasing costs of the health plan, and thus indirectly depend on the drug pricing of the patent-holding producer. The reason is that the combination of imperfect physician agency and market power leads to prescription choices being potentially distorted both at the intensive and at the extensive margins.

The distortion at the *extensive* margin stems from two inefficiencies that go in opposite directions. One is caused by the imperfect agency of the prescribing physician. If the physician makes treatment decisions without taking treatment costs fully into account (i.e., if $\beta < 1$), this leads to overconsumption of drug treatments if the drugs are priced at marginal cost and the market is not fully covered. This is, however, counteracted by the market power of the patent-holding producer, which leads to drug prices above marginal costs and therefore contributes to underconsumption of drug treatments, all else equal. If the former effect dominates, which will

[15] The proof of Proposition 5 is trivial and thus omitted.

typically be the case if drug quality is sufficiently low, the introduction of the follow-on drug might be inefficient due to overconsumption of drug treatments, where some patients are given drug treatments even if the therapeutic benefit is lower than the cost of producing the drugs.

The market power of the patent-holding producer not only implies that drugs are priced higher than marginal costs, but it also leads to price differences between drugs of different quality, even though the marginal production costs are the same. This causes a distortion at the *intensive* margin in market structure MM, leading to an allocative inefficiency in the sense that the share of patients that are prescribed the higher-quality drug is suboptimally low. This might also have implications for the welfare-optimal launch time of the follow-on drug. If this drug is of lower quality than the original drug, immediate launch would lead to market structure MM in the first period with $p_f^{MM} < p_o^{MM}$, which in turn would imply lower health benefits for some patients who switch to the lower-quality (but cheaper) follow-on drug. A similar inefficiency would not occur if the follow-on drug is instead launched in the second period, since $p_f^{GM} > p_o^{GM} = c$.

Although the welfare-optimal launch time under free pricing is *a priori* ambiguous, for the reasons explained above, the following parameter restrictions allow us to derive a clear-cut conclusion:[16]

Proposition 6 *Suppose that β is sufficiently close to one and that the follow-on drug is at least of the same quality as the original drug. In this case, under free pricing, total welfare is maximised by immediate launch of the follow-on drug.*

The intuition behind this result is relatively straightforward. If the follow-on drug is launched immediately instead of being delayed, the producer sells two drugs under patent protection instead of one in the first period, while in the second period both drugs face direct generic competition instead of just one. When both drugs are available at prices equal to marginal cost, the only potential source of inefficiency is the imperfect agency of the prescribing physician. Since this source of inefficiency vanishes as β approaches one, welfare is clearly higher with generic competition for both drugs in the second period if β is sufficiently close to one. In this case, a sufficient condition for immediate launch to be a welfare-optimal outcome is that welfare is higher also in the first period; in other words, that welfare is higher in MM than in M. In this period there is a potential distortion of prescription choices at the intensive margin,

[16] A formal proof of Proposition 6 is given in Appendix B.

as previously explained. However, this distortion potentially makes welfare lower in MM than in M only if $w < v$. On the contrary, if the follow-on drug is of at least as high quality as the original drug, every patient that switches to from the original to the follow-on drug gets a higher therapeutic benefit. If in addition the physician is sufficiently close to being a perfect agent for the health plan, welfare is unambiguously higher in MM than in M, thus ensuring that immediate launch of the follow-on drug is the welfare-maximising outcome under free pricing. Notice that the parameter restriction given in Proposition 6 is a sufficient but not necessary condition for launch delay to be welfare detrimental. The result applies for a substantially wider set of parameters, but which is somewhat hard to characterise in a clear way.

When seen in conjunction with Proposition 1, the result in Proposition 6 implies that free pricing yields welfare-optimal launch time incentives for a wide set of parameters. Thus, and perhaps surprisingly, market power in itself is not likely to cause inefficiencies in the timing of follow-on drug launches. Our results in Propositions 3 and 4 reveal instead that *price regulation* might lead to inefficient outcomes in the form of drug launch delays.

Discussions of how price regulation affects welfare are usually related to the well-known trade-off between dynamic and static efficiency, where the typical concern is that (excessively strict) price regulation might harm dynamic efficiency by providing insufficient incentives for developing new drugs. In our analysis we identify another (though related) potential inefficiency caused by price regulation, namely delayed launch of drugs that have already been developed. Moreover, our results suggest that stricter price regulation might be harmful not only in terms of overall welfare, but it might also in some cases reduce the surplus of the health plan. This might happen if a lower price cap leads to delayed launch of the follow-on drug, and the reduction in purchasing costs due to a lower price cap is outweighed by the reduction in health benefits due to the launch delay. This would always be the case if, for example, the price cap is reduced from slightly above to slightly below the price cap thresholds identified in Propositions 3 and 4.

7 Extension: brand-loyal patients

Our main analysis is based on the simplifying assumption that generic competition drives drug prices down to marginal cost, implying that positive profits from drug sales are only made during periods of exclusivity. Although the observed low brand-name market shares for off-patent drugs in the real world suggest that this might be a reasonable approximation, it is

also a well-established fact that many brand-name drugs survive in the market (with positive market shares) after patent expiry despite being priced significantly higher than their generic competitors, which suggests that there might be an element of (perceived) vertical differentiation between brand-name and generic drugs.

In this section we check the robustness of our main result in Proposition 1, under free pricing, by extending our model to accommodate the possibility of profitable off-patent sales of brand-name drugs. We do so by assuming that, for each value of the therapeutic mismatch variable x, the prescribing physician considers the brand-name drug and its generic versions to be perfect substitutes only for a share $1-\lambda$ of the patients. For the remaining share λ, the generic versions are considered to be therapeutically inferior to the extent that generic substitution is never an option for these patients. We will henceforth refer to these two patient segments as the *generic substitution segment* and the *brand-loyal segment*, respectively. Our analysis is simplified by assuming that $\lambda < \frac{1}{3}$. This is arguably an innocuous assumption, given that brand-name market shares in most real-world off-patent pharmaceutical markets are well below one third. As a further simplifying assumption, we also restrict attention to the case of $w = v$, implying that the original and follow-on drugs are horizontally differentiated only.

Obviously, the existence of a brand-loyal segment potentially affects drug prices and profits only in market structures in which at least one of the brand-name drugs faces direct generic competition. Thus, the prices and profits in M and MM remain unaffected.

However, in case of immediate launch of the follow-on drug in the first period, the assumption of a brand-loyal segment implies that the firm will also earn positive profits in the second period, when both drugs face generic competition. In this market structure, which we refer to as GG, the brand-name producer cannot profitably compete with the generic producers in the generic substitution segment and will therefore set prices to maximise profits from the brand-loyal segment only. Since the distribution of therapeutic benefits across patients in this segment is, by assumption, identical to the distribution of therapeutic benefits in the overall patient population, the producer's pricing incentives in GG are identical to the ones in MM. Thus, the profit-maximising price of the two brand-name drugs in market structure GG is given by

$$p_o^{GG}(\lambda) = p_f^{GG}(\lambda) = \begin{cases} \frac{v+c\beta}{2\beta} & if \quad v \leq \beta c + \tau \\ \frac{2v-\tau}{2\beta} & if \quad v > \beta c + \tau \end{cases}. \tag{25}$$

However, since only patients in the brand-loyal segment will be prescribed a brand-name drug

at these prices, demand for the original and the follow-on drug is obviously lower in market structure GG than in market structure MM. Profits in GG are therefore given by

$$\pi^{GG}(\lambda) = \begin{cases} \lambda\frac{(v-\beta c)^2}{2\tau\beta} & if \quad v \leq \beta c + \tau \\ \lambda\left(\frac{2v-\tau}{2\beta} - c\right) & if \quad v > \beta c + \tau \end{cases} . \tag{26}$$

Thus, when the producer sells two horizontally differentiated brand-name drugs that lose patent protection at the same time, generic entry implies that brand-name drug prices stay at the same level, while demand for these drugs drop by $100\,(1-\lambda)$ per cent.

The presence of a brand-loyal patient segment also affects prices and profits in market structure GM. We show in Appendix A that the profit-maximsing prices of the original drug and the follow-on drug are given by, respectively,

$$p_o^{GM}(\lambda) = \begin{cases} \frac{v+c\beta}{2\beta} & if \quad \beta c < v \leq \beta c + \tau \\ \frac{4(1+\lambda)v-2(1-\lambda)\beta c-(3+\lambda)\tau}{2(1+3\lambda)\beta} & if \quad \beta c + \tau < v \leq \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \\ c + \frac{\tau}{(1-\lambda)\beta} & if \quad v > \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \end{cases} , \tag{27}$$

and

$$p_f^{GM}(\lambda) = \begin{cases} \frac{v+c\beta}{2\beta} & if \quad \beta c < v \leq \beta c + \frac{2}{3}\tau \\ \frac{2v-\tau}{\beta} - c & if \quad \beta c + \frac{2}{3}\tau < v \leq \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau \\ \frac{(1-\lambda)\tau+2(\lambda v+\beta c)}{2(1+\lambda)\beta} & if \quad \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau < v \leq \beta c + \tau \\ \frac{2\beta(1-\lambda)c+8\lambda v+(1-5\lambda)\tau}{2(1+3\lambda)\beta} & if \quad \beta c + \tau < v \leq \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \\ c + \frac{(1+\lambda)\tau}{2(1-\lambda)\beta} & if \quad v > \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \end{cases} , \tag{28}$$

while the total profits are given by

$$\pi^{GM}(\lambda) = \begin{cases} \frac{1+\lambda}{4\tau\beta}(v-\beta c)^2 & if \quad \beta c < v \leq \beta c + \frac{2}{3}\tau \\ \frac{4(3(v-\beta c)-\tau)\tau-(8-\lambda)(v-\beta c)^2}{4\beta\tau} & if \quad \beta c + \frac{2}{3}\tau < v \leq \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau \\ \frac{2\lambda(1+3\lambda)(v-\beta c)^2+(1-\lambda)(4\lambda(v-\beta c)+(1-\lambda)\tau)\tau}{8(1+\lambda)\beta\tau} & if \quad \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau < v \leq \beta c + \tau \\ \frac{\left(8\lambda(5-\lambda)(v-\beta c)-\left(18\lambda-\lambda^2-1\right)\tau\right)\tau-16(1-\lambda)\lambda(v-\beta c)^2}{8(1+3\lambda)\beta\tau} & if \quad \beta c + \tau < v \leq \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \\ \frac{(1+3\lambda)\tau}{8(1-\lambda)\beta} & if \quad v > \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau \end{cases} . \tag{29}$$

The optimal prices are stepwise linear functions of the maximum therapeutic benefit v. If this benefit is sufficiently low, $v < \beta c + \frac{2}{3}$, both patient segments are only partially covered and there is no *de facto* substitution between different drugs in either segment. However, if

the therapeutic benefit is somewhat higher, $v \in \left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$, the optimal prices yield full coverage in the generic substitution segment, but the pricing of the follow-on drug is restricted in the sense that $p_f^{GM}(\lambda)$ is kept just low enough for this segment to be fully covered. For even higher therapeutic benefits, $v \in \left(\beta c + \frac{3+\lambda}{2(2+\lambda)}, \beta c + \tau\right)$, there is still partial coverage in the brand-loyal segment and full coverage in the generic substitution segment, but the price $p_f^{GM}(\lambda)$ is now an unrestricted best-response to the price $p_o = c$ in the latter segment. For yet higher therapeutic benefits, $v \in \left(\beta c + \tau, \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau\right)$, both segments are fully covered but the price of the original drug, $p_o^{GM}(\lambda)$, is restricted to a level that is just low enough to make the brand-loyal segment fully covered. And finally, if the therapeutic benefit is sufficiently high, $v > \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau$, the profit-maximising pair of prices is an (unrestricted) interior solution in which both segments are fully covered.

As previously mentioned, the presence of brand-loyal patients implies that the firm earns positive profits also after the end of the exclusivity period; in other words, the firm earns positive profits beyond the two periods considered in our analysis. Notice, however, that launch delay only implies that the market structure GG is postponed by one period compared with the case of immediate launch. Thus, in order to assess the profitability of launch delay, we only need to compare the profits over the two periods in which the market structures are different under launch delay and immediate launch. In case of immediate launch, the producer faces market structure MM in the first of these two periods and market structure GG in the second. In the case of launch delay, on the other hand, the producer faces market structure M in the first period and GM in the second. Thus, the profitability of launch delay is now given by

$$\Delta\pi(\lambda) := \pi^{M}(\lambda) + \pi^{GM}(\lambda) - \pi^{MM}(\lambda) - \pi^{GG}(\lambda). \tag{30}$$

The next Proposition describes the incentives for launch delay in the presence of brand-loyal patients under free pricing:[17]

Proposition 7 *Suppose that $w = v$ and that a share $\lambda < \frac{1}{3}$ of patients are brand loyal and thus not susceptible to generic substitution. Under free pricing, the patent holder has still no incentive to delay the launch of the follow-on drug.*

This result confirms the robustness of Proposition 1. Even if the presence of a brand-loyal patient segment allows the producer to earn positive profits also in face of direct generic

[17] See Appendix B for a formal proof.

competition, is has never any incentive to delay the launch of a horizontally differentiated follow-on drug under free pricing. The presence of brand-loyal patients has two counteracting effects on the profitability of launch delay, in the sense that it leads to higher second-period profits regardless of whether the launch of the follow-on drug is delayed or not.

If the launch is delayed, the producer earns higher profits in market structure GM for potentially two different reasons. First, because it earns positive profits from selling the original drug despite direct generic competition. Moreover, if v is so high that the brand-loyal market is fully covered, the producer benefits from being able to control the price of both of the two horizontally differentiated drugs sold to the brand-loyal consumers (cf. the intuition for Lemma 1), implying that the optimal price of the follow-on drug in market structure GM is higher with than without brand-loyal patients. Notice, however, that the assumed inability to price discriminate between different types of patients implies that a price increase to extract more surplus from inframarginal brand-loyal patients comes at the cost of a lower market share in the generic substitution segment.

On the other hand, if the follow-on drug is launched immediately, the presence of brand-loyal patients implies that the producer earns positive profits also in the second period, in market structure GG, where the producer earns positive profits from the sales of both the original drug and the follow-on drug. And furthermore, the prices of the two horizontally differentiated drugs can be set to maximise the surplus from brand-loyal patients without consideration for any loss of market share in the generic substitution segment. For these reasons, the profit gain in market structure GM does not sufficiently outweigh the foregone profits in market structure GG to make launch delay profitable.

