Sustainable Reimbursements: The Case for Two-Price Programs

Sylvain Chassang[c], Valentina Mantua[a,b], Erik Snowberg[d,e,f], Entela Xoxi[a], and Luca Pani[b,g,*]

[a]The Italian Medicines Agency (AIFA), [b]European Medicines Agency (EMA),
[c]New York University, [d]California Institute of Technology, [e]University of British Columbia,
[f]National Bureau of Economic Research, [g]University of Miami
[*]Corresponding Author: Luca Pani

Background / Importance:
High prices and the uncertainties around new pharmaceutical compounds are the greatest threat to the sustainability of the current pace of pharmaceutical innovation.

Methods:
The Economic theory of procurement contracting suggests two-price programs will substantially reduce the uncertainty to both payers and pharmaceutical companies. Tailoring these theories to pharmaceuticals suggests two different reimbursement programs: a *patient-based* program when there is uncertainty about how individual patients will respond to the compound, and a *treatment-based* program when there is uncertainty about the number of patients. Simulations are used to project the effects of adoption of these reimbursement programs.

Findings:
Two-price programs substantially reduce uncertainty for both payers and pharmaceutical companies. This is accomplished while still rewarding pharmaceutical companies that innovate and create value for patients.

Interpretation:
Adoption of two-price reimbursement programs will be an important tool in combating spiraling pharmaceutical costs and sustaining the current pace of innovation.

Funding:
This research received no outside funding.

Contributions:
All authors contributed to all parts of the design and execution of this research. The views expressed in this manuscript are personal and may not be understood or quoted as being made on behalf of or reflecting the position of EMA or AIFA or any of their committees or working parties. They are not intended to be an official and/or binding regulatory position.

Recent breakthroughs in the treatment of advance stage cancers such as melanoma and non-small cell lung cancer have completely changed the outlook for some patients with those diseases. In a sizeable share of cases these patients have been given an extended stay on what was formerly a death sentence. [1] However, the very success of these drugs presents large payers with financial challenges that threaten their sustainability. High prices set under the assumption of temporary treatment—examples ranging from $300K a year (nivolumab + imipilumab) to $1M per patient, per year (pembrolizumab) in the U.S. [1, 2], with similar numbers applying to the U.K.—become truly unsustainable when an acute condition is transformed into a chronic one, causing the patient population to grow exponentially [3].

Reimbursements to practitioners have been the focus of change, such as the attempt to move away from the fee-for-service model in the US. [4, 5, 6] Yet pharmaceutical reimbursements are almost uniformly done on a fee-per-dose basis. To make new therapies sustainable, a paradigm shift is needed: from reimbursement approaches which only address uncertainties in clinical outcomes, to ones that can also deal with uncertainties regarding the length of treatment, rates of patient survival, and the size of future patient populations. We argue that two-price reimbursement programs, with high initial prices and low continuation prices, reduce risk to both pharmaceutical companies and payers, and offer a sustainable, profitable way to guarantee patient access.

Key Sources of Uncertainty

The challenge of reimbursement for innovative treatments is the uncertainty over both their effects, and the size of the (future) patient base. [3] In most cases, treatments are approved on the basis of a relatively small benefit to soft endpoints. [7] However, in some cases treatments can have radical, lasting effects on a sizeable share of treated patients. For example, an unknown proportion of patients with a formerly quickly-fatal disease may go on living under treatment for an extended period of time, and this will not be well reflected in a surrogate endpoint such as proportion of patients experiencing progression-free survival at 6 or 12 months. Indeed, in the case of immune-checkpoint inhibitors, a substantial number of the patients from the original trials are still alive years later. If this occurs, treated populations can grow exponentially.

Changing demographics and medical practices also create significant uncertainty over the size of future patient populations. For example, the number of patients taking PCSK9 inhibitors to reduce LDL cholesterol is difficult to predict as it depends crucially on the number of patients who do not respond to statins, the number that develop statin intolerance, and the number who choose to switch because of differential side effects between the treatments. [8] Moreover, if treatment initiation is moved to before the symptomatic phase, as has been suggested for disease modifying drugs for neurodegenerative conditions, the treatable population expands massively and unpredictably. This further leads to significant uncertainty in even defining relevant outcomes, precluding price negotiations based on results.

