Theory of constraints for publicly funded health systems

Somayeh Sadat · Michael W. Carter · Brian Golden

Received: 4 January 2012 / Accepted: 23 July 2012 / Published online: 21 August 2012 © Springer Science+Business Media, LLC 2012

Abstract Originally developed in the context of publicly traded for-profit companies, theory of constraints (TOC) improves system performance through leveraging the constraint(s). While the theory seems to be a natural fit for resource-constrained publicly funded health systems, there is a lack of literature addressing the modifications required to adopt TOC and define the goal and performance measures. This paper develops a system dynamics representation of the classical TOC’s system-wide goal and performance measures for publicly traded for-profit companies, which forms the basis for developing a similar model for publicly funded health systems. The model is then expanded to include some of the factors that affect system performance, providing a framework to apply TOC’s process of ongoing improvement in publicly funded health systems. Future research is required to more accurately define the factors affecting system performance and populate the model with evidence-based estimates for various parameters in order to use the model to guide TOC’s process of ongoing improvement.

Keywords Theory of constraints · Publicly funded health systems · Goal · Performance measures · System dynamics

1 Introduction

Theory of constraints (TOC), first developed in manufacturing by Eliyahu Goldratt [1], views systems and processes as series of dependent events. Analogous to the fact that a chain is only as strong as its weakest link, all systems have at least one and at most a few constraints that limit system performance [2]. TOC emphasizes that improving the constraint’s performance directly results in enhancing total system performance [3]. Thus, TOC provides insight on process improvement efforts by focusing the improvement efforts on the root cause of the problem (the constraint), as opposed to picking random or “low-hanging” fruits.

TOC’s process of ongoing improvement has five steps. First, system constraint, which keeps the system from achieving more of its goal, is identified. Second, since the constraint’s output is the limiting factor of the whole system’s output, the constraint is exploited to squeeze the most out of it. For physical constraints, such as the production capacity of a machine in a production line, the exploitation step ensures that the constraint is utilized most effectively. Constraining policies, such as those forbidding overtime, are replaced by policies that allow the exploitation of the constraint. The third step is to subordinate everything else to the constraint. At this stage, non-constraint resources are operated in a way to support the plan to exploit the constraint, for instance by ensuring that the constraint machine does not have to wait for parts produced by non-constraint machines (which may be producing parts that the constraint does not need yet). If the system performance is not enhanced to the desired extent, in the fourth stage the constraint is elevated by increasing its capacity (potentially a costly stage and therefore a strategic decision). For instance, an identical machine to the constraint machine may be purchased and installed to increase the constraint stage capacity, and through that the system output. Recognizing the fact that the system constraint has probably shifted, the final stage is to overcome inertia and start over from the first step to identify the new system constraint. [4]
A prerequisite to implementing TOC’s process of ongoing improvement is measuring the performance of different stages of the system with regards to the goal. Therefore, a goal along with appropriate performance measures is needed. The classical TOC literature has fully discussed the goal and performance measures for publicly traded for-profit companies [5]. But the equivalent literature for publicly funded health systems is scarce, and does not cover all of the performance measures. The purpose of this paper is to fill the gap in the literature by finding the equivalents of TOC’s classical goal and performance measures for publicly funded health systems and build a conceptual framework for implementing TOC’s process of ongoing improvement in this setting. As such, the purpose of this paper is to see how TOC sees the publicly funded health system with regards to defining the goal, performance measures, and the process of ongoing improvement.

To set the context, a system dynamics (SD) representation of the goal and performance measures for publicly traded for-profit companies is developed in the next section. This representation forms the foundation for a similar model, capable of defining the relationships between the suggested goal and performance measures, for publicly funded health systems. The model is then expanded to include some of the factors that affect system performance, providing a framework to apply TOC’s process of ongoing improvement in publicly funded health systems. For brevity reasons, we use the term “health system” in lieu of “publicly funded health system” in the rest of this paper.

2 An SD model of TOC’s classical goal and performance measures

TOC defines the goal of the system from the perspective of its owners. In classical TOC literature dealing with publicly traded for-profit companies, shareholders are the owners and they define the goal as making money, both now and in the future.

The performance of the company is measured with regards to the goal and through the global performance measures of cash flow, net profit, and return on investment. Moreover, a set of operational performance measures, namely throughput, inventory, and operating expenses, are defined that enable company management to assess the effects of local actions on global performance measures. In-depth explanations of these performance measures are discussed in the literature [5].