8 Policy implications and concluding remarks

If new versions of a drug offer improved treatment outcomes or better adherence, they should be introduced in the market as soon as they have been recognised. Instead, we see that follow-on drugs tend to be introduced shortly before the patent protection of original versions expires. This is often seen as evidence of life cycle management practices that strategically delay new versions to extend exclusivity periods beyond what is justified by the original innovation. Our analysis challenges the conventional view that patent protection and market power inherently lead to strategic delays in the introduction of follow-on drugs. Instead, we demonstrate that

only under certain regulatory constraints, particularly price cap regulations, may pharmaceutical companies find it optimal to delay the launch of valuable line extensions until the patent expiration of the original drug. Without these regulatory constraints, we have shown that pharmaceutical companies have incentives to launch extensions in competition with their still-protected original drug, even if this leads to full cannibalisation of the original drug's sales. This finding suggests that inefficiencies in drug launch timing are more attributable to price regulation policies than to the patents themselves.

Thus, policymakers should carefully consider the potential adverse effects of these regulatory constraints on pharmaceutical innovation and patient welfare. Empirical studies have found that welfare losses due to delayed entry can be large, and that removal of exclusivity protection for line extensions can be welfare improving, even if this means that therapeutic valuable line extensions are not developed. Future policy deliberations might benefit from exploring regulatory frameworks that balance the need to control drug prices with the imperative to ensure timely access to improved therapeutic options. According to our analysis, relaxing price caps for patented drugs could have the positive effect of triggering immediate launch of improvements. More research is needed to explore the trade-off between price control and earlier access to valuable new versions of existing drugs. There is an obvious cost of laxer price control to stimulate early access, and this needs to be compared with the cost of not granting exclusivity and missing out on line extensions altogether.

Appendix A: Supplementary calculations

This Appendix contains supplementary calculations and details underlying the price and profit expressions presented in the paper.

Market structure M under free pricing

In case of launch delay, the producer is a single-product monopolist in the first period, selling only the original drug. By maximising (1) for each potential patient in the market, demand for the original drug, denoted by y_o, is given by

$$y_o = \begin{cases} 1 & if \quad p_o \leq \frac{v-\tau}{\beta} \\ \frac{v-\beta p_o}{\tau} & if \quad p_o > \frac{v-\tau}{\beta} \end{cases} . \tag{A1}$$

The profit-maximising solution is either on the price-elastic or on the price-inelastic part of the demand curve. Assume first that the solution is on the price-elastic part. The profit-maximising price is then given by the solution to the following problem:

$$\max_{p_o} (p_o - c) \left(\frac{v - \beta p_o}{\tau} \right), \tag{A2}$$

which yields

$$p_o^M = \frac{v + \beta c}{2\beta}. \tag{A3}$$

By using (A1), it is easily verified that this price yields an interior solution (i.e., $y_o < 1$) if $v < v^M$, where v^M is given by (3). If $v > v^M$, the profit-maximising solution is a corner solution in which the producer sets a price that makes the physician indifferent between prescribing or not the drug to the patient with the lowest therapeutic benefit (i.e., the patient with $x = 1$). This price is therefore implicitly given by

$$v - \beta p_o^M - \tau = 1, \tag{A4}$$

and explicitly given by

$$p_o^M = \frac{v - \tau}{\beta}. \tag{A5}$$

At the profit-maximising price, the sales of the original drug are then given by

$$y_o^M = \begin{cases} \frac{v - \beta c}{2\tau} & if \quad v \leq v^M \\ 1 & if \quad v > v^M \end{cases}, \tag{A6}$$

from which we can easily verify that positive demand requires $v > \beta c$.

Market structure MM under free pricing

If the follow-on drug is launched immediately, the producer is a multi-product monopolist, selling both the original drug and the follow-on drug in the first period. The profit-maximising prices might in principle yield either full or partial market coverage.

Assume first that the profit-maximising prices are such that the market is fully covered. In this case, the physician is indifferent between prescribing the original and the follow-on drug to

a patient characterised by $\widehat{x}$, which is implicitly given by

$$v - \beta p_o - \tau \widehat{x} = w - \beta p_f - \tau \left(1 - \widehat{x}\right), \tag{A7}$$

and explicitly given by

$$\widehat{x} = \frac{1}{2} + \frac{v - w - \beta \left(p_o - p_f\right)}{2\tau}, \tag{A8}$$

which in turn implies that demand for the original and follow-on drug are given by $y_o = \widehat{x}$ and $y_f = 1 - \widehat{x}$, respectively.

Under full market coverage, the profit-maximising pair of prices must necessarily be such that the therapeutic benefit is equal to the perceived treatment cost for the marginal patient (i.e., the patient for which $x = \widehat{x}$). If not, the producer could increase both prices by the same proportion without affecting demand for either drug, which would clearly be profitable. Thus, the profit-maximising prices must be such that

$$v - \beta p_o - \tau \widehat{x} = 0, \tag{A9}$$

which implies

$$p_o = \frac{w + v - \tau}{\beta} - p_f. \tag{A10}$$

The profit-maximising price pair is then given by (A10) and the solution to the following problem:

$$\max_{p_f} \left(p_o - c\right) y_o + \left(p_f - c\right) y_f \quad \text{subject to (A10)}. \tag{A11}$$

The unique solution to this problem is

$$p_o^{MM} = \frac{3w + v - 2\tau}{4\beta} \tag{A12}$$

and

$$p_f^{MM} = \frac{3v + w - 2\tau}{4\beta}. \tag{A13}$$

In this case, the sales of the two drugs are given by

$$y_o^{MM} = 1 - y_f^{MM} = \frac{1}{2} - \left(\frac{w - v}{4\tau}\right) \tag{A14}$$

Alternatively, the profit-maximising prices might be such that the market is only partially covered. If this is the case, demand for the two drugs are given by

$$y_o = \frac{v - \beta p_o}{\tau} \tag{A15}$$

and

$$y_f = \frac{v - \beta p_f}{\tau}. \tag{A16}$$

The profit-maximising prices are then given as the solutions to the following two independent maximisation problems:

$$\max_{p_o} \pi_o = (p_o - c)\, y_o \tag{A17}$$

and

$$\max_{p_f} \pi_f = (p_f - c)\, y_f. \tag{A18}$$

These solutions are given by

$$p_o^{MM} = \frac{v + \beta c}{2\beta} \tag{A19}$$

and

$$p_f^{MM} = \frac{w + \beta c}{2\beta}, \tag{A20}$$

which yield the following sales of the two drugs:

$$y_o^{MM} = \frac{v - \beta c}{2\tau}, \tag{A21}$$

$$y_f^{MM} = \frac{w - \beta c}{2\tau}. \tag{A22}$$

This solution actually results in a partially covered market if $y_o^{MM} + y_f^{MM} < 1$,which is true for $v \leq v^{MM}$, where v^{MM} is given by (7). Otherwise, if $v > v^{MM}$, the profit-maximising prices are given by (A12)-(A13) and the market is fully covered. Furthermore, it can easily be verified from (A14) and (A21)-(A22) that positive demand for both drugs requires $\min\{v, w\} > \beta c$ and $|w - v| < 2\tau$.

Market structure GM with free pricing

In case of launch delay, the follow-on drug faces competition from generic versions of the (now off-patent) original drug in the second period. By assumption, generic competition leads to

$p_o = c$. The profit-maximising price of the follow-on drug is therefore a best-response to this price. Once more, this might yield either a fully or partially covered market.

Assume first that the profit-maximising price for the follow-on drug results in a fully covered market. In this case, demand for the follow-on drug is given by $y_f = 1 - \widehat{x}$, with $\widehat{x}$ given by (A8) for $p_o = c$. Profit maximisation therefore yields a price of

$$p_f^{GM} = c + \frac{\tau + w - v}{2\beta}, \tag{A23}$$

which in turn yields sales of the two drug versions equal to

$$y_o^{GM} = \frac{v - w + 3\tau}{4\tau} \tag{A24}$$

and

$$y_f^{GM} = 1 - y_o^{GM} = \frac{w - v + \tau}{4\tau}. \tag{A25}$$

This solution yields positive sales for both drugs if $w - 3\tau < v < w + \tau$. This solution also requires that the therapeutic benefit is at least as high as the perceived treatment cost for the marginal patient (and thus for all patients). This is true if

$$v - \beta p_o - \tau y_o \geq 0 \Leftrightarrow v > v_2^{GM}, \tag{A26}$$

where v_2^{GM} is given by (15). If this condition does not hold, i.e., if $v < v_2^{GM}$, the profit-maximising price of the follow-on drug is one that is just low enough to ensure that the therapeutic benefit is equal to the perceived treatment cost for the marginal patient. This price is implicitly given by

$$v - \beta c - \tau \widehat{x} = 0, \tag{A27}$$

where $\widehat{x}$ is given by (A8) for $p_o = c$, and explicitly given by

$$p_f^{GM} = \frac{v + w - \tau}{\beta} - c, \tag{A28}$$

which implies that the sales of the follow-on drug is equal to

$$y_f^{GM} = 1 - \left(\frac{v - \beta c}{\tau} \right). \tag{A29}$$

Alternatively, the profit-maximising price of the follow-on drug might yield only a partially covered market. In this case, the demand for the follow on drug is given by (A16), and the profit-maximising price is therefore given by (A20). In this solution, the sales of the two drug versions are

$$y_o^{GM} = \frac{v - \beta c}{\tau} \tag{A30}$$

and

$$y_f^{GM} = \frac{w - \beta c}{2\tau}. \tag{A31}$$

At these prices, the market is only partially covered if

$$y_o^{GM} + y_f^{GM} < 1, \tag{A32}$$

which is true if $v \leq v_1^{GM}$. Since $v_1^{GM} < v_2^{GM}$, the profit-maximising price of the follow-on drug in the second period, in case of launch delay, is fully characterised by (13).

Drug pricing constrained by the health plan's inclusion decision

Consider the health plan's choice of including the follow-on drug in the first period under free pricing. If the drug is not included, the market structure in the first period remains M, with a price of the original drug price given by (2). The health plan's surplus is then given by

$$H^M = \int_0^{y_o^M} \left(v - p_o^M - \tau x\right) dx = \begin{cases} \frac{((3\beta-2)v-(2-\beta)\beta c)(v-c\beta)}{8\beta\tau} & if \quad v \leq v^M \\ \frac{(2-\beta)\tau-2(1-\beta)v}{2\beta\tau} & if \quad v > v^M \end{cases}, \tag{A33}$$

where y_o^M is given by (A6) and v^M is given by (3). It is easily verified that $H^M > 0$ if β is sufficiently close to one, which we assume to be the case.[18] On the other hand, in case of inclusion, the first-period market structure is MM, with drug prices under free pricing given by (5)-(6). In this case, the health plan's surplus is given by

$$\begin{aligned} H^{MM} &= \int_0^{y_o^{MM}} \left(v - p_o^{MM} - \tau x\right) dx + \int_{y_o^{MM}}^1 \left(v - p_f^{MM} - \tau\left(1 - x\right)\right) dx \\ &= \begin{cases} \frac{(3\beta-2)\left(v^2+w^2\right)+2\beta c((2-\beta)\beta c-\beta(v+w))}{8\beta\tau} & if \quad v \leq v^{MM} \\ \frac{(3\beta-2)(w-v)^2+4((2-\beta)\tau-2(1-\beta)(v+w))\tau}{16\beta\tau} & if \quad v > v^{MM} \end{cases}, \end{aligned} \tag{A34}$$

[18] A closer inspection of () reveals that the threshold level of β, above which $H^M > 0$, is higher than $\frac{2}{3}$.

where y_o^{MM} is given by (A14) and v^{MM} is given by (7). If launched in the first period, the health plan will therefore include the follow-on drug if

$$\Delta H_1 := H^{MM} - H^M. \tag{A35}$$

Since $\beta c < v^{MM} < v^M$, there are three different intervals of v to consider. Suppose first that $v \in \left(\beta c, v^{MM}\right)$, in which case the effect of the follow-on drug on the health plan's surplus under free pricing is given by

$$\Delta H_1 = \frac{\left(\left(3\beta - 2\right) w - \beta \left(2 - \beta\right) c\right) \left(w - \beta c\right)}{8\beta\tau} > 0, \tag{A36}$$

where the positive sign holds for all values of β for which $H^M > 0$. Thus, the health plan will include the follow-on drug in the first period under free pricing for all $v \in \left(\beta c, v^{MM}\right)$.

Consider next the case of $v \in \left(v^{MM}, v^M\right)$, in which the effect of the follow-on drug on the health plan's surplus is given by

$$\Delta H_1 = \frac{\left(3\beta - 2\right) \left(\left(w - v\right) \left(v + w\right) - 2wv\right) + 2 \left(2 \left(v - c\right) + \beta c\right) \beta^2 c + 4 \left(\left(2 - \beta\right) \tau - 2 \left(1 - \beta\right) \left(v + w\right)\right) \tau}{16\beta\tau}, \tag{A37}$$

which has an ambiguous sign. However, it is easily verified that

$$\lim_{v \to v^{MM}} \Delta H_1 = \frac{2 \left(\left(3\beta - 2\right) w - \beta \left(2 - \beta\right) c\right) \left(w - \beta c\right)}{16\beta\tau} > 0, \tag{A38}$$

where the positive sign once more holds for all β for which $H^M > 0$. Thus, by continuity, the health plan will include the follow-on drug in the first period under free pricing if v is sufficiently close to the lower bound v^{MM}.

Finally, consider the case of $v > v^M$, in which the effect of the follow-on drug on the health plan's surplus is given by

$$\Delta H_1 = \frac{\left(2\tau + w - v\right) \left[\left(3\beta - 2\right) \left(w - v\right) - 2 \left(2 - \beta\right) \tau\right]}{16\beta\tau} < 0. \tag{A39}$$

The negative sign is confirmed by imposing the equilibrium condition $w - v < 2\tau$, and the condition $\beta > \frac{2}{3}$, which is necessary for $H^M > 0$. Thus, the health plan will not include the follow-on drug in the first period under free pricing if $v > v^M$. Notice that a sufficiently high value of v can be interpreted as a sufficiently low degree of horizontal differentiation, since the

latter is measured by the parameter τ *relative to* the maximum therapeutic benefit v (or w).