Standard fee-per-dose reimbursement programs expose both pharmaceutical companies and payers to significant financial risks. If the treatment turns out to be unexpectedly effective, or more widely used than expected, the payer may have difficulty covering a rapidly expanding patient population. [2, 3] There is also uncertainty from the pharmaceutical company's point of view: the drug may be used much less than expected. This may occur because a competitor quickly gains approval, because the treatment turns out to be much less effective than trial results would predict (for example, Drotrecogin Alfa), or due to the detection of previously unknown side-effects (for example, Tetrazepam). [9, 10] These significant financial risks cause pharmaceutical companies to maximize profit on successful compounds, which limits payers' ability to provide access to the most innovative treatments. [3, 7]

Two-Price Reimbursement Programs

Two-price reimbursement programs reduce the payer's financial uncertainties by ensuring a low cost of continuing treatment, while simultaneously reducing a pharmaceutical company's uncertainties by guaranteeing the bulk of their profit up front. These programs are inspired by cost-plus and price-volume discount programs that have been extensively studied by economists. [11, 12, 13, 14]

These programs start with high prices for initial doses, and revert to lower prices after some time has passed. High initial prices guarantee that profits will be realized up front, in a predictable way. Low continuation prices limit the potential financial liability of payers, guaranteeing that payments do not grow exponentially, even if the treated population does. There will be little difference in cost for the payer if a patient lives two, five, or ten years while continuing treatment.

Patient-Based and Treatment-Based Programs

In the *patient-based program*, the payer pays a large initial price for the first dose *to that patient* at some point after treatment starts, with a much smaller continuation price being paid for doses *to that patient* thereafter. A lag between the start of treatment and the start of reimbursement can be used as a check against over-prescription. The simplest implementation verifies that the treatment is appropriate to a patient's condition; reimbursement will not occur if treatment is halted, or if the patient dies, during the initial lag. More sophisticated effectiveness checks are possible during this initial period, but rely on having an effective and trusted monitoring system, such as the certified registries developed by AIFA. [3, 15] The patient-based program resembles other suggestions for capitated pricing [16], but varies in some of the details, especially the initial lag in high prices that can be used to assess patient-level efficacy.

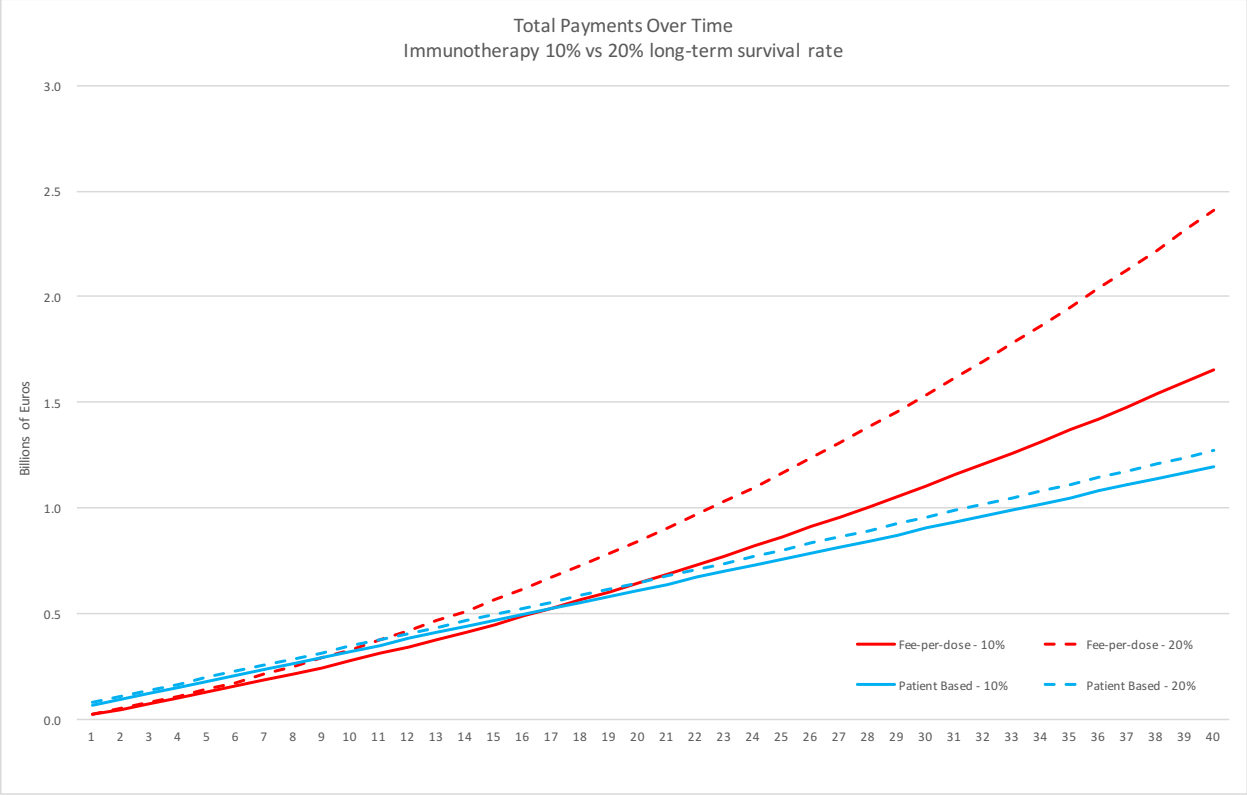
In *treatment-based programs*, the larger initial fee is paid on some number of initial doses covered by the payer, and the smaller continuation price is paid on all subsequent doses. Once again, the high initial fee may be delayed, in this case to give the payer time to understand the efficacy of treatment in their covered population. Moreover, the payments themselves could be negotiated to depend on outcomes in that population. If this requires too long a delay, initial payments could be made immediately, with claw-backs—or additional payments—indexed to pre-defined levels of efficacy. With this initial lag and high-payments indexed to efficacy, the treatment-based program resembles suggestions for value based pricing [1], although the lower continuation price would allow for greater access.