There is a hierarchy amongst the global performance measures. At the bottom is the survival performance measure of cash flow. One usually would not invest in a company expected to go bankrupt. At the next level, the more the company makes money (the absolute measure of net profit), the more appealing it is for investment. But the ultimate performance measure is the relative performance measure of return on investment: how much money is made relative to the amount invested. [5]

To shed more light on the descriptive explanations of these measures in the literature and better demonstrate the relationships between TOC’s classical goal and performance measures, we developed a system dynamics (SD) stock-and-flow diagram as depicted in Fig. 1. Stock-and-flow diagrams provide easy to understand illustrations of the relationships between various variables and the change in their values over time. Fundamentals of system dynamics and stock-and-flow diagrams are well described in the literature [6]. Basically, there are three types of variables in a stock-and-flow diagram: stock (demonstrated by boxes), flow (demonstrated by valves going into and/or out of stock variables), and auxiliary variables (the rest of the variables connected to other variables through simple arrows). Stocks are entities that accumulate or deplete over time and flow variables are the rate of change in stocks. A flow variable

![Fig. 1 TOC’s classic goal and performance measures](image-url)
with a positive value going into (out of) a stock variable appreciates (depreciates) the value of that stock. A flow variable with a negative value going into (out of) a stock variable would depreciate (appreciate) the value of that stock. Auxiliary variables are used to better demonstrate the relationships between stock and flow variables, as well as other auxiliary variables of interest. Throughout our descriptions of the models below, the terms in italics refer to specific model variables in various figures.

The model in Fig. 1 consists of three stock variables (Total Cash, Inventory, and Total Dividend Pay Out) and a number of flow variables representing the rate of change in each of the stock variables.

**Total Cash** is the cash on hand that the company has retained and not yet invested or paid out in any form. **Inventory** is all the money the company has invested in purchasing things it intends to sell. While at first sight this appears to only include the inventory of materials, it also encompasses any other facility or equipment that the company possesses. The company can and will sell these possessions when they are not required anymore. However, it should be noted that the value of inventory is calculated only based on the purchase price. The potential extra money that the company might get from selling the inventory (above the purchase price) is the throughput in TOC language: “the rate at which the company generates money through sales” [5]. **Throughput** is another input to the Total Cash stock. If the cash generated through the sale of inventory is less than the initial purchase price, the Investment Gained Back flow would still remain equal to the purchase value of part of the inventory that is being sold, but the Throughput flow would have a negative value.

**Inventory** can also depreciate through waste of investments (e.g. scrapped, perished, or aged goods). This is shown by the Investment Waste flow. As the money wasted is not gained back through sales, this rate does not flow back to the Total Cash stock.

Finally, the **Total Cash** stock is also reduced by the **Overhead cost** flow, which is the rate of spending money to operate the company (e.g.: rent, salaries, and interest on debt). These expenditures are not gained back through sales, and as such are not considered investments.

The value of each stock variable at each time T is formulated as the accumulation of its input minus its output flows, plus its initial value:

\[
\text{Total Cash}(T) = \int_{0}^{T} (\text{Net New Financing}(t) - \text{Dividend Pay Out}(t) - \text{Investment}(t) + \text{Investment Gained Back}(t) + \text{Throughput}(t) - \text{Overhead Cost}(t)) dt + \text{Total Cash}(0)
\]

\[
\text{Total Dividend Payout}(T) = \int_{0}^{T} \text{Dividend Payout}(t) dt + \text{Total Dividend Payout}(0)
\]

\[
\text{Inventory}(T) = \int_{0}^{T} (\text{Investment}(t) - \text{Investment Gained Back}(t) - \text{Investment Waste}(t)) dt + \text{Inventory}(0)
\]
Definitions of the various flow rates depend on the special circumstances of the company. The above variables model company’s basic flow of cash. Since the stock variables are expressed in dollars, all the flow variables are expressed in dollars per some period of time, for instance dollars per year. The model also illustrates TOC’s local and global performance measures. Inventory and Throughput variables represent two of TOC’s three local performance measures. The third local performance measure, operating expense, is modeled as an auxiliary variable (titled Operating Expense) and formulated as follows (the term in brackets shows the unit of measurement):

\[
\text{Operating Expense} \left[ \frac{\text{\$}}{\text{Year}} \right] = \text{Overhead Cost} + \text{Investment Waste}
\]

The global survival performance measure of Cash Flow is represented in the model by the Total Cash stock variable: if the value of this stock variable hits zero, the company would go bankrupt. The other two global performance measures are represented with auxiliary variables Net Profit and Return on Investment respectively, and formulated as follows (the terms in brackets show the units of measurement):

\[
\text{Net profit} \left[ \frac{\text{\$}}{\text{Year}} \right] = \text{Throughput} - \text{Operating expense}
\]

\[
\text{Return on Investment}[\text{Annual}] = \frac{\text{Net profit}}{\text{Inventory}}
\]

A final point to discuss here is discounting the value of money over time as commonly practiced in engineering economics. Auxiliary variables that represent the equivalents of any of the monetary variables (those measured in $ or $/Year) at time zero can be easily added to the model and formulated by using the continuous compound interest rate formula: \( P = \frac{F}{r} \), where \( r \) is the interest rate per period of time and \( N \) is the number of time periods between \( P \) (present value) and \( F \) (future value) [7]. For simplicity, we have not added these auxiliary variables to the model.