We also want to show that, for $v > v^M$, inclusion of the follow-on drug will always increase the health plan's surplus if it is instead launched in the second-period. In this case, inclusion of the drug leads to market structure GM, with the unconstrained profit-maximising price given by (13). The health plan's surplus is then given by

$$\begin{aligned} H^{GM} &= \int_0^{y_o^{GM}} (v - c - \tau x)\, dx + \int_{y_o^{GM}}^1 \left(w - p_o^{GM} - \tau x\right) dx \\ &= \frac{(3\beta - 2)(w - v)^2 + (2\beta(5v + 3w - 8c) - (2 + 5\beta)\tau - 4(w - v))\tau}{16\beta\tau}, \end{aligned} \tag{A40}$$

where y_o^{GM} is given by (A24). If instead the follow-on drug is not included, the market will consist only of the original drug and its generic competitors, which we will label as market structure G, with a drug price equal to c. For $v > v^M$, this will yield a fully covered market and the health plan's surplus is then given by

$$H^G = \int_0^1 (v - c - \tau x)\, dx = v - c - \frac{\tau}{2}. \tag{A41}$$

From (A40)-(A41) we derive

$$\Delta H_2 := H^{GM} - H^G = \frac{(w - v + \tau)^2 (3\beta - 2)}{16\beta\tau} > 0, \tag{A42}$$

where the positive sign is once more established by applying the condition $\beta > \frac{2}{3}$, which is necessary for H^M. Thus, for $v > v^M$, the health plan will accept the follow-on drug at the unconstrained profit-maximising price if it is launched in the second period.

Price cap regulation

Suppose that the producer faces a binding price cap, given by $\overline{p}$. In market structure M, the demand for the original drug is found by setting $p_o = \overline{p}$ in (A1), which immediately yields the profit expression given by (22).

In market structure MM, the patient for which the physician is indifferent between prescribing the original drug and the follow-on drug is characterised by $x = \frac{1}{2}$ when $w = v$. The therapeutic benefit net of perceived treatment costs for this patient is given by $v - \beta\overline{p} - \frac{\tau}{2}$, which is non-negative, implying that the market is fully covered (with total demand equal to

one), if $\overline{p} \leq \frac{2v-\tau}{2\beta}$. Otherwise, if $\overline{p} > \frac{2v-\tau}{2\beta}$, some patients are not prescribed drug treatment and total demand for the producer is twice the demand under partial market coverage in market structure M. This yields total profits as given by (23).

Finally, in market structure GM, the patient for whom the physician is indifferent between prescribing the follow-on drug and (a generic copy of) the original drug is characterised by $x = \frac{1}{2} + \frac{\beta}{2\tau}(\overline{p} - c)$ when $w = v$. For this patient, the therapeutic benefit net of perceived treatment costs is given by $v - \beta\overline{p} - \tau\left(\frac{1}{2} - \frac{\beta}{2\tau}(\overline{p} - c)\right)$, which is non-negative, implying that the market is fully covered, if $\overline{p} \leq \frac{2v-\tau-\beta c}{\beta}$, in which case demand for the follow-on drug is given by $\frac{1}{2} - \frac{\beta}{2\tau}(\overline{p} - c)$. Otherwise, if $\overline{p} > \frac{2v-\tau-\beta c}{\beta}$, some patients will not be prescribed any of the drugs available in the market, and the demand for the follow-on drug is identical to the demand of the original drug in market structure M. This yields total profits as given by (24).

Brand-loyal patient segment

Suppose that, for each value of x, generic substitution is only an option for a share $1 - \lambda$ of the patients, whereas the remaining share λ consists of brand-loyal patients.

Consider first the case in which p_o and p_f are such that both the brand-loyal segment and the generic substitution segment are only partially covered. In this case, there are two marginal patients in the former segment, characterised by

$$x_{b1} = \frac{v - \beta p_o}{\tau} \tag{A43}$$

and

$$x_{b2} = 1 - \left(\frac{v - \beta p_f}{\tau}\right) \tag{A44}$$

while the marginal patient in the latter segment is characterised by $x_g = x_{b2}$. Total demand for the original drug, which comes only from the brand-loyal segment, is then given by

$$q_o = \lambda x_{b1} = \frac{\lambda(v - \beta p_o)}{\tau}, \tag{A45}$$

while total demand for the follow-on drug, which comes from both segments, is given by

$$q_f = \lambda(1 - x_{b2}) + (1 - \lambda)(1 - x_g) = \frac{v - \beta p_f}{\tau}. \tag{A46}$$

Profit maximisation yields

$$p_o^{GM}(\lambda) = p_f^{GM}(\lambda) = \frac{v + \beta c}{2\beta}, \tag{A47}$$

which implies

$$x_{b1}^{GM} = \frac{v - \beta c}{2\tau}, \quad x_{b2}^{GM} = x_g^{GM} = 1 - \left(\frac{v - \beta c}{2\tau}\right), \tag{A48}$$

and

$$\pi^{GM}(\lambda) = \frac{(1+\lambda)(v - \beta c)^2}{4\beta\tau}. \tag{A49}$$

Since generic version of the original drug is available at price c, these will be prescribed to all patients in the generic substitution segment characterised by $x \leq \frac{v-\beta c}{\tau}$. This market segment will therefore be partially covered only if $x_g^{GM} > \frac{v-\beta c}{\tau}$, which holds if $v < \beta c + \frac{2}{3}\tau$. Thus, the profit-maximising prices are given by (A47) for $v \in \left(c, \beta c + \frac{2}{3}\tau\right)$.

If $v > \beta c + \frac{2}{3}\tau$, the prices given by (A47) are no longer optimal. Suppose that $v > \beta c + \frac{2}{3}\tau$, but still low enough for the brand-loyal segment to be partially covered at the profit-maximising prices. In this case, the marginal patient in the generic substitution segment derives the net therapeutic benefit of being prescribed the follow-on drug or (a generic version of) the original drug, and is thus characterised by

$$x_g' = \frac{1}{2} + \frac{\beta(p_f - c)}{2\tau}. \tag{A50}$$

The marginal patients in the brand-loyal segment are still characterised by (A43)-(A44), implying that demand for the original drug is still given by (A45). On the other hand, total demand for the follow-on drug is now given by

$$q_f = \lambda(1 - x_{b2}) + (1 - \lambda)\left(1 - x_g'\right). \tag{A51}$$

Profit-maximisation yields $p_o^{GM}(\lambda)$ equal to (A47) and

$$p_f^{GM}(\lambda) = \frac{(1-\lambda)\tau + 2(\lambda v + \beta c)}{2(1+\lambda)\beta}, \tag{A52}$$

which implies

$$x_{b1}^{GM} = \frac{v - \beta c}{2\tau}, \quad x_{b2}^{GM} = \frac{(3+\lambda)\tau - 2(v - \beta c)}{2(1+\lambda)\tau}, \quad x_g'^{GM} = \frac{(3+\lambda)\tau + 2\lambda(v - \beta c)}{4(1+\lambda)\tau} \tag{A53}$$

and

$$\pi^{GM}(\lambda) = \frac{2\lambda(1+3\lambda)(v-\beta c)^2 + (1-\lambda)(4\lambda(v-\beta c) + (1-\lambda)\tau)\tau}{8(1+\lambda)\beta\tau}. \tag{A54}$$

This is the profit-maximising solution if two conditions hold. First, the brand-loyal segment must remain partially covered, which requires

$$x_{b1}^{GM} < x_{b2}^{GM} \Leftrightarrow v < \beta c + \tau. \tag{A55}$$

Second, the marginal patient in the generic substitution segment must have a therapeutic benefit that is at least as high as the prescribing physician's perceived treatment cost, which requires

$$v - \beta c - \tau x_g'^{GM} \geq 0 \Leftrightarrow v \geq \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau. \tag{A56}$$

Thus, the profit-maximising prices are given by (A47) and (A52) for $v \in \left(\beta c + \frac{3+\lambda}{2(2+\lambda)}\tau, \beta c + \tau\right)$.

However, for $v \in \left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$, the producer's pricing of the follow-on drug is constrained by the need to provide the marginal patient in the generic substitution segment with a perceived treatment cost that does not exceed the therapeutic benefit. Within this parameter set, the optimal price of the follow-on drug is therefore given by

$$v - \beta p_f - \tau\left(1 - x_g'\right) = 0 \Leftrightarrow p_f^{GM}(\lambda) = \frac{2v-\tau}{\beta} - c, \tag{A57}$$

while $p_o(\lambda)$ is still given by (A47). This yields

$$x_{b1}^{GM} = \frac{v-\beta c}{2\tau}, \quad x_{b2}^{GM} = x_g'^{GM} = \frac{v-\beta c}{\tau} \tag{A58}$$

and

$$\pi^{GM}(\lambda) = \frac{4(3(v-\beta c)-\tau)\tau - (8-\lambda)(v-\beta c)^2}{4\beta\tau}. \tag{A59}$$

This solution requires that the brand-loyal segment remains partially covered, i.e., that $x_{b1}^{GM} < x_{b2}^{GM}$, which always holds for $v > \beta c$. Thus, the profit-maximising prices are given by (A47) and (A57) for $v \in \left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$.

For $v > \beta c + \tau$, the above analysis shows that the profit-maximising solution must necessarily imply that both segments are fully covered. This means that the marginal patient in the brand-

loyal segment is characterised by

$$x_b' = \frac{1}{2} + \frac{\beta\left(p_f - p_o\right)}{2\tau}, \tag{A60}$$

while the marginal patient in the generic substitution segment is still characterised by (A40). Total demand for the two drugs are then given by

$$q_o = \lambda x_b' \tag{A61}$$

and

$$q_f = \lambda\left(1 - x_b'\right) + (1-\lambda)\left(1 - x_g'\right). \tag{A62}$$

Profit maximisation yields

$$p_o\left(\lambda\right) = c + \frac{\tau}{(1-\lambda)\,\beta} \tag{A63}$$

and

$$p_f\left(\lambda\right) = c + \frac{(1+\lambda)\,\tau}{2\,(1-\lambda)\,\beta}, \tag{A64}$$

which implies

$$x_b'^{GM} = \frac{1}{4}, \quad x_g'^{GM} = \frac{3-\lambda}{4\,(1-\lambda)} \tag{A65}$$

and

$$\pi^{GM} = \frac{(1+3\lambda)\,\tau}{8\,(1-\lambda)\,\beta}. \tag{A66}$$

This solution requires that the marginal patients in both segments have a therapeutic benefit that is at least as high as the perceived treatment costs, which implies

$$v - \beta p_o - \tau x_b'^{GM} \geq 0 \Leftrightarrow v \geq \beta c + \frac{(5-\lambda)\,\tau}{4\,(1-\lambda)} \tag{A67}$$

and

$$v - \beta p_f - \tau\left(1 - x_g'^{GM}\right) \geq 0 \Leftrightarrow v \geq \beta c + \frac{(3-\lambda)\,\tau}{4\,(1-\lambda)}. \tag{A68}$$

Since $\frac{(5-\lambda)\tau}{4(1-\lambda)} < \tau < \frac{(3-\lambda)\tau}{4(1-\lambda)}$ for $\lambda < \frac{1}{3}$, it follows that the condition in (A68) always holds within the parameter set under consideration, i.e., $v > \beta c + \tau$, whereas the condition in (A67) does not hold if $v \in \left(\beta c + \tau, \beta c + \frac{(5-\lambda)\tau}{4(1-\lambda)}\right)$.

In the latter parameter set, where $v \in \left(\beta c + \tau, \beta c + \frac{(5-\lambda)\tau}{4(1-\lambda)}\right)$, the producer's pricing of the

original drug is constrained by the need to provide the marginal patient in the brand-loyal segment with a perceived treatment cost that does not exceed the therapeutic benefit, which implies that the original drug is priced according to

$$v - \beta p_o - \tau\left(x_b'\right) = 0 \Leftrightarrow p_o = \frac{2v - \tau}{\beta} - p_f. \tag{A69}$$

Profit-maximisation with respect to p_f, under the condition that p_o is set according to (A69), then yields

$$p_f^{GM}(\lambda) = \frac{2\beta(1-\lambda)c + 8\lambda v + (1-5\lambda)\tau}{2(1+3\lambda)\beta}, \tag{A70}$$

which in turn implies

$$p_o^{GM}(\lambda) = \frac{4(1+\lambda)v - 2(1-\lambda)\beta c - (3+\lambda)\tau}{2(1+3\lambda)\beta}. \tag{A71}$$

In this solution,

$$x_b'^{GM} = \frac{((3+\lambda)\tau - 2(1-\lambda)(v - c\beta))}{2(1+3\lambda)\tau}, \quad x_g'^{GM} = \frac{((3+\lambda)\tau + 8\lambda(v-\beta c))}{4(1+3\lambda)\tau} \tag{A72}$$

and

$$\pi^{GM}(\lambda) = \frac{\left(8\lambda(5-\lambda)(v-\beta c) - \left(18\lambda - \lambda^2 - 1\right)\tau\right)\tau - 16(1-\lambda)\lambda(v-\beta c)^2}{8(1+3\lambda)\beta\tau}. \tag{A73}$$

Thus, the profit-maximising prices are given by (A70)-(A71) if $v \in \left(\beta c + \tau, \beta c + \frac{(5-\lambda)\tau}{4(1-\lambda)}\right)$ and by (A63)-(A64) if $v > \beta c + \frac{(5-\lambda)\tau}{4(1-\lambda)}$. Notice also that, in both of these solutions, $x_b'^{GM} \in (0,1)$ and $x_g'^{GM} \in (0,1)$ if $\lambda < \frac{1}{3}$.

Appendix B: Proofs

Proof of Lemma 1

A comparison of (2) and (5) reveals that there are three different parameter regimes to consider (notice that $v^M > v^{MM}$). If $v \leq v^{MM}$, the market is only partially covered in the first period, regardless of whether the follow-on drug is launched or not. In this case, it is obvious that the introduction of the follow-on drug has no effect on the price of the original drug, since there is no *de facto* substitution between the two drug versions. However, if $v^{MM} < v \leq v^M$, the effect

a follow-on drug launch on the price of the original drug is given by

$$p_o^{MM} - p_o^M = \frac{v + w - 2(\tau + \beta c)}{4\beta} > 0 \text{ for } v \in \left(v^{MM}, v^M\right). \tag{B1}$$

And finally, if $v > v^M$, this price effect is given by

$$p_o^{MM} - p_o^M = \frac{w - v + 2\tau}{4\beta} > 0 \text{ for } v \in \left(v^M, \overline{v}^{MM}\right), \tag{B2}$$

where $\overline{v}^{MM}$ is the upper bound on v above which the follow-on drug cannot profitably survive in the market.