In both programs, there are three aspects to negotiate. Both programs require the negotiation of the high initial price and lower continuation price. In the patient-based program the lag between treatment start and payment of the initial price must also be negotiated; the analog in the treatment-based program is the number of doses that will be reimbursed at the higher initial price. These details will determine total payments, but the structure of the programs leads to fundamental differences in how uncertainties affect both the payer and pharmaceutical company. We explore these differences through examples, which leads to a general discussion of the properties of the two programs.

Examples

Two examples serve to illustrate how different reimbursement programs insure payers and pharmaceutical companies against different types of uncertainties.

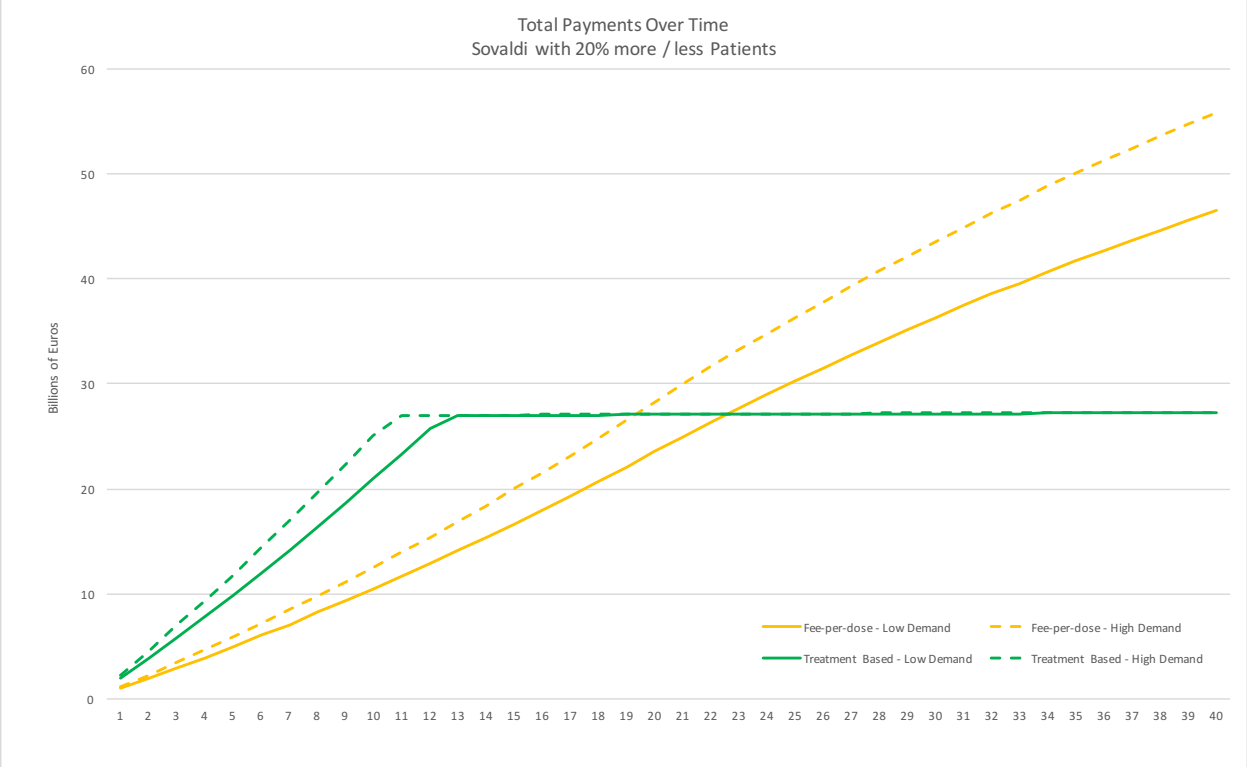
First, Figure 1 shows simulated total payments from fee-per-dose (using list prices) and a patient-based program for nivolumab + imipilumab as a treatment for metastatic melanoma. [17] This figure makes payment projections based on the patient populations and survival profile in Italy. The critical uncertainty is what percentage of patients will survive for a very long time on this treatment. In the simulations, two different values are used representing optimistic (20%) and pessimistic (10%) rates. The fee-per-dose program is based on the negotiated price of $25,000 in the US [1]. The initial and continuation prices in the patient-based program are set so spending and revenues under both programs is relatively equal by year 5—the 20th quarter. This choice focuses on the long term effects of uncertainty under the different programs.