Having created an SD model of TOC’s definition of the goal and performance measures for publicly traded for-profit companies, we now move forward to define the equivalent goal and the performance measures for health systems.

### 3 Defining the goal for health systems

In developing the goal for health systems, we follow the same logic in deriving the goal for publicly traded for-profit companies in that the goal should be defined by the owners of the system. We see tax payers as the true owners because similar to shareholders investing in publicly traded for-profit companies, tax payers are investing in health systems.

We propose that tax payers fund health services so that the society enjoys longer and higher-quality lives in general. Therefore, we define the goal as: *to increase the quality and quantity of lives, both now and in the future.* We admit that this is a simplistic overall goal for the health system, but it serves our purpose of finding an equivalent for TOC’s simplistic overall goal for the publicly traded for-profit companies.

A unit of measurement is needed for evaluating the goal. We suggest a unit that simultaneously captures gains from reduced morbidity (quality gain) and reduced mortality (quantity gain) and combines the two into a single measure, such as quality adjusted life years (QALY) [8]. It is not our intention to endorse QALY as the “right” unit of measurement. As we will discuss in the discussions section of the paper, there are other competing measures developed in the health economics literature. To explicate our model, however, we had to use a unit of measurement and we selected QALY due to its prevalence in the health economics literature. In fact, in discussing a different topic (the accuracy of QALY as a utility), Drummond [8] also says: “The view is that the QALY is a good basic definition of what we are trying to achieve in health care, and maximizing QALYs is quite an appropriate goal”.

Our goal statement is similar in concept to two suggested goal statements in the TOC literature on health care, namely “to maximize the units of health” [3] and “to make more health, today as well as in the future” [9]. However, we believe that our expression of the goal allows for a more accurate development of TOC’s operational and global performance measures for health systems, as described in the following section. We differ in the goal statement with those suggesting treating more patients both now and in the future [10]. We do not believe society funds the system to maximize the number of patients treated. In particular, the society might prefer that there are no patients in the first place to be treated (i.e. favor prevention over treatment where possible and economically feasible). We do not agree with a goal statement of “providing safe care while thinking of money only as an operating expense – something to be reduced” [11] for two reasons: First, we do not see the goal as only providing care; we see prevention as part of the goal as well. Second, in accordance with classical TOC literature that does not include operating expenses in the goal statement, we believe that the goal statement should be kept separate from any constraints that need to be overcome to achieve the goal. For this latter reason, two remaining goal statements mentioned in the TOC literature on health care are not our preferred ones either, namely “to maximize life expectancy...
and quality of life at an acceptable cost to society” [12] and “maximize quality medical services provided to its customers, subject to budgetary constraints” [13].

4 An SD model of TOC’s performance measures for health systems

Except for Schaeferes et al. [9], none of the studies in the literature try to define all the performance measures in a systematic way and relate them back to the goal. Schaeferes et al. [9] formulate the goal with some units of life time expectancy multiplied by an index of life quality, and try to link the performance measures to this goal. Sadat [14] discusses in detail a number of ambiguities and possible flaws with their formulations of the throughput and inventory. Moreover, the fact that they keep the definition of operating expenses the same as that of the classical TOC literature (money spent on running the organization) results in them not being able to define the global performance measures of net profit (since their defined throughput has a different unit of measurement than operating expenses and as such the two terms can’t be subtracted) and, as a result, return on investment. As such, the only paper to date trying to define the relationship between TOC’s goal and performance measures in health systems is not as accurate or comprehensive as needed to be able to provide insight on the application of TOC to health systems.

We use SD’s stock-and-flow diagrams to define the relationships among TOC’s goal and the various performance measures for health systems. Our goal is to define the equivalents of the variables in Fig. 1 for health systems. We accomplish this goal in two steps. First, we model the goal and the performance measures from the perspective of one individual interacting with the health system by investing or spending goal units (QALYs) with the expectation to gain more goal units (QALYs) in the future. Next, we scale up the model to incorporate all the individuals in the society and formulate the goal and performance measures at the health system level. Both steps are demonstrated in Fig. 2. The term “total” that appears in parenthesis in the title of all variables only applies to the societal model. The individual model variable names do not require the term “total”. For instance, the stock variable in the middle is named Remaining QALYs for the individual model, and Total Remaining QALYs for the societal model.

Similar to Fig. 1, Fig. 2 consists of three stock variables (Remaining QALYs, Invested QALYs, and QALYs Lived Outside the Health System) and a number of flow variables representing the rate of change in each of the stock variables.