Proof of Proposition 1

Notice first that, if $v > \overline{v}^{GM}$, the follow-on drug will be driven out of the market by competition from generic copies of the original drug if launched in the second period. Moreover, necessary conditions for positive sales of the original drug and of the follow-on drug are $v > \beta c$ and $w > \beta c$, respectively. Thus, whether or not to delay the launch of the follow-on drug to the second period is a relevant decision only for parameter values belonging to the set Φ, defined by

$$\Phi := \left\{(w, v, \tau, \beta, c) \mid \beta c < v < \overline{v}^{GM} \text{ and } w > \beta c\right\}. \tag{B3}$$

The proof of Proposition 1 is then conducted by partitioning Φ into 45 non-intersecting subsets that are characterised in (w, v)-space by using eight different threshold values of v, defined by (3), (7), (9), (10), (14), (15), (17), and (18), and by determining the sign of $\Delta\pi$ in each of these subsets.

Regime (i): Suppose that $w \in \left(\beta c, \beta c + \frac{1}{2}\tau\right)$. In this case, it is straightforward to verify that

$$\overline{v}^{MM} > v^M > v^{MM} > \overline{v}^{GM} > v_2^{GM} > v_1^{GM} > \beta c > \underline{v}^{MM} > \underline{v}^{GM}, \tag{B4}$$

which implies that there are three relevant ranges of v.

Case 1: If $\beta c < v < v_1^{GM}$, the relevant profit expressions are given by

$$\pi^M = \frac{(v - \beta c)^2}{4\beta\tau}, \tag{B5}$$

$$\pi^{MM} = \frac{v^2 + w^2}{4\tau\beta} - \frac{(v + w - \beta c)\, c}{2\tau} \tag{B6}$$

and

$$\pi^{GM} = \frac{(w - \beta c)^2}{4\beta\tau}, \tag{B7}$$

which implies that the profit gain of delaying the launch of the follow-on drug is given by $\Delta\pi = 0$.

Case 2: If $v_1^{GM} < v < v_2^{GM}$, the follow-on drug faces competition from generic versions of the original drug if launched in the second period, so the second-period profits are in this case given by

$$\pi^{GM} = \frac{(v + w - \tau - 2c\beta)\left[\tau - (v - c\beta)\right]}{\beta\tau}, \tag{B8}$$

whereas π^M and π^{MM} are still given by (B5) and (B6), respectively. The profit gain of delaying the launch of the follow-on drug is now given by

$$\Delta\pi = -\frac{(2v + w - 2\tau - 3c\beta)^2}{4\beta\tau} < 0. \tag{B9}$$

Case 3: If $v_2^{GM} < v < \overline{v}^{GM}$, the second-period profits in case of delayed launch change to

$$\pi^{GM} = \frac{(w - v + \tau)^2}{8\beta\tau}, \tag{B10}$$

while the other profit expressions remain the same. The profit gain of delayed launch is now given by

$$\Delta\pi = -\frac{\phi}{8\beta\tau}, \tag{B11}$$

where

$$\phi := w^2 - v^2 + 2wv - 2\left(2w - c\beta\right)\beta c - \left(2\left(w - v\right) + \tau\right)\tau. \tag{B12}$$

By using the fact that ϕ is monotonically increasing in v over the interval $\left(v_2^{GM}, \overline{v}^{GM}\right)$, and that $\lim_{v \to v_2^{GM}} \phi > 0$, we can verify that $\phi > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v_2^{GM}, \overline{v}^{GM}\right)$.

Regime (ii): Suppose that $w \in \left(\beta c + \frac{\tau}{2}, \beta c + \tau\right)$. In this regime, the ranking of v-thresholds is given by

$$\overline{v}^{MM} > v^M > \overline{v}^{GM} > v^{MM} > v_2^{GM} > v_1^{GM} > \beta c > \underline{v}^{MM} > \underline{v}^{GM}, \tag{B13}$$

which implies that there are now four relevant intervals of v.

Case 1: If $\beta c < v < v_1^{GM}$, the analysis is identical to Case 1 in Regime (i), with the conclusion that $\Delta\pi < 0$.

Case 2: If $v_1^{GM} < v < v_2^{GM}$, the analysis is identical to Case 2 in Regime (i), with the conclusion that $\Delta\pi < 0$.

Case 3: If $v_2^{GM} < v < v^{MM}$, the profit gain of delayed launch is given by (B9), with the only difference that this expression is defined for a range of v with a lower upper bound (i.e., $v^{MM} < \overline{v}^{GM}$). However, since $\phi > 0$ for all $v \in \left(v_2^{GM}, \overline{v}^{GM}\right)$, it follows that $\phi > 0$ also for $v \in \left(v_2^{GM}, v^{MM}\right)$, implying that $\Delta\pi < 0$.

Case 4: If $v^{MM} < v < \overline{v}^{GM}$, the first-period profits if the follow-on drug is launched immediately are given by

$$\pi^{MM} = \frac{w + v - \tau}{2\beta} + \frac{(w - v)^2}{8\tau\beta} - c, \tag{B14}$$

while π^M and π^{GM} are given by (B5) and (B10), respectively. The profit gain of delayed launch is now given by

$$\Delta\pi = -\frac{\rho}{8\beta\tau}, \tag{B15}$$

where

$$\rho := (6v + 2w - 5\tau)\,\tau - 2\,(v - 2c\beta)\,v - 2\,(4\tau + c\beta)\,\beta c. \tag{B16}$$

Since $\partial^2\rho/\partial v^2 < 0$, ρ reaches a minimum value at either the lower or the upper bound of v. It is straightforward to verify that $\lim_{v\to v^{MM}} \rho > 0$ and $\lim_{v\to\overline{v}^{GM}} \rho$ for $w \in \left(\beta c + \frac{\tau}{2}, \beta c + \tau\right)$, implying that $\rho > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v^{MM}, \overline{v}^{GM}\right)$.

Regime (iii): Suppose that $w \in \left(\beta c + \tau, \beta c + \frac{3}{2}\tau\right)$, which implies the following ranking of v-thresholds:

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > v^{MM} > v_2^{GM} > v_1^{GM} > \beta c > \underline{v}^{MM} > \underline{v}^{GM}. \tag{B17}$$

There are now five relevant intervals of v.

Case 1: If $\beta c < v < v_1^{GM}$, the analysis is identical to Case 1 in Regime (i), with the conclusion that $\Delta\pi < 0$.

Case 2: If $v_1^{GM} < v < v_2^{GM}$, the analysis is identical to Case 2 in Regime (i), with the conclusion that $\Delta\pi < 0$.

Case 3: If $v_2^{GM} < v < v^{MM}$, the analysis is identical to Case 3 in Regime (ii), with the conclusion that $\Delta\pi < 0$.

Case 4: If $v^{MM} < v < v^M$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B14) and (B10), respectively, so the profitability of launch delay is given by (B15). We know that ρ

in (B15) is concave in v. It can also be verified that $\lim_{v\to v^M}\rho > 0$ and $\lim_{v\to v^{MM}}\rho > 0$ for $w \in \left(\beta c+\tau, \beta c+\frac{3}{2}\tau\right)$, which implies that $\rho > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v^{MM}, v^M\right)$.

Case 5: If $v^M < v < \overline{v}^{GM}$, the expressions for π^{MM} and π^{GM} are given by (B18) and (B12), respectively, while first-period profits when the launch of the follow-on drug is delayed are given by

$$\pi^M = \frac{v-\tau}{\beta} - c. \tag{B18}$$

The profit gain of launch delay is then given by

$$\Delta\pi = -\frac{(2(w-v)+3\tau)}{8\beta} < 0 \text{ if } v < \overline{v}^{GM}. \tag{B19}$$

Regime (iv): Suppose that $w \in \left(\beta c+\frac{3}{2}\tau, \beta c+2\tau\right)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > v_2^{GM} > v^{MM} > v_1^{GM} > \beta c > \underline{v}^{MM} > \underline{v}^{GM}, \tag{B20}$$

which implies that there are five relevant intervals of v.

Case 1: If $\beta c < v < v_1^{GM}$, the analysis is identical to that of Case 1 in Regime (i), with the conclusion that $\Delta\pi < 0$.

Case 2: If $v_1^{GM} < v < v^{MM}$, the profit gain of launch delay is the same as for Case 2 in Regime (i) and thus given by (B11), which is negative.

Case 3: If $v^{MM} < v < v_2^{GM}$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B14) and (B8), respectively. The profit gain of launch delay is then given by

$$\Delta\pi = -\frac{\mu}{8\beta\tau}, \tag{B21}$$

where

$$\mu := 6vw + 7v^2 + w^2 - 4(3v + w - \tau - 4\beta c)\tau - 2(10v + 4w - 7c\beta)\beta c. \tag{B22}$$

It is relatively straightforward to verify that μ is monotonically increasing in v for $w \in \left(\beta c+\frac{3}{2}\tau, \beta c+2\tau\right)$, and that $\lim_{v\to v^{MM}}\mu > 0$, implying that $\mu > 0$, and thus $\Delta\pi < 0$, for $v \in \left(v^{MM}, v_2^{GM}\right)$.

Case 4: If $v_2^{GM} < v < v^M$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B14) and (B10), respectively, so the profitability of launch delay is given by (B15), whose sign is given by the opposite of the sign of ρ, which we know is concave in v. It is easily verified

that $\lim_{v\to v_2^{GM}} \rho > 0$ and $\lim_{v\to v^M} \rho > 0$, implying that $\rho > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v_2^{GM}, v^M\right)$.

Case 5: If $v^M < v < \overline{v}^{GM}$, the analysis is identical to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (v): Suppose that $w \in \left(\beta c + 2\tau, \beta c + \frac{9}{4}\tau\right)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > v_2^{GM} > \underline{v}^{MM} > \beta c > v_1^{GM} > v^{MM} > \underline{v}^{GM}. \tag{B23}$$

There are thus four relevant intervals of v.

Case 1: If $\beta c < v < \underline{v}^{MM}$, the follow-on drug replaces the original drug for all patients (with a fully covered market) if launched in the first period, which means that first-period profits in this case is given by

$$\pi^{MM} = \frac{w-\tau}{\beta} - c, \tag{B24}$$

while π^M and π^{GM} are given by (B5) and (B8), respectively. The profit gain of launch delay is then given by

$$\Delta\pi = -\frac{(v - c\beta)\,(3v + 4w - 8\tau - 7\beta c)}{4\beta\tau}. \tag{B25}$$

The sign of this expression depends on the sign of the second factor in the numerator, which we temporarily denote by $(\cdot)$, and which is monotonically increasing in both v and w. It is easily verified that $\lim_{v\to\beta c, w\to\beta c+2\tau}(\cdot) = 0$, implying that $(\cdot) > 0$, and thus $\Delta\pi < 0$, for $v \in \left(\beta c, \underline{v}^{MM}\right)$ and $w \in \left(\beta c + 2\tau, \beta c + \frac{9}{4}\tau\right)$.

Case 2: If $\underline{v}^{MM} < v < v_2^{GM}$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B14) and (B8), respectively, which means that the profit gain of launch delay is given by (B21), whose sign is given by the opposite of the sign of μ, defined by (B22). Since μ is monotonically increasing in v, and since $\lim_{v\to\underline{v}^{MM}} \mu > 0$, it follows that $\mu > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(\underline{v}^{MM}, v_2^{GM}\right)$.

Case 3: If $v_2^{GM} < v < v^M$, the analysis is similar to Case 4 in Regime (iv), with the conclusion that $\Delta\pi < 0$.

Case 4: If $v^M < v < \overline{v}^{GM}$, the analysis is similar to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (vi): Suppose that $w \in \left(\beta c + \frac{9}{4}\tau, \beta c + \frac{5}{2}\tau\right)$, which implies that the ranking of

v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > \underline{v}^{MM} > v_2^{GM} > \beta c > v_1^{GM} > v^{MM} > \underline{v}^{GM}. \tag{B26}$$

There are thus four relevant ranges of v.

Case 1: If $\beta c < v < v_2^{GM}$, the profit gain of launch delay is given by (B25), whose sign is negative if $3v + 4w - 8\tau - 7\beta c > 0$. The only relevant difference from the analysis of Case 1 in Regime (v) is that the lower bound of w is now higher, which only reinforces the conclusion that $3v + 4w - 8\tau - 7\beta c > 0$ and thus $\Delta\pi < 0$, for the entire relevant parameter range.

Case 2: If $v_2^{GM} < v < \underline{v}^{MM}$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B24) and (B10), respectively, which implies that the profit gain of launch delay is given by

$$\Delta\pi = -\frac{\eta}{8\beta\tau}, \tag{B27}$$

where

$$\eta := (2v + 6w - 9\tau)\,\tau - 2c\beta\,(4\tau - 2v + c\beta) - 3v^2 + 2vw - w^2. \tag{B28}$$

It is easily verified that η is monotonically increasing in v over the interval $\left(v_2^{GM}, \underline{v}^{MM}\right)$, and it is also straightforward to check that $\lim_{v \to v_2^{GM}} \eta > 0$ for $w \in \left(\beta c + \frac{9}{4}\tau, \beta c + \frac{5}{2}\tau\right)$, which implies that $\eta > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v_2^{GM}, \underline{v}^{MM}\right)$.

Case 3: If $\underline{v}^{MM} < v < v^M$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B14) and (B10), respectively, so the profit gain of launch delay is given by (B15), whose sign is given by the opposite of the sign of ρ, defined by (B16). We know that ρ is concave in v, and it is easily verified that $\lim_{v \to \underline{v}^{MM}} \rho > 0$ and $\lim_{v \to v^M} \rho > 0$ for $w \in \left(\beta c + \frac{9}{4}\tau, \beta c + \frac{5}{2}\tau\right)$, implying that $\rho > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(\underline{v}^{MM}, v^M\right)$.

Case 4: If $v^M < v < \overline{v}^{GM}$, the analysis is similar to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (vii): Suppose that $w \in \left(\beta c + \frac{5}{2}\tau, \beta c + \frac{8}{3}\tau\right)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > \underline{v}^{MM} > v_2^{GM} > \beta c > v_1^{GM} > \underline{v}^{GM} > v^{MM}. \tag{B29}$$

There are once more four relevant ranges of v.

Case 1: If $\beta c < v < v_2^{GM}$, the analysis is similar to Case 1 in Regime (vi), with the only

relevant difference that the lower bound of w is now higher, which only reinforces the conclusion that $\Delta\pi < 0$ for the entire relevant parameter range.

Case 2: If $v_2^{GM} < v < \underline{v}^{MM}$, the analysis is similar to Case 2 in Regime (vi). It is easily verified that η is monotonically increasing in v over the interval $\left(v_2^{GM}, \underline{v}^{MM}\right)$ and that $\lim_{v \to v_2^{GM}} \eta > 0$ also for $w \in \left(\beta c + \frac{5}{2}\tau, \beta c + \frac{8}{3}\tau\right)$, which implies that $\eta > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(v_2^{GM}, \underline{v}^{MM}\right)$.