Two patterns in the figure are worth noting. First, under either assumption about the long-term-survival rate, in the fee-per-dose program total spending grows exponentially, whereas the total spending in the patient-based program is roughly linear. This is due to the fact that in the patient-based program, payments in a given year are almost completely determined by the flow of

new patients, which is roughly constant across years. In contrast, under the fee-per-dose program payments in a given year depends on the stock of existing patients, which is growing. Second, in the patient-based program misestimating the percent of long-term survivors has very little effect on cumulative spending (5%), whereas a 10% error in the percentage of long-term survivors leads to a 45% difference in cumulative spending over 10 years in the fee-per-dose program. The patient-based program also provides some insurance to the company as initial profits are higher. However, there is still some uncertainty over the company's total profit and payer's total costs, depending on the effective life of the drug.

Our second example uses the Hepatitis C treatment Sovaldi (Sofosbuvir), using figures from the Italian experience. [15] Here, the critical uncertainty is in the number of patients. As new indications are approved, the patient population may grow unexpectedly. [18, 19] The existence of a "cure" for Hepatitis C may also change testing decisions, as well as risky behavior. [20] Thus, Figure 2 compares a fee-per-dose program with a treatment-based one for two projections of future patient populations. Both of these use current patient populations in Italy to project future demand. One takes this projection and lowers it by10%, and the other raises it by 10%.



The treatment-based program almost completely eliminates financial risk for both the payer and pharmaceutical company as profits and costs are both extremely predictable and quickly

realized. However, other risks exist for the payer that should lead to lower negotiated prices than in the patient-based program. If the treatment works less well than expected, there is little recourse for the payer through renegotiation. If a more effective treatment is introduced, the payer may stick with the existing treatment because of its much lower continuation price. Finally, it requires the payer to have access to a large amount of financing up front, although this may be mitigated by the pharmaceutical company agreeing to delayed payment. This, coupled with the reduction in risk to the pharmaceutical company may allow the payer to bargain for lower prices, although this may be offset by the fact that treatment-based programs eliminate the possibility of truly "blockbuster" financial performance.

Why Two-price Programs Work

The total cost of a drug using a fee-per-dose program is given by:

Total Cost = [Number of Patients x Number of Doses per Patient] x Price per Dose.

The sources of uncertainty discussed above affect one of the two factors in brackets above: either the number of doses that will be needed per patient (immunotherapies, due to unknown survival times), or the number of patients (PCSK9 inhibitors, Hepatitis C). The two-price programs automatically adjust the price per dose so that no matter how these uncertainties resolve, the total cost will be insulated.

Both the patient-based and treatment-based programs adjust the price per dose in similar ways: the initial price is low (or zero), followed by a period of high prices, and then a period of low prices. However, the timing of these prices differ based on whether the program is insuring against a large number of doses per patient, or a large number of patients (and thus doses), in the total population.