Remaining QALYs is the equivalent to the Total Cash stock in Fig. 1. Remaining QALYs is the QALYs ahead of an individual at any point of time in life. The individual usually does not know the exact value of this variable at any point in time, but much like a company deciding how to spend its cash (e.g. whether on investments or paying dividends), the individual decides how to spend the Remaining QALYs on different activities. For the sake of our model, these activities are divided into activities that are part of the health system and those not related to the health system, illustrated on the right and left side of the Remaining QALYs stock variable in Fig. 2 respectively.

To the left of the Remaining QALYs stock is interactions outside the health system. Similar to the Net New Financing in Fig. 1, the Net New QALYs Gain External to the Health System is the rate of acquiring QALYs other than through interactions with the health system. This flow mainly represents birth (when we are all given some unknown Remaining QALYs). Additionally, some activities such as exercising in the gym that are not usually considered part of the

---

Fig. 2 Health system’s goal & performance measures
publicly funded health systems also might result in some QALYs gain. The additional QALYs gain (the additional QALYs above and beyond the amount of QALYs spent on activities such as exercising) is represented by the **Net QALYs Gain External to the Health System** flow variable. It is possible that activities such as exercising do more harm than good, and in such instances the value for **Net QALYs Gain External to the Health System** would be negative. In any case, the actual QALYs spent on such activities are part of the **Living QALYs Outside the Health System** flow (as described below). It is only the additional gain (or possibly loss) from this investment external to the health system that is modeled with the **Net QALYs Gain External to the Health System** flow.

The **Living QALYs Outside the Health System** is equivalent to the **Dividend Pay Out** flow in Fig. 1. It simply represents living life outside the health system (e.g. not including the time hospitalized). Assuming no interaction with the health system or activities outside the health system that might result in **Net QALYs Gain External to the Health System** over a period of time (e.g. a year), this flow simply is the rate of losing QALYs while living. An individual in perfect health would lose one QALY per year, while someone half as healthy would lose 0.5 QALYs per year. It is the accumulation of this flow over time, represented by the stock variable **QALYs Lived Outside the Health System** (equivalent to the **Total Dividend Pay Out** in Fig. 1) that individuals like the health system to maximize. That is, individuals interact with the health system with the hope to experience more **QALYs Lived Outside the Health System** before death.

Therefore, similar to the case in Fig. 1, an individual might choose to forgo some QALYs that could have been enjoyed immediately through **Living QALYs Outside the Health System** flow and invest those QALYs (through the **Investing QALYs in Health Programs** flow, which is the equivalent of the **Investment** flow in Fig. 1) with the hope to gain more QALYs in the future. We use the term health program very broadly here, including any activity or intervention offered by the health system that is expected to increase the quality and/or length of life such as health education and prevention, treatment, rehabilitation, and even palliative care (intended to solely increase the quality of life). All those QALYs invested in health programs for which the benefits have not been realized yet are represented by the **Invested QALYs** stock variable (equivalent to the **Inventory** stock variable in Fig. 1).

Similar to Fig. 1, the QALYs gain is represented by two different flow variables, **Gaining Back Invested QALYs** and **Gain of Additional QALYs due to Investment**, equivalent to **Investment Gained Back** and **Throughput** flows in Fig. 1, respectively. Similar to **Throughput** in Fig. 1, the **Gain of Additional QALYs due to Investment** could be negative if the QALYs gained from an investment is fewer than those invested. An example of this would be a treatment that does more harm than good, for instance due to a medical error or lack of effectiveness, and could even potentially result in death (i.e. the **Gain of Additional QALYs due to Investment** draining all the **Remaining QALYs**).

The final flow in Fig. 2, **Waste of QALYs due to Investment**, is the rate at which **Invested QALYs** are wasted when one participates in a health program. Examples include time spent to travel to the health program location or to wait around during the various stages of participating in a health program. These are wasted QALYs since no health benefit is expected as a result of losing them. However, given the current circumstances, including a shortage of capacity or possible inefficiencies in the system, the individual has to lose them in order to be able to participate in health programs. In other words, this flow is equivalent to the **Investment Waste** flow in Fig. 1. Since Fig. 2 does not have an equivalent to **Overhead Cost** flow of Fig. 1, **Waste of QALYs due to Investment** is the equivalent to the **Operating Expense** of Fig. 1 as well. It should be noted that ineffective interactions with the health systems, for instance therapies that do more harm than good, are still modeled with a negative **Gain of Additional QALYs due to Investment**.

The **Waste of QALYs due to Investment** flow is the loss of QALYs due to participating in a health program, be it a highly effective one (with a large positive **Gain of Additional QALYs due to Investment**) or an ineffective one (with a negative **Gain of Additional