Case 3: If $\underline{v}^{MM} < v < v^M$, the analysis is similar to Case 3 in Regime (vi). It is easily verified that $\lim_{v \to \underline{v}^{MM}} \rho > 0$ and $\lim_{v \to v^M} \rho > 0$ also for for $w \in \left(\beta c + \frac{5}{2}\tau, \beta c + \frac{8}{3}\tau\right)$, implying that $\rho > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(\underline{v}^{MM}, v^M\right)$.

Case 4: If $v^M < v < \overline{v}^{GM}$, the analysis is similar to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (viii): Suppose that $w \in \left(\beta c + \frac{8}{3}\tau, \beta c + 3\tau\right)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > \underline{v}^{MM} > v_2^{GM} > \beta c > \underline{v}^{GM} > v_1^{GM} > v^{MM}. \tag{B30}$$

There are once more four relevant intervals of v.

Case 1: If $\beta c < v < v_2^{GM}$, the analysis is similar to Case 1 in Regime (vi), with the only relevant difference that the lower bound of w is now higher, which only reinforces the conclusion that $\Delta\pi < 0$ for the entire relevant parameter range.

Case 2: If $v_2^{GM} < v < \underline{v}^{MM}$, the analysis is similar to Case 2 in Regime (vi), with the conclusion that $\Delta\pi < 0$ for the relevant parameter range.

Case 3: If $\underline{v}^{MM} < v < v^M$, the analysis is similar to Case 3 in Regime (vi), with the conclusion that $\Delta\pi < 0$ for the relevant parameter range.

Case 4: If $v^M < v < \overline{v}^{GM}$, the analysis is similar to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (ix): Suppose that $w \in (\beta c + 3\tau, \beta c + 4\tau)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > v^M > \underline{v}^{MM} > \underline{v}^{GM} > \beta c > v_2^{GM} > v_1^{GM} > v^{MM}. \tag{B31}$$

There are four relevant intervals of v.

Case 1: If $\beta c < v < \underline{v}^{GM}$, the follow-on drug not only replaces the original drug for all

patients if launched in the first period, but it also fully outcompetes the generic versions of the original drugs if launched in the second period. In this case, the second-period profits in case of launch delay is given by

$$\pi^{GM} = \frac{w - v - \tau}{\beta}, \tag{B32}$$

while π^M and π^{MM} are given by (B5) and (B24), respectively. The profit gain of launch delay is in this case given by

$$\Delta\pi = -\frac{(v - c\beta)(\beta c + 4\tau - v)}{4\beta\tau}. \tag{B33}$$

The sign of this expression depends on the sign of $\beta c+4\tau-v$, which is monotonically decreasing in v. It is straightforward to verify that $\lim_{v\to\underline{v}^{GM}}(\beta c + 4\tau - v) > 0$, implying that $\beta c+4\tau-v > 0$, and thus $\Delta\pi < 0$, if for all $v \in \left(\beta c, \underline{v}^{GM}\right)$.

Case 2: If $\underline{v}^{GM} < v < \underline{v}^{MM}$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B24) and (B10), respectively, which means that the profit gain of launch delay is given by (B27), whose sign is given by the opposite of the sign of η. Since $\partial^2\eta/\partial v^2 < 0$, η reaches a minimum on the interval $\left(\underline{v}^{GM}, \underline{v}^{MM}\right)$ at either the lower or the upper bound. It is relatively straightforward to verify that $\lim_{v\to\underline{v}^{GM}}\eta > 0$ and $\lim_{v\to\underline{v}^{MM}}\eta$ for $w \in (\beta c + 3\tau, \beta c + 4\tau)$, implying that $\eta > 0$, and thus $\Delta\pi < 0$, for all $v \in \left(\underline{v}^{GM}, \underline{v}^{MM}\right)$.

Case 3: If $\underline{v}^{MM} < v < v^M$, the analysis is similar to Case 3 in Regime (vi), with the conclusion that $\Delta\pi < 0$ for the relevant parameter range.

Case 4: Case 4: If $v^M < v < \overline{v}^{GM}$, the analysis is similar to Case 5 in Regime (iii), with the conclusion that $\Delta\pi < 0$.

Regime (x): Suppose that $w \in (\beta c + 4\tau, \beta c + 5\tau)$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM} > \overline{v}^{GM} > \underline{v}^{MM} > v^M > \underline{v}^{GM} > \beta c > v_2^{GM} > v_1^{GM} > v^{MM}. \tag{B34}$$

In this regime there are four relevant intervals of v.

Case 1: If $\beta c < v < \underline{v}^{GM}$, the analysis is similar to Case 1 of Regime (ix), with the conclusion that $\Delta\pi < 0$ for the relevant parameter range.

Case 2: If $\underline{v}^{GM} < v < v^M$, the expressions for π^M, π^{MM} and π^{GM} are given by (B5), (B24) and (B10), respectively, which means that the profit gain of launch delay is given by (B27), whose sign is the opposite of the sign of η, which is defined by (B28). Since η is strictly concave

in v, η reaches its minimum value at either the lower or upper bound of v, and it is easily verified that $\lim_{v\to\underline{v}^{GM}}\eta>0$ and $\lim_{v\to v^{M}}\eta>0$ also for $w\in(\beta c+4\tau,\beta c+5\tau)$, implying that $\eta>0$, and thus $\Delta\pi<0$, for all $v\in\left(\underline{v}^{GM},v^{M}\right)$.

Case 3: If $v^{M}<v<\underline{v}^{MM}$, the expressions for π^{M}, π^{MM} and π^{GM} are given by (B18), (B24) and (B10), respectively, which implies that the profit gain of delayed launch is given by

$$\Delta\pi=-\frac{\varphi}{8\beta\tau}, \tag{B35}$$

where

$$\varphi:=\left(6\left(w-v\right)-\tau\right)\tau-\left(w-v\right)^{2}. \tag{B36}$$

It is easily verified that φ is monotonically decreasing in v over the interval $\left(v^{M},\underline{v}^{MM}\right)$, and that $\lim_{v\to\underline{v}^{MM}}\varphi>0$, implying that $\varphi>0$, and thus $\Delta\pi<0$, for all $v\in\left(v^{M},\underline{v}^{MM}\right)$.

Case 4: If $\underline{v}^{MM}<v<\overline{v}^{GM}$, the analysis is identical to Case 5 in Regime (iii), with the conclusion that $\Delta\pi<0$.

Regime (xi): Suppose that $w>\beta c+5\tau$, which implies that the ranking of v-thresholds is given by

$$\overline{v}^{MM}>\overline{v}^{GM}>\underline{v}^{MM}>\underline{v}^{GM}>v^{M}>\beta c>v_{2}^{GM}>v_{1}^{GM}>v^{MM}. \tag{B37}$$

In this final regime there are four relevant ranges of v.

Case 1: If $\beta c<v<v^{M}$, the profit gain of launch delay is given by (B33), which is clearly negative for $\beta c<v<v^{M}$.

Case 2: If $v^{M}<v<\underline{v}^{GM}$, the expressions for π^{M}, π^{MM} and π^{GM} are given by (B18), (B24) and (B32), respectively, which implies that the profit gain of launch delay is given by

$$\Delta\pi=-\frac{\tau}{\beta}<0. \tag{B38}$$

Case 3: If $\underline{v}^{GM}<v<\underline{v}^{MM}$, the analysis is similar to Case 3 of Regime (x), with the conclusion that $\Delta\pi<0$.

Case 4: If $\underline{v}^{MM}<v<\overline{v}^{GM}$, the analysis is identical to Case 5 in Regime (iii), with the conclusion that $\Delta\pi<0$.

Proof of Proposition 2

From the analysis in Appendix A, we know that, if β is sufficiently close to one and $w = v > v^M$, drug pricing is constrained in market structure MM but not in the other market structures. Thus, the profits in M and GM are given by, respectively,

$$\pi^M = \frac{v - \tau}{\beta} - c \tag{B39}$$

and

$$\pi^{GM} = \frac{\tau}{8\beta}. \tag{B40}$$

In market structure MM, the pricing of the two drugs is constrained by the requirement that inclusion of the follow-on drug does not reduce the health plan's surplus. If the follow-on drug is not included, the health plan's surplus, H^M, is given by (A33). In case the follow-on drug is included, the health plan's surplus, as a function of drug prices, is given by

$$H^{MM}(p_o, p_f) = \int_0^{\widehat{x}} (v - p_o - \tau s)\, ds + \int_{\widehat{x}}^1 (v - p_f - \tau(1 - s))\, ds, \tag{B41}$$

where $\widehat{x}$ is is given by (A8) for $w = v$. The health plan will include the follow-on drug in the first period if the drug prices are such that

$$\Delta H_1(p_o, p_f) := H^{MM}(p_o, p_f) - H^M = \frac{(4(v - \tau) + \beta(\tau - 2(p_o + p_f)))\tau + (2 - \beta)\beta^2 (p_o - p_f)^2}{4\beta\tau} \geq 0. \tag{B42}$$

Since $w = v$, both the producer's profits and the health plan's surplus are maximised for $p_o = p_f$. In the set of drug prices for which (B42) holds, the profit-maximising pair of prices are therefore given by

$$p_o^{MM} = p_f^{MM} = \frac{v}{\beta} - \frac{(4 - \beta)\tau}{4\beta}, \tag{B43}$$

yielding a profit of

$$\pi^{MM} = \frac{v}{\beta} - \frac{(4 - \beta)\tau}{4\beta} - c. \tag{B44}$$

The profit gain of launch delay is then given by

$$\Delta\pi = \pi^M + \pi^{GM} - \pi^{MM} = -\frac{(2\beta - 1)\tau}{8\beta} < 0. \tag{B45}$$

Thus, launch delay is not profitable even if drug pricing is constrained in case of immediate launch.

Proof of Proposition 3

If $w = v$, equilibrium drug prices in market structure MM and GM under free pricing are found by evaluating (5)-(6) and (13) for $w = v$, respectively, yielding

$$p_o^{MM} = p_f^{MM} = \begin{cases} \frac{v+c\beta}{2\beta} & if \quad v \leq \tau + \beta c \\ \frac{2v-\tau}{2\beta} & if \quad v > \tau + \beta c \end{cases} \tag{B46}$$

and

$$p_f^{GM} = \begin{cases} \frac{1}{2\beta}(v + c\beta) & if \quad \beta c < v \leq \beta c + \frac{2}{3}\tau \\ \frac{2v-\tau}{\beta} - c & if \quad \beta c + \frac{2}{3}\tau < v \leq \beta c + \frac{3}{4}\tau \\ c + \frac{\tau}{2\beta} & if \quad v > \beta c + \frac{3}{4}\tau \end{cases}, \tag{B47}$$

while p_o^M is still given by (2). This implies that there are five relevant threshold levels of v, given by

$$\beta c + 2\tau > \beta c + \tau > \beta c + \frac{3}{4}\tau > \beta c + \frac{2}{3}\tau > \beta c. \tag{B48}$$

The proof of Proposition 3 relies on an evaluation of the profit gain of launch delay within each of the intervals of v defined by these thresholds, and under each possible scenario in which the price cap binds in at least one of the three possible market structures.

(i) Suppose that the maximum therapeutic benefit is given by $v \in \left(\beta c, \beta c + \frac{2}{3}\tau\right)$. In this case, it follows from (2) and (B46)-(B47) that

$$p_o^M = p_o^{MM} = p_f^{MM} = p_f^{GM} = \frac{v + \beta c}{2\beta}. \tag{B49}$$

Thus, if the price cap binds, it binds under all three market structures. In case of a binding price cap, i.e., $\overline{p} < \frac{v+\beta c}{2\beta}$, it follows from (22)-(24) that the ranking of relevant $\overline{p}$-thresholds is given by

$$\frac{v + \beta c}{2\beta} > c > \frac{2v - \tau}{2\beta} > \frac{2v - \tau - c\beta}{\beta} > \frac{v - \tau}{\beta} \text{ if } v \in \left(\beta c, \beta c + \frac{1}{2}\tau\right) \tag{B50}$$

and

$$\frac{v + \beta c}{2\beta} > \frac{2v - \tau - c\beta}{\beta} > \frac{2v - \tau}{2\beta} > c > \frac{v - \tau}{\beta} \text{ if } v \in \left(\beta c + \frac{1}{2}\tau, \beta c + \frac{2}{3}\tau\right), \tag{B51}$$

which implies that there are two relevant sub-intervals of v. Suppose first that $v \in \left(\beta c, \beta c + \frac{1}{2}\tau\right)$, which from (B50) implies that the only relevant interval of $\overline{p}$ is given by $\overline{p} \in \left(c, \frac{v+\beta c}{2\beta}\right)$. In this case, profits in each of the three possible market structures are given by

$$\pi^{M}(\overline{p}) = (\overline{p} - c)\left(\frac{v - \beta\overline{p}}{\tau}\right), \tag{B52}$$

$$\pi^{MM}(\overline{p}) = (\overline{p} - c)\left(\frac{2(v - \beta\overline{p})}{\tau}\right) \tag{B53}$$

and

$$\pi^{GM}(\overline{p}) = (\overline{p} - c)\left(\frac{v - \beta\overline{p}}{\tau}\right), \tag{B54}$$

which implies that the profit gain of launch delay is $\Delta\pi = 0$.