Considerations when Choosing a Two-Price Reimbursement Program

The optimal program depends on the key uncertainty being addressed, and other details. Patient based contracts can expose payers to more substantial financial liabilities, but they are fairly straightforward to implement. Treatment based contracts offer more insurance to both pharmaceutical companies and payers, but are trickier. All in all, we believe that while both payers and pharmaceutical companies familiarize themselves with these new programs, the more robust patient based contracts seem like the right first step.

Negotiated terms will depend on too many factors to enumerate here. However, we can give some general guidance on how to design these two-price programs to achieve specific goals, and mitigate specific uncertainties. The main considerations, and tools for dealing with them, are listed in the table.

For both programs, setting the lower continuation price involves a trade-off. If that price is set too low, it may cause the pharmaceutical company to underinvest in production, which could lead to shortages. If that price is set too high, there is little incentive for the pharmaceutical company to take common-sense cost-cutting measures. In addition, high continuation prices may create incentives for pharmaceutical companies to try to keep patients on a treatment for a longer period of time, or to treat patients unlikely to benefit from the treatment. The latter risk is particularly acute in the patient-based program, but can be mitigated through a longer delay before high reimbursement costs apply, or evaluation of effects through monitoring.

Consideration / Concern	**Design Solution**
Quality of Treatment	-Higher continuation price
Cost of production of treatment	-Lower continuation price
Over-prescription and over-treatment	-Lower continuation price -Later initial payment in patient-based program -Fewer and higher initial payments in treatment-based payments
Large uncertainty about efficacy of related compounds and / or for related conditions	-Patient-based program -Treatment-based program with quantity threshold proportional to treated population
Large uncertainty about efficacy of current treatment	-Delayed high cost repayments contingent on measured performance

Other concerns are program-specific. Patient-based programs may expose the payer to significant financial uncertainty when there is little clarity about upcoming changes to the treatment landscape for a given condition. The patient-based program may end up being quite costly if the current treatment has a long effective life because replacement treatments do not come to fruition. In this case, a treatment-based program may be preferred by the payer.

In treatment-based programs, there is an incentive for the pharmaceutical company to incrementally change treatments to benefit from renewed high initial payments. For example, the drug Kalydeco was initially approved only for use on patients with cystic fibrosis caused by a specific genetic mutation. [21] Later, Vertex moved to have the same drug approved for eight related mutations. [22] Had a treatment-based program been negotiated for the first approved use, it would not account for the additional approved conditions, and a second program would need to be negotiated. However, this can, and should, be anticipated and addressed, perhaps by the contract covering payments for any future (albeit likely unspecified) new indications for the treatment. If the pharmaceutical company refuses such a proposal, then it's a sign that they are in fact planning to seek such approval.

Another concern with treatment-based programs is that they may increase renegotiation by pharmaceutical companies, and limit the scope for renegotiation by payers. Indeed, once treatment is delivered at the low continuation price, pharmaceutical companies can renegotiate by claiming higher than anticipated costs of manufacturing. Conversely, payers cannot renegotiate for lower prices if the realized efficacy is lower than anticipated as they have paid most of the expected total amount.

While these concerns could potentially be mitigated, they would require regulators to have different tools. One way to address potential renegotiation by pharmaceutical companies would be to include a contractual provision granting payers a license to produce (with the low price playing the role of a royalty) treatment in the event that the pharmaceutical company is not able to deliver at the agreed-upon price. Payments indexed on realized efficacy, including potential clawbacks, could be included in the program. AIFA's monitoring infrastructure and success in clawing back reimbursements when treatments are less efficacious than anticipated shows this is feasible. But, if these concerns cannot be confidently addressed, patient-based programs are preferable to treatment-based programs.

References

[1] Saltz, LB. Perspectives on cost and value in cancer care. *JAMA Oncology* 2016; **2**: 19-21.

[2] Bach PB. Limits on Medicare's ability to control rising spending on cancer drugs, *N Engl J Med* 2009; **360**: 626-33.

[3] Pani L. Sustainable innovation: medicines and the challenges for the future of our national health service. Milano: Edra, 2016.

[4] Obama, B. United States health care reform: progress to date and next steps. *JAMA* 2016; **316**: 525-32.

[5] Steinbrook R. The end of fee-for-service medicine? Proposals for payment reform in Massachusetts. *N Engl J Med* 2009; **361**: 1036-38.

[6] Schroeder SA, Frist W. Phasing out fee-for-service payment. *N Engl J Med* 2013; **368**: 2029-32.

[7] Bach PB. Indication-specific pricing for cancer drugs. *JAMA* 2014; **312**: 1629-30.

[8] Nissen SE, Stroes E, Dent-Acosta RE, Rosenson RS, Lehman SJ, Sattar N, et al. Efficacy and tolerability of evolocumab vs ezetimibe in patients with muscle-related statin intolerance: the GAUSS-3 randomized clinical trial. *JAMA* 2016*;* **315**: 1580-90.

[9] Ranieri VM, Thompson BT, Barie PS, Dhainaut JF, Douglas IS, Finfer S, et al. Drotrecogin alfa (activated) in adults with septic shock. *N Engl J Med* 2012; **366**: 2055-64.

[10] http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/referrals/Tetrazepam_containing_medicinal_products/human_referral_prac_000015.jsp&mid=WC0b01ac05805c516f (accessed November 17, 2016)

[11] Arrow K. Economic welfare and the allocation of resources for invention. *The rate and direction of inventive activity: Economic and social factors*. Princeton University Press 1962: 609-26.

[12] Laffont JJ, Tirole J. Using cost observation to regulate firms. *Journal of Political Economy* 1986; **94**: 614-41.

[13] Kremer M. Patent buy-outs: a mechanism for encouraging innovation. *NBER* 1997.

[14] Chu LY, Sappington DEM. Simple cost-sharing contracts. *The American Economic Review* 2007; **97**: 419-28.

[15] Xoxi E, Tomino C, de Nigro L, Pani L. The Italian post-marketing registries. *Pharmaceutical Programming* 2012; **5**: 57-60.

[16] Goldman DP, Jena AB, Philipson T, Sun E. Drug licenses: a new model for pharmaceutical pricing. *Health Aff* 2008; **27**: 122-9

[17] Larkin J, Hodi FS, Wolchok JD. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N Engl J Med* 2015; **373**: 1270-71

[18] Kowdley KV, Lawitz E, Crespo I, Hassanein T, Davis MN, DeMicco M, et al. Sofosbuvir with pegylated interferon alfa-2a and ribavirin for treatment-naive patients with hepatitis C genotype-1 infection (ATOMIC): an open-label, randomised, multicentre phase 2 trial. *Lancet* 2013; **381**: 2100-7.

[19] Sulkowski MS, Gardiner DF, Rodriguez-Torres M, Reddy KR, Hassanein T, Jacobson I, et al. Daclatasvir plus sofosbuvir for previously treated or untreated chronic HCV infection. *N Engl J Med* 2014*;* **370**: 211-21

[20] Crepaz N, Hart TA, Marks G. Highly active antiretroviral therapy and sexual risk behavior. *JAMA* 2004; **292**: 224-36.

[21] McPhail GL, Clancy JP. Ivacaftor: the first therapy acting on the primary cause of cystic fibrosis. *Drugs Today* 2013; **49**: 253-60.

[22] Pettit RS, Fellner C. CFTR modulators for the treatment of cystic fibrosis. *PT* 2014; **39**: 500-11.

Appendix: Details of Simulations

In order to focus on the issues addressed by two-price programs, we have made a number of assumptions. The purpose of this section is to spell out the assumptions made, and to discuss why changing those assumptions would not materially affect the results. The Excel spreadsheets used to produce these simulations are available from the authors upon request, and will allow the interested reader to change parameters to see the results for themselves.

In general, we have used real data from the Italian experience and published sources wherever possible. The assumptions that are made are done so to increase transparency of the mechanisms at work in our proposed payment programs.

Fee-per-dose Benchmark

As should be clear from the text the fee-per-dose benchmark we simulate assumes a single price for each dose purchased by a payer. The most common way this assumption is violated in practice is through the negotiation of price-volume discounts. However, if the price-volume discounts are based on the amount purchased in a particular time frame, for example, a year, in most cases this would perform just like a fee-per-dose program as we have modeled it, but the cost per dose would be replaced by the average cost per dose over the year. On the other hand, if the price-volume discount calculated those discounts over the entire length of time that the drug was being purchased by the payer, this would be very similar to a treatment-based program, except that the continuation price would be going down over time. Such price-volume discounts are, in our experience, exceedingly rare.

Patient-based Program

The simulated patient-based program is based on the treatment of metastatic melanoma with nivolumab + imipilumab. The pricing in the fee-per-dose program is based on the monthly list price of these two compounds of 20,000 €.