Suppose next that $v \in \left(\beta c + \frac{1}{2}\tau, \beta c + \frac{2}{3}\tau\right)$, which from (B51) implies that there are three relevant sub-intervals of $\overline{p}$. If $\overline{p} \in \left(c, \frac{2v-\tau}{2\beta}\right)$, then the profits in market structure MM are given by

$$\pi^{MM}(\overline{p}) = \overline{p} - c, \tag{B55}$$

and profits in market structure GM are given by

$$\pi^{GM}(\overline{p}) = (\overline{p} - c)\left(\frac{\tau - \beta(\overline{p} - c)}{2\tau}\right), \tag{B56}$$

while $\pi^{M}(\overline{p})$ is still given by (B52). This means that the profit gain of launch delay is given by

$$\Delta\pi(\overline{p}) = (\overline{p} - c)\frac{2v - \beta(3\overline{p} - c) - \tau}{2\tau} > (<) \text{ if } \overline{p} < (>)\,\widehat{p}, \tag{B57}$$

where

$$\widehat{p} := \frac{2v + c\beta - \tau}{3\beta}. \tag{B58}$$

It is easily verified that $\widehat{p} \in \left(c, \frac{2v-\tau}{2\beta}\right)$ for $v \in \left(\beta c + \frac{1}{2}\tau, \beta c + \frac{2}{3}\tau\right)$. Thus:

$$\Delta\pi(\overline{p}) > 0 \text{ if } v \in \left(\beta c + \frac{1}{2}\tau, \beta c + \frac{2}{3}\tau\right) \text{ and } \overline{p} \in (c, \widehat{p}). \tag{B59}$$

If instead $v \in \left(\beta c + \frac{1}{2}\tau, \beta c + \frac{2}{3}\tau\right)$ and $\overline{p} \in \left(\frac{2v-\tau}{2\beta}, \frac{2v-\tau-c\beta}{\beta}\right)$, then $\pi^{M}(\overline{p})$, $\pi^{MM}(\overline{p})$ and $\pi^{GM}(\overline{p})$ are given by (B52), (B53) and (B56), respectively, which implies that the profit gain

of launch delay is given by

$$\Delta\pi\left(\overline{p}\right)=\left(\overline{p}-c\right)\frac{\tau-2v+\beta\left(c+\overline{p}\right)}{2\tau}<0\text{ for }v\in\left(\beta c+\frac{1}{2}\tau,\beta c+\frac{2}{3}\tau\right)\text{ and }\overline{p}>c. \tag{B60}$$

Finally, if $v\in\left(\beta c+\frac{1}{2}\tau,\beta c+\frac{2}{3}\tau\right)$ and $\overline{p}\in\left(\frac{2v-\tau-c\beta}{\beta},\frac{1}{2\beta}\left(v+c\beta\right)\right)$, the three relevant profit expressions are given by (B52)-(B54) and $\Delta\pi\left(\overline{p}\right)=0$.

(ii) Suppose that the maximum therapeutic benefit is given by $v\in\left(\beta c+\frac{2}{3}\tau,\beta c+\frac{3}{4}\tau\right)$. In this case, it follows from (2) and (B46)-(B47) that

$$p_f^{GM}=\frac{2v-\tau}{\beta}-c>p_o^M=p_o^{MM}=p_f^{MM}=\frac{v+\beta c}{2\beta}. \tag{B61}$$

(a) Suppose that $\overline{p}\in\left(\frac{v+\beta c}{2\beta},\frac{2v-\tau}{\beta}-c\right)$, which implies that the price cap binds in market structure GM, but not in market structures M or MM. This reduces the profitability of GM without affecting the profitability of M and MM, so launch delay cannot possibly be profitable given the result in Proposition 1.

(b) Suppose that $\overline{p}\in\left(c,\frac{v+\beta c}{2\beta}\right)$, which implies that the price cap binds in all three market structures. It follows from (22)-(24) that the ranking of relevant $\overline{p}$-thresholds is given by

$$\frac{2v-\tau-c\beta}{\beta}>\frac{v+\beta c}{2\beta}>\frac{2v-\tau}{2\beta}>c>\frac{v-\tau}{\beta}, \tag{B62}$$

which in turn implies that there are two relevant subintervals of $\overline{p}$. If $\overline{p}\in\left(c,\frac{2v-\tau}{2\beta}\right)$, profits in market structure M, MM and GM are given by (B52), (B55) and (B56), respectively, which implies that the profit gain of launch delay is given by (B57), and is therefore positive if $\overline{p}<\widehat{p}$ as defined by (B58). It is easily verified that $\widehat{p}\in\left(c,\frac{2v-\tau}{2\beta}\right)$ for $v\in\left(\beta c+\frac{2}{3}\tau,\beta c+\frac{3}{4}\tau\right)$. Thus:

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v\in\left(\beta c+\frac{2}{3}\tau,\beta c+\frac{3}{4}\tau\right)\text{ and }\overline{p}\in\left(c,\widehat{p}\right). \tag{B63}$$

On the other hand, if $\overline{p}\in\left(\frac{2v-\tau}{2\beta},\frac{v+\beta c}{2\beta}\right)$, profits in market structure MM is instead given by (B53), which implies that the profit gain of launch delay is given by (B60), which is negative for $v\in\left(\beta c+\frac{2}{3}\tau,\beta c+\frac{3}{4}\tau\right)$ and $\overline{p}>c$.

(iii) Suppose that the maximum therapeutic benefit is given by $v\in\left(\beta c+\frac{3}{4}\tau,\beta c+\tau\right)$. In

this case, it follows from (2) and (B46)-(B47) that

$$p_f^{GM} = c + \frac{\tau}{2\beta} > p_o^M = p_o^{MM} = p_f^{MM} = \frac{v + \beta c}{2\beta}. \tag{B64}$$

(a) Suppose that $\overline{p} \in \left(\frac{v+\beta c}{2\beta}, c + \frac{\tau}{2\beta}\right)$, which means that the price cap binds in market structure GM, but not in M or MM. Once more, this reduces the profitability of GM without affecting the profitability of M and MM, so launch delay cannot possibly be profitable given the result in Proposition 1.

(b) Suppose that $\overline{p} \in \left(c, \frac{v+\beta c}{2\beta}\right)$, which means that the price cap binds in all three possible market structures. From (22)-(24), the ranking of $\overline{p}$-thresholds is given by (B62), which means that there are two relevant sub-intervals of $\overline{p}$. If $\overline{p} \in \left(c, \frac{2v-\tau}{2\beta}\right)$, the profits in market structure M, MM and GM are given by (B52), (B55) and (B56), respectively, which implies that the profit gain of launch delay is given by (B57), and is therefore positive if $\overline{p} < \widehat{p}$ as defined by (B58). It is easily verified that $\widehat{p} \in \left(c, \frac{2v-\tau}{2\beta}\right)$ for $v \in \left(\beta c + \frac{3}{4}\tau, \beta c + \tau\right)$. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \frac{3}{4}\tau, \beta c + \tau\right) \text{ and } \overline{p} \in \left(c, \widehat{p}\right). \tag{B65}$$

On the other hand, if $\overline{p} \in \left(\frac{2v-\tau}{2\beta}, \frac{v+\beta c}{2\beta}\right)$, profits in market structure MM is instead given by (B53), which implies that the profit gain of launch delay is given by (B60), which is negative for $v \in \left(\beta c + \frac{3}{4}\tau, \beta c + \tau\right)$ and $\overline{p} > c$.

(iv) Suppose that the maximum therapeutic benefit is given by $v \in (\beta c + \tau, \beta c + 2\tau)$. In this case, it follows from (2) and (B46)-(B47) that

$$p_o^{MM} = p_f^{MM} = \frac{2v - \tau}{2\beta} > p_o^M = \frac{v + \beta c}{2\beta} > p_f^{GM} = c + \frac{\tau}{2\beta}. \tag{B66}$$

(a) Suppose that $\overline{p} \in \left(\frac{v+\beta c}{2\beta}, \frac{2v-\tau}{2\beta}\right)$, which implies that the price cap binds in market structure MM but not in the other two market structures. In this case, the profits in market structure MM is given by (B55), while profits in market structures M and GM, using (4) and (16), are given by, respectively,

$$\pi^M = \frac{(v - \beta c)^2}{4\beta\tau} \tag{B67}$$

and

$$\pi^{GM} = \frac{\tau}{8\beta}. \tag{B68}$$

The profit gain of launch delay is therefore given by

$$\Delta\pi\left(\overline{p}\right)=\frac{2c^{2}\beta^{2}+2v\left(v-2c\beta\right)-8\tau\beta\left(\overline{p}-c\right)+\tau^{2}}{8\beta\tau}>(<)\,0\text{ if }\overline{p}<(>)\,\widehat{p}',\tag{B69}$$

where

$$\widehat{p}':=\frac{2c\beta\left(4\tau+c\beta\right)+2v\left(v-2c\beta\right)+\tau^{2}}{8\tau\beta}.\tag{B70}$$

It is relatively easy to verify that $\widehat{p}'<\frac{2v-\tau}{2\beta}$ for all $v\in(\beta c+\tau,\beta c+2\tau)$, while $\widehat{p}'>\frac{v+\beta c}{2\beta}$ if v is sufficiently close to the upper bound $\beta c+2\tau$. Thus:

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v\in(\underline{v},\beta c+2\tau)\text{ and }\overline{p}\in\left(\frac{v+\beta c}{2\beta},\widehat{p}'\right),\tag{B71}$$

where $\overline{v}$ is implicitly defined by $2c^{2}\beta^{2}+2\underline{v}\left(\underline{v}-2c\beta\right)-8\tau\beta\left(\overline{p}-c\right)+\tau^{2}=0$.

(b) Suppose that $\overline{p}\in\left(c+\frac{\tau}{2\beta},\frac{v+\beta c}{2\beta}\right)$, which implies that the price cap binds in M and MM but not in GM. It follows from (22)-(24) that the ranking of relevant $\overline{p}$-thresholds is given by

$$\frac{2v-\tau}{2\beta}>\frac{v+\beta c}{2\beta}>c+\frac{\tau}{2\beta}>\frac{v-\tau}{\beta}\quad\text{if}\quad v\in\left(\beta c+\tau,\beta c+\frac{3}{2}\tau\right)\tag{B72}$$

and

$$\frac{2v-\tau}{2\beta}>\frac{v+\beta c}{2\beta}>\frac{v-\tau}{\beta}>c+\frac{\tau}{2\beta}\quad\text{if}\quad v\in\left(\beta c+\frac{3}{2}\tau,\beta c+2\tau\right),\tag{B73}$$

which implies that there are two relevant sub-intervals of v. Suppose first that $v\in\left(\beta c+\tau,\beta c+\frac{3}{2}\tau\right)$. In this case, profits in market structures M and MM are given by (B52) and (B55), while profits in market structure GM are given by (B68). The profit gain of launch delay is therefore given by

$$\Delta\pi\left(\overline{p}\right)=\frac{8\beta\left(v-\tau-\beta\overline{p}\right)\left(\overline{p}-c\right)+\tau^{2}}{8\beta\tau}.\tag{B74}$$

The sign of this expression depends on the sign of the numerator, which is monotonically decreasing in $\overline{p}$ for $v\in\left(\beta c+\tau,\beta c+\frac{3}{2}\tau\right)$ and $\overline{p}\in\left(c+\frac{\tau}{2\beta},\frac{v+\beta c}{2\beta}\right)$, and also monotonically increasing in v. It is relatively straightforward to verify that $\lim_{v\to\beta c+\tau}\Delta\pi<0$ for all $\overline{p}\in\left(c+\frac{\tau}{2\beta},\frac{v+\beta c}{2\beta}\right)$, while $\lim_{v\to\beta c+\frac{3}{2}\tau}\Delta\pi>(<)\,0$ if $\overline{p}$ is sufficiently close to the lower (upper) bound. By continuity and monotonicity, it follows that

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v\in\left(\underline{v}',\beta c+\frac{3}{2}\tau\right)\text{ and }\overline{p}\in\left(c+\frac{\tau}{2\beta},\widehat{p}''\right),\tag{B75}$$

where

$$\underline{v}' := \tau + \beta\overline{p} - \frac{\tau^2}{8\beta\left(\overline{p} - c\right)} \tag{B76}$$

and where $\widehat{p}''$ is implicitly defined by

$$8\beta\left(v - \tau - \beta\widehat{p}''\right)\left(\widehat{p}'' - c\right) + \tau^2 = 0. \tag{B77}$$

Suppose next that $v \in \left(\beta c + \frac{3}{2}\tau, \beta c + 2\tau\right)$, which implies that there are two relevant sub-intervals of $\overline{p}$. If $\overline{p} \in \left(c + \frac{\tau}{2\beta}, \frac{v-\tau}{\beta}\right)$, the profits are the same in market structure M as in market structure MM, and given by (B55), while the profits in GM are given by (B68). This clearly implies that launch delay is profitable, since launching the follow-on drug alongside the original drug in the first period yields zero additional profits. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \frac{3}{2}\tau, \beta c + 2\tau\right) \text{ and } \overline{p} \in \left(c + \frac{\tau}{2\beta}, \frac{v-\tau}{\beta}\right). \tag{B78}$$

On the other hand, if $\overline{p} \in \left(\frac{v-\tau}{\beta}, \frac{v+\beta c}{2\beta}\right)$, the profits in M and MM are given by (B52) and (B55), while profits in market structure GM are given by (B68), so the profit gain of launch delay is given by (B74), whose sign depend on the sign of the numerator, which is monotonically increasing in v and monotonically decreasing in $\overline{p}$. It is relatively straightforward to show that $\Delta\pi > 0$ at the the lower bound of v if $\overline{p}$ is sufficiently close to the lower bound, while $\Delta\pi > 0$ at the upper bound of v for all $\overline{p} \in \left(\frac{v-\tau}{\beta}, \frac{v+\beta c}{2\beta}\right)$. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \frac{3}{2}\tau, \beta c + 2\tau\right) \text{ and } \overline{p} \in \left(\frac{v-\tau}{\beta}, \min\left\{\widehat{p}'', \frac{v+\beta c}{2\beta}\right\}\right), \tag{B79}$$

where $\widehat{p}''$ is implicitly defined by (B77).

(c) Suppose that $\overline{p} \in \left(c, c + \frac{\tau}{2\beta}\right)$, which means that the price cap binds in all possible market structures, and that the ranking of $\overline{p}$-thresholds is given by

$$\frac{2v - \tau - c\beta}{\beta} > \frac{2v-\tau}{2\beta} > \frac{v+\beta c}{2\beta} > c + \frac{\tau}{2\beta} > \frac{v-\tau}{\beta} > c \text{ if } v \in \left(\beta c + \tau, \beta c + \frac{3}{2}\right) \tag{B80}$$

and

$$\frac{2v - \tau - c\beta}{\beta} > \frac{2v-\tau}{2\beta} > \frac{v+\beta c}{2\beta} > \frac{v-\tau}{\beta} > c + \frac{\tau}{2\beta} > c \text{ if } v \in \left(\beta c + \frac{3}{2}, \beta c + 2\tau\right), \tag{B81}$$

which implies that there are two relevant sub-intervals of v. Suppose first that $v \in \left(\beta c + \tau, \beta c + \frac{3}{2}\right)$, which implies in turn that there are two relevant sub-intervals of $\overline{p}$. If $\overline{p} \in \left(c, \frac{v-\tau}{\beta}\right)$, the profits are the same in market structure M as in market structure MM, and given by (B55), while the profits in GM are given by (B56). Once more, this clearly implies that launch delay is profitable, since launching the follow-on drug alongside the original drug in the first period yields zero additional profits. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \tau, \beta c + \frac{3}{2}\tau\right) \text{ and } \overline{p} \in \left(c, \frac{v-\tau}{\beta}\right). \tag{B82}$$

On the other hand, if $\overline{p} \in \left(\frac{v-\tau}{\beta}, c + \frac{\tau}{2\beta}\right)$, the profits in market structure M is instead given by (B52), which implies that the profit gain of launch delay is given by (B57), whose sign depends on whether the price cap is above or below the threshold level $\widehat{p}$ given by (B58). It is relatively straightforward to verify that $\widehat{p} > \frac{v-\tau}{\beta}$ for $v \in \left(\beta c + \tau, \beta c + \frac{3}{2}\right)$, while $\widehat{p} < (>)\, c + \frac{\tau}{2\beta}$ if $v < (>)\, \beta c + \frac{5}{4}\tau$. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \tau, \beta c + \frac{3}{2}\tau\right) \text{ and } \overline{p} \in \left(\frac{v-\tau}{\beta}, \min\left\{\widehat{p}, c + \frac{\tau}{2\beta}\right\}\right), \tag{B83}$$

where $\widehat{p}$ is given by (B58).

Suppose next that $v \in \left(\beta c + \frac{3}{2}, \beta c + 2\tau\right)$. In this case, profits are the same in market structure M as in market structure MM, and given by (B55), while the profits in GM are given by (B56). Again, this clearly implies that launch delay is profitable, since launching the follow-on drug alongside the original drug in the first period yields zero additional profits. Thus:

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \frac{3}{2}, \beta c + 2\tau\right) \text{ and } \overline{p} \in \left(c, c + \frac{\tau}{2\beta}\right). \tag{B84}$$

(v) Finally, suppose that the maximum therapeutic benefit is given by $v > \beta c + 2\tau$. In this case, it follows from (2) and (B46)-(B47) that

$$p_o^{MM} = p_f^{MM} = \frac{2v-\tau}{2\beta} > p_o^{M} = \frac{v-\tau}{\beta} > p_f^{GM} = c + \frac{\tau}{2\beta}. \tag{B85}$$

(a) Suppose that $\overline{p} \in \left(\frac{v-\tau}{\beta}, \frac{2v-\tau}{2\beta}\right)$, which implies that the price cap binds in MM but not in M and GM. In this case, the profits in MM are given by (B55), the profits in GM are given by (B68), while the profits in M are given by (4). The profit gain of launch delay is then given

by

$$\Delta\pi\left(\overline{p}\right)=\frac{8\left(v-\beta\overline{p}\right)-7\tau}{8\beta}>(<)\,0\text{ if }\overline{p}<(>)\,\frac{8v-7\tau}{8\beta}. \tag{B86}$$

It is easily verified that $\frac{v-\tau}{\beta}<\frac{8v-7\tau}{8\beta}<\frac{2v-\tau}{2\beta}$ for $v>\beta c+2\tau$. Thus:

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v>\beta c+2\tau\text{ and }\overline{p}\in\left(\frac{v-\tau}{\beta},\frac{8v-7\tau}{8\beta}\right). \tag{B87}$$

(b) Suppose that $\overline{p}\in\left(c+\frac{\tau}{2\beta},\frac{v-\tau}{\beta}\right)$, which implies that the price cap binds in M and MM but not in GM, and that the ranking of relevant $\overline{p}$-thresholds is given by

$$\frac{2v-\tau}{2\beta}>\frac{v-\tau}{\beta}>c+\frac{\tau}{2\beta}. \tag{B88}$$

In this case, the profits are the same in market structure M as in market structure MM, and given by (B55), while the profits in GM are given by (B68). This clearly implies that launch delay is profitable, since launching the follow-on drug alongside the original drug in the first period yields zero additional profits. Thus:

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v>\beta c+2\tau\text{ and }\overline{p}\in\left(c+\frac{\tau}{2\beta},\frac{v-\tau}{\beta}\right). \tag{B89}$$

(c) Suppose that $\overline{p}\in\left(c,c+\frac{\tau}{2\beta}\right)$, which implies that the price cap binds in all three possible market structures. Also in this case, the profits are the same in market structure M as in market structure MM, and given by (B55), while the profits in GM are now given by (B56). Yet again, this clearly implies that launch delay is profitable, since launching the follow-on drug alongside the original drug in the first period yields zero additional profits. Thus:

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v>\beta c+2\tau\text{ and }\overline{p}\in\left(c,c+\frac{\tau}{2\beta}\right) \tag{B90}$$

The proof is completed by noting that the combination of (B59), (B63) and (B65) yields

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v\in\left(\beta c+\frac{\tau}{2},\beta c+\frac{2}{3}\tau\right)\text{ and }\overline{p}\in\left(c,\widehat{p}\right), \tag{B91}$$

the combination of (B82) and (B83) yields

$$\Delta\pi\left(\overline{p}\right)>0\text{ if }v\in\left(\beta c+\tau,\beta c+\frac{3}{2}\tau\right)\text{ and }\overline{p}\in\left(c,\min\left\{\widehat{p},c+\frac{\tau}{2\beta}\right\}\right), \tag{B92}$$

the combination of (B78), (B79) and (B84) yields

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v \in \left(\beta c + \frac{3}{2}\tau, \beta c + 2\tau\right) \text{ and } \overline{p} \in \left(c, \min\left\{\widehat{p}'', \frac{v+\beta c}{2\beta}\right\}\right), \tag{B93}$$

and the combination of (B87), (B89) and (B90) yields

$$\Delta\pi\left(\overline{p}\right) > 0 \text{ if } v > \beta c + 2\tau \text{ and } \overline{p} \in \left(c, \frac{8v-7\tau}{8\beta}\right). \tag{B94}$$

It follows from (B91)-(B94) that, for every value of $v > \beta c + \frac{\tau}{2}$, there exists a threshold value of the price cap that is strictly higher than marginal cost, such that $\Delta\pi\left(\overline{p}\right) > 0$ if the price cap lies between marginal cost and this threshold value.

Proof of Proposition 4

If $w > v + 3\tau$, it follows from (11) and (19) that the equilibrium drug prices in market structure MM and GM under free pricing are given by $p_f^{MM} = \frac{w-\tau}{\beta}$ and $p_f^{GM} = c + \frac{w-v-\tau}{\beta}$, respectively, while from (2) we know that the profit-maximising in M is given by $p_o^M = \frac{v+\beta c}{2\beta}$ if $v < \beta c + 2\tau$ and by $p_o^M = \frac{v-\tau}{\beta}$ if $v > \beta c + 2\tau$.

(a) Consider first the case of $v \in (\beta c, \beta c + 2\tau)$, which implies that

$$p_f^{MM} = \frac{w-\tau}{\beta} > p_f^{GM} = c + \frac{w-v-\tau}{\beta} > p_o^M = \frac{v+\beta c}{2\beta}. \tag{B95}$$

(i) Suppose that $\overline{p} \in \left(c + \frac{w-v-\tau}{\beta}, \frac{w-\tau}{\beta}\right)$, implying that the price cap binds in MM but not in GM or M.

The profits in the three possible market structures are then given by

$$\pi^M = \frac{(v-\beta c)^2}{4\beta\tau}, \tag{B96}$$

$$\pi^{MM}\left(\overline{p}\right) = \overline{p} - c \tag{B97}$$

and

$$\pi^{GM} = \frac{w-v-\tau}{\beta}, \tag{B98}$$

and the profit gain of launch delay is given by

$$\Delta\pi\left(\overline{p}\right)=\frac{\left(v-\beta c\right)^{2}+4\tau\left(w-v-\tau-\beta\left(\overline{p}-c\right)\right)}{4\beta\tau}>\left(<\right)0\text{ if }\overline{p}<\left(>\right)\widehat{p}^{\prime\prime\prime},\tag{B99}$$

where

$$\widehat{p}^{\prime\prime\prime}:=c+\frac{w-v-\tau}{\beta}+\frac{\left(v-\beta c\right)^{2}}{4\tau\beta}.\tag{B100}$$

It is easily verified that $\widehat{p}^{\prime\prime\prime}\in\left(c+\frac{w-v-\tau}{\beta},\frac{w-\tau}{\beta}\right)$.

(ii) Suppose that $\overline{p}\in\left(\frac{v+\beta c}{2\beta},c+\frac{w-v-\tau}{\beta}\right)$, implying that the price cap binds in MM and GM but not in M. In this case, the profits in GM are given by

$$\pi^{GM}\left(\overline{p}\right)=\overline{p}-c,\tag{B101}$$

while the profits in M and MM are still given by (B96) and (B97). Since, $\pi^{MM}\left(\overline{p}\right)=\pi^{GM}\left(\overline{p}\right)$, launch delay is clearly profitable, since it gives the firm one period of additional profits equal to π^{M}.

(iii) Suppose that $\overline{p}\in\left(c,\frac{v+\beta c}{2\beta}\right)$, implying that the price cap binds in all three possible market structures. Once more, since $\pi^{MM}\left(\overline{p}\right)=\pi^{GM}\left(\overline{p}\right)$, launch delay is profitable.

(b) Consider next the case of $v\in\left(\beta c+2\tau,\beta c+4\tau\right)$. In this case, the price ranking (under free pricing) is given by

$$p_{f}^{MM}=\frac{w-\tau}{\beta}>p_{f}^{GM}=c+\frac{w-v-\tau}{\beta}>p_{o}^{M}=\frac{v+\beta c}{2\beta}.\tag{B102}$$

(i) Suppose that $\overline{p}\in\left(c+\frac{w-v-\tau}{\beta},\frac{w-\tau}{\beta}\right)$, implying that the price cap binds in MM but not in M and GM. In this case, the profits in MM and GM are given by (B97) and (B98), respectively, while the profits in M is now given by

$$\pi^{M}=\frac{v-\tau}{\beta}-c,\tag{B103}$$

which in turn means that the profit gain of launch delay is

$$\Delta\pi\left(\overline{p}\right)=\frac{w-2\tau}{\beta}-\overline{p}>\left(<\right)0\text{ if }\overline{p}<\left(>\right)\frac{w-2\tau}{\beta}\in\left(c+\frac{w-v-\tau}{\beta},\frac{w-\tau}{\beta}\right).\tag{B104}$$

(ii) Suppose that $\overline{p}\in\left(\frac{v+\beta c}{2\beta},c+\frac{w-v-\tau}{\beta}\right)$, implying that the price cap binds in MM and

GM but not in M. This means that the profits in GM change to (B101), implying once more that $\pi^{GM}(\overline{p}) = \pi^{MM}(\overline{p})$, which clearly makes launch delay profitable.

(iii) Suppose that $\overline{p} \in \left(c, \frac{v+\beta c}{2\beta}\right)$, implying that the price cap binds in all three possible market structures. Once more, $\pi^{GM}(\overline{p}) = \pi^{MM}(\overline{p})$ and launch delay is profitable.

(c) Finally, consider the case of $v > \beta c + 4\tau$, which implies that the price ranking (under free pricing) is given by

$$p_f^{MM} = \frac{w-\tau}{\beta} > p_o^M = \frac{v+\beta c}{2\beta} > p_f^{GM} = c + \frac{w-v-\tau}{\beta}. \tag{B105}$$

(i) Suppose that $\overline{p} \in \left(\frac{v+\beta c}{2\beta}, \frac{w-\tau}{\beta}\right)$, implying that the price cap binds in MM but not in M and GM. The analysis is the same as in case (b)-(i), and the profitability of launch delay is given by the condition in (B104).

(ii) Suppose that $\overline{p} \in \left(c + \frac{w-v-\tau}{\beta}, \frac{v+\beta c}{2\beta}\right)$, implying that the price cap binds in M and MM but not in GM. In this case, the profits in M is given by $\overline{p} - c$ and is therefore the same as in MM, implying that launch delay is profitable.

(iii) Suppose that $\overline{p} \in \left(c, c + \frac{w-v-\tau}{\beta}\right)$, implying that the price cap binds in all three possible market structures. In this case profits are equal to $\overline{p} - c$ in all three market structures and launch delay is therefore clearly profitable.

Proof of Proposition 6

Let $W^s := \pi^s + H^s$ be the total welfare in market structure s. Immediate launch of the follow-on drug is then welfare-optimal if

$$\Delta W := \left(W^{MM} - W^M\right) + \left(W^{GG} - W^{GM}\right) > 0, \tag{B106}$$

where GG denotes the market structure in which both the original and the follow-on drugs face direct generic competition, leading to prices equal to marginal cost for both drugs. It is immediately obvious that $W^{GG} > W^{GM}$ if $\beta = 1$. In this case, there are no distortion in prescription choices, neither at the intensive nor at the extensive margin in market structure GG, while too few patients are prescribed the follow-on drug in market structure GM because $p_f^{GM} > c$. By continuity, it must be true that $W^{GG} > W^{GM}$ also for values of β sufficiently close to one.

Given that β is sufficiently close to one, a sufficient condition for $\Delta W > 0$ is therefore

$W^{MM} > W^{M}$. Using (4) and (A33), total welfare in M is given by

$$W^{M} = \begin{cases} \frac{(v-\beta c)(3v-(4-\beta)c)}{8t} & if \quad v \leq v^{M} \\ v - c - \frac{t}{2} & if \quad v > v^{M} \end{cases}, \tag{B107}$$

whereas, using (8) and (A34), total welfare in MM is given by

$$W^{MM} = \begin{cases} \frac{3\left(v^2+w^2\right)-2c((2+\beta)(v+w)-\beta(4-\beta)c)}{8t} & if \quad v \leq v^{MM} \\ \frac{3(v-w)^2+4(2(v+w)-4c-t)t}{16t} & if \quad v > v^{MM} \end{cases}. \tag{B108}$$

Suppose first that $v < v^{MM}$. In this case we have

$$W^{MM} - W^{M} = \frac{(w - \beta c)\left(3w - (4-\beta)c\right)}{8t}, \tag{B109}$$

which is clearly positive for all $w > c$ if $\beta \to 1$. Suppose next that $v \in \left(v^{MM}, v^{M}\right)$, in which case we have

$$W^{MM} - W^{M} = \frac{\omega}{16t}, \tag{B110}$$

where

$$\omega := (8t + 3(w - v))(v + w) + 4((2+\beta)cv - (4c+t)t) - 6vw - 2(4-\beta)\beta c^2. \tag{B111}$$

The sign of (B110) is given by the sign of ω. From (B111) it is easily verified that $\partial\omega/\partial w > 0$ for $w \geq v$. Evaluating ω at $w = v$ yields

$$\lim_{w \to v} \omega = 4(4(v-c) - t)t - 2(3v - 2(2+\beta)c)v - 2(4-\beta)\beta c^2. \tag{B112}$$

It is easily verified that $\partial^2 \left(\lim_{w\to v} \omega\right)/\partial v^2 < 0$, implying that $\lim_{w\to v} \omega$ reaches its lowest value on the interval $\left(v^{M}, v^{MM}\right)$ either at the lower or at the upper bound of v. Evaluating the expression at both bounds yields

$$\lim_{v \to v^{M}} \left(\lim_{w \to v} \omega\right) = (3t + 4(\beta - 1)c)t, \tag{B113}$$

which is positive if β is sufficiently close to one, and

$$\lim_{v \to v^{MM}} \left(\lim_{w \to v} \omega \right) = 2t^2 > 0. \tag{B114}$$

If β is sufficiently close to one, $\lim_{w \to v} \omega > 0$ for all $v \in \left(v^{MM}, v^{M}\right)$, which in turn means that $\omega > 0$, and thus $W^{MM} > W^{M}$, for all $w \geq v$.