The number of patients is determined using very simple survival dynamics. Without treatment, patients die at a constant rate over quarters, so that all patients diagnosed at a given point in time are all dead by a specific point in time in the future. Thus, the survival dynamics depend only on a single parameter. For the purposes of this simulation, 50% of patients are assumed to die each quarter so that the median survival time after diagnosis / progression is 3

months, and all patients have died in two quarters. More complicated survival dynamics would increase the number of parameters in the simulation, and introduce uncertainty associated with each of them, without materially affecting the results.

Survival once on the immunotherapy nivolumab + imipilumab is also modeled simply. Essentially, for most patients, the treatment is assumed to make no difference in the length of their survival after diagnosis / progression. However, for some minority of patients, the treatment is assumed to drastically increase their life expectancy—in the case of the simulation, all patients who respond in this way are assumed to survive until the end of the simulation. The major difference between the two scenarios considered in the simulation is that in one case the number that respond extremely positively to treatment is 10 and 20%.

The number of patients, and the percent of those patients that survive a very long time are based on the number of patients in the Italian system, and preliminary evidence from clinical studies, which last at most a year. The assumptions here would need to be drastically changed in response to two focal possible changes. The first would be new indications for the treatment, such as other cancers, or earlier stage cancers of the same type. The second would be bio-markers that predict more accurately which patients will respond extremely positively to treatment. The former would unambiguously increase the total number of patients. While this would drastically increase the cost of either the fee-per-dose or treatment-based program, it would keep the relative costs of the two programs the same. More accurate bio-markers would decrease the cost of the fee-per-dose program towards the cost of the patient-based program. Perfectly predictive biomarkers would make these two programs essentially the same (for properly calibrated prices).

As noted in the text, prices in the patient-based program are calibrated so that the total cost to the payer in the fifth year is approximately the same. This is done using an initial waiting period before initial payment in the patient-based program of 3 months (so that 50% of the patients that do not respond to treatment do not cost the payer anything), and an initial payment

of 240,000 € (four times the list price of one-quarter of treatment), followed by a continuation payment of 2,000 €.

Treatment-based Program

The simulated treatment-based program is based on the Hepatitis C treatment Sovaldi (Sofosbuvir). We have focused only on the list price of taking this compound for 12 weeks, ignoring the fact that this drug is often combined with others, and the length of treatment is often as much as 24 weeks. As such, the fee-per-dose we use here is 135,000 € (per patient, per 12 weeks of treatment). Accurately assessing the treatment mix in a particular population would change the estimate of the fee-per-dose (per patient) price, but unless the pricing of these additional compounds were negotiated under the same treatment-based program as Sovaldi this would not change the cost savings (or additional costs over the short term) of our simulated treatment-based program. On the other hand, the fact that many patients are treated up to 24 weeks means that the potential savings of the treatment-based program may be significantly understated.

As described in the text, the prices in the treatment-based program are calibrated to deliver increased costs to the payer in the short term, and the same total cost by around year five. These numbers are thus implied by all the other numbers in our simulation. So if the price per patient is too high in the fee-per-dose formulation is too low or too high, so too will be the initial price in this simulation. For reference, the initial price of 270,000 € (double the list price used in the fee-per-dose program) to be paid on the first 100,000 patients. As such, using a more accurate mixture of treatment lengths would just require re-calibrating the prices in the treatment-based program by doubling the (more accurate, based on mix of length-of-treatment, and standard discounts) price per patient in the fee-per-dose program. The assumed continuation price of 1,000 € may be negotiated to be much lower based on the fact that Sovaldi is a small-molecule drug, and likely not all that costly to produce.

The baseline number of patients in the simulation are based on the Italian experience, with a number of assumptions about future patient trends. The first full year the treatment with Sovaldi was available was 2015. In the first half of 2016 the number of new patients grew by about 10%. We assume that this growth is constant across quarters, and will continue for 5 years, followed by a gradual decline in new patients due to reduced transmission and other factors. In

order to give the range of possible costs in the simulation we modify this baseline number by both adding and subtracting 10% in each quarter. This allows us a reasonable range of mis-estimation of 20% of the patient population. Making this range wider would have two effects: first the gap in total cost between the two scenarios in the fee-per-dose program would be widened; and second, the amount of time it would take for the treatment-based program to get to the very flat portion of the total cost curve would be widened.