Finally, suppose that $v > v^{MM}$, in which case we have

$$W^{MM} - W^{M} = \frac{\left(2t + 3\left(w - v\right)\right)\left(2t + w - v\right)}{16t}, \tag{B115}$$

which is clearly positive for all $w \geq v$.

Proof of Proposition 7

Given the assumption $\lambda < \frac{1}{3}$, the ranking of all the relevant v-values is given by

$$\beta c < \beta c + \frac{2}{3}\tau < \beta c + \frac{3+\lambda}{2\left(2+\lambda\right)}\tau < \beta c + \tau < \beta c + \frac{\left(5-\lambda\right)}{4\left(1-\lambda\right)}\tau < \beta c + 2\tau. \tag{B116}$$

We will consider each interval of v in turn.

(i) Suppose that $v \in \left(\beta c, \beta c + \frac{2}{3}\tau\right)$. In this case, the profits in the four different market structures are given by

$$\pi^{M}\left(\lambda\right) = \frac{\left(v - \beta c\right)^2}{4\beta\tau}, \tag{B117}$$

$$\pi^{MM}\left(\lambda\right) = \frac{\left(v - \beta c\right)^2}{2\beta\tau}, \tag{B118}$$

$$\pi^{GM}\left(\lambda\right) = \frac{\left(1+\lambda\right)\left(v - \beta c\right)^2}{4\beta\tau} \tag{B119}$$

and

$$\pi^{GG}\left(\lambda\right) = \frac{\lambda\left(v - \beta c\right)^2}{2\beta\tau}, \tag{B120}$$

which implies that the profit gain of launch delay is given by

$$\Delta\pi\left(\lambda\right) = -\frac{\lambda\left(v - \beta c\right)^2}{4\beta\tau} < 0. \tag{B121}$$

(ii) Suppose that $v \in \left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$. In this case, the profit in market structure

GM changes to

$$\pi^{GM}(\lambda) = \frac{4(3(v-\beta c)-\tau)\tau - (8-\lambda)(v-\beta c)^2}{4\beta\tau}, \tag{B122}$$

which implies that the profit gain of launch delay is given by

$$\Delta\pi(\lambda) = -\frac{\kappa}{4\beta\tau}, \tag{B123}$$

where $\kappa := (9+\lambda)(v-\beta c)^2 + 4(\tau - 3(v-\beta c))\tau$. It is easily verified that κ is monotonically increasing in v over the interval $\left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$ and that $\lim_{v\to\beta c+\frac{2}{3}\tau}\kappa > 0$, implying that $\Delta\pi(\lambda) < 0$ for all $v \in \left(\beta c + \frac{2}{3}\tau, \beta c + \frac{3+\lambda}{2(2+\lambda)}\tau\right)$.

(iii) Suppose that $v \in \left(\beta c + \frac{3+\lambda}{2(2+\lambda)}\tau, \beta c + \tau\right)$. In this case, the profit in market structure GM changes again to

$$\pi^{GM}(\lambda) = \frac{2\lambda(1+3\lambda)(v-\beta c)^2 + (1-\lambda)(4\lambda(v-\beta c)+(1-\lambda)\tau)\tau}{8(1+\lambda)\beta\tau}, \tag{B124}$$

which implies that the profit gain of launch delay is

$$\Delta\pi(\lambda) = -\frac{\psi}{8(1+\lambda)\beta\tau}, \tag{B125}$$

where $\psi := 2\left(1+2\lambda-\lambda^2\right)(v-\beta c)^2 - (1-\lambda)((1-\lambda)\tau + 4\lambda(v-\beta c))\tau$. It is easily verified that ψ is monotonically increasing in v over the interval $\left(\beta c + \frac{3+\lambda}{2(2+\lambda)}\tau, \beta c + \tau\right)$ and that $\lim_{v\to\beta c+\frac{3+\lambda}{2(2+\lambda)}\tau}\psi > 0$, implying that $\Delta\pi(\lambda) < 0$ for all $v \in \left(\beta c + \frac{3+\lambda}{2(2+\lambda)}\tau, \beta c + \tau\right)$.

(iv) Suppose that $v \in \left(\beta c + \tau, \beta c + \frac{(5-\lambda)}{4(1-\lambda)}\tau\right)$. In this case, the profit in GM changes again to

$$\pi^{GM}(\lambda) = \frac{\left(8\lambda(5-\lambda)(v-\beta c) - \left(18\lambda-\lambda^2-1\right)\tau\right)\tau - 16(1-\lambda)\lambda(v-\beta c)^2}{8(1+3\lambda)\beta\tau}, \tag{B126}$$

while the profits in MM and GG change to

$$\pi^{MM}(\lambda) = \frac{2v-\tau}{2\beta} - c \tag{B127}$$

and

$$\pi^{GG}(\lambda) = \lambda\left(\frac{2v-\tau}{2\beta} - c\right), \tag{B128}$$

respectively. The profit gain of launch delay is then given by

$$\Delta\pi(\lambda) = -\frac{\xi}{8(1+3\lambda)\beta\tau}, \tag{B129}$$

where $\xi := \left(8\left(1-\lambda+4\lambda^2\right)(v-\beta c) - \left(5-2\lambda+13\lambda^2\right)\tau\right)\tau - 2\left(1-5\lambda+8\lambda^2\right)(v-\beta c)^2$. It is easily verified that ξ is strictly concave in v, implying that ξ reaches its lowest value on the interval $\left(\beta c+\tau, \beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau\right)$ at either the lower or the upper bound. It is relatively straightforward to verify that $\lim_{v\to\beta c+\tau}\xi > 0$ and $\lim_{v\to\beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau}\xi > 0$ if $\lambda < \frac{1}{3}$, implying that $\Delta\pi(\lambda) < 0$ for all $v \in \left(\beta c+\tau, \beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau\right)$.

(v) Suppose that $v \in \left(\beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau, 2\tau\right)$. In this case, the profit in market structure GM changes once more to

$$\pi^{GM}(\lambda) = \frac{(1+3\lambda)\tau}{8(1-\lambda)\beta}, \tag{B130}$$

so the profit gain of launch delay becomes

$$\Delta\pi(\lambda) = -\frac{\theta}{8(1-\lambda)\beta\tau}, \tag{B131}$$

where $\theta := \left(8(1-\lambda)(1+\lambda)(v-\beta c) - \left(5+3\lambda-4\lambda^2\right)\tau\right)\tau - 2(1-\lambda)(v-\beta c)^2$. It is easily confirmed that θ is monotonically increasing in v over the interval $\left(\beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau, 2\tau\right)$ and that $\lim_{v\to\beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau}\theta > 0$ if $\lambda < \frac{1}{3}$, implying that $\Delta\pi(\lambda) < 0$ for all $v \in \left(\beta c+\frac{(5-\lambda)}{4(1-\lambda)}\tau, 2\tau\right)$.

(vi) Finally, suppose that $v > \beta c+2\tau$. In this case, the profit in market structure M changes to

$$\pi^{M}(\lambda) = \frac{v-\tau}{\beta} - c, \tag{B132}$$

which implies that the profit gain of launch delay is given by

$$\Delta\pi(\lambda) = -\frac{\sigma}{8(1-\lambda)\beta}, \tag{B133}$$

where $\sigma := 8\lambda(1-\lambda)(v-\beta c) + \left(3-11\lambda+4\lambda^2\right)\tau$. It is immediately evident that σ is monotonically increasing in v, and it can easily be verified that $\lim_{v\to\beta c+2\tau}\sigma > 0$ if $\lambda < \frac{1}{3}$, implying that $\Delta\pi(\lambda) < 0$ also for all $v > \beta c+2\tau$.

References

[1] Bala, R., & Bhardwaj, P. (2010). Detailing vs. direct-to-consumer advertising in the prescription pharmaceutical industry. *Management Science*, 56(1), 148–160.

[2] Bardey, D., & Bourgeon, J-M. (2011). Health care network formation and policyholders' welfare. *The B.E. Journal of Economic Analysis & Policy*, 11(2).

[3] Bardey, D., B. Jullien & Lozachmeur, J.-M.. (2016). Health insurance and diversity of treatment. *Journal of Health Economics*, 47, 50–63.

[4] Branstetter, L., Chatterjee, C., & Higgins, M.J. (2016). Regulation and welfare: evidence from paragraph IV generic entry in the pharmaceutical industry. *The RAND Journal of Economics*, 47(4), 857–890.

[5] Brekke, K.R., Dalen, D.M., & Straume, O.R. (2022). Paying for pharmaceuticals: uniform pricing versus two-part tariffs. J*ournal of Health Economics*, 83, 102613.

[6] Brekke, K.R., Dalen, D.M., & Straume, O.R. (2023). The price of cost-effectiveness thresholds under therapeutic competition in pharmaceutical markets. *Journal of Health Economics*, 90, 102778.

[7] Brekke, K.R., Dalen, D.M., & Straume, O.R. (2024). Competing with precision: incentives for developing predictive biomarker tests. *The Scandinavian Journal of Economics*, 126(1), 60–97.

[8] Carlton, D.W., Flyer, F.A., & Shefi, Y. (2016). Does the FTC's theory of product hopping promote competition? *Journal of Competition Law & Economics*, 12(3), 495–506.

[9] Carrier, M.A. (2010). A real-world analysis of pharmaceutical settlements: the missing dimension of product hopping. *Florida Law Review*, 62, 1009.

[10] Carrier, M.A., & Shadowen, S.D. (2016). Product hopping: a new framework. *Notre Dame Law Review*, 92(1), 167–230.

[11] Cockburn, I.M., Lanjouw, J.O., & Schankerman, M. (2016). Patents and global diffusion of new drugs. *American Economic Review,* 106(1), 136–164.

[12] Danzon, P.M., & Chao, L.-W. (2000). Does regulation drive out competition in pharmaceutical markets? *The Journal of Law and Economics*, 43(2), 311–358.

[13] Dickson, S., Gabriel, N., & Hernandez, I. (2023). Changes in net prices and spending for pharmaceuticals after the introduction of new therapeutic competition, 2011–19. *Health Affairs*, 42(8), 1062–1070.

[14] Ellison, S.F., Cockburn, I., Griliches, A. & Hausman., J. (1997). Characteristics of demand for pharmaceutical products: an examination of four cephalosporins. *RAND Journal of Economics*, 28, 426–446.

[15] Ellison, G., & Ellison, S.F. (2011). Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. *American Economic Journal: Microeconomics*, 3(1), 1–36.

[16] Engelberg, A.B., Kesselheim, A.S., & Avorn, J. (2009). Balancing innovation, access, and profits—market exclusivity for biologics. *New England Journal of Medicine*, 361(20), 1917–1919.

[17] Fowler, A.C. (2017). Pharmaceutical line extensions in the united states: A primer on definitions and incentives. *Value of Medical Research White Paper Series, NBER-IFS.*

[18] Fowler, A. C. (2019). Hurry up or wait? Strategic delay in the introduction of pharmaceutical line extensions. Working Paper.

[19] González, P., Macho-Stadler, I., & Pérez-Castrillo, D. (2016). Private versus social incentives for pharmaceutical innovation. *Journal of Health Economics*, 50, 286–297.

[20] Hemphill, C.S., & Sampat, B.N. (2011). When do generics challenge drug patents? *Journal of Empirical Legal Studies*, 8(4), 613–649.

[21] Hemphill, C.S., & Sampat, B.N. (2012). Evergreening, patent challenges, and effective market life in pharmaceuticals. *Journal of Health Economics*, 31(2), 327–339.

[22] Huckfeldt, P.J., & Knittel, C.R. (2011). Pharmaceutical use following generic entry: Paying less and buying less. *National Bureau of Economic Research*, Working Paper No. 17046.

[23] Huskamp, H.A., Busch, A.B., Domino, M.E., & Normand, S.-L.T. (2009). Antidepressant reformulations: who uses them, and what are the benefits? *Health Affairs*, 28(3), 734–745.

[24] Kyle, M.K. (2007). Pharmaceutical price controls and entry strategies. *The Review of Economics and Statistics,* 89(1), 88–99.

[25] Lakdawalla, D.N. (2018). Economics of the pharmaceutical industry. *Journal of Economic Literature*, 56, 397–449.

[26] Lipatov, V., Neven, D., & Siotis, G. (2021). Preempting the entry of near perfect substitutes. *Journal of Competition Law & Economics*, 17(1), 194–210.

[27] Miraldo, M. (2009). Reference pricing and firms' pricing strategies. *Journal of Health Economics*, 28, 176–197.

[28] Narasimhan, C., & Zhang, Z.J. (2000). Market entry strategy under firm heterogeneity and asymmetric payoffs. *Marketing Science*, 19(4), 313–327.

[29] Oi, W.Y. (1996). The welfare implications of invention. In: *The Economics of New Goods* (pp. 109-142). University of Chicago Press.

[30] Phadke, I. (2024). Rewarding incremental innovation: evidence from pharmaceutical line extensions. Working Paper. Last downloaded January 29, 2025 from https://iphadke.github.io/

[31] Price II, W.N. (2020). The cost of novelty. *Columbia Law Review*, 120, 769.

[32] Shadowen, S.D., Leffler, K.B., & Lukens, J.T. (2009). Anticompetitive product changes in the pharmaceutical industry. *Rutgers Law Journal*, 41, 1.

[33] Shapiro, B.T. (2016). Estimating the cost of strategic entry delay in pharmaceuticals: The case of Ambien CR. *Quantitative Marketing and Economics*, 14(3), 201–231.

[34] Stremersch, S., & Lemmens, A. (2009), Sales growth of new pharmaceuticals across the globe: The role of regulatory regimes. *Marketing Science,* 28(4), 690–708.

[35] Yin, N. (2023). Pharmaceuticals, incremental innovation and market exclusivity. *International Journal of Industrial Organization*, 87, 102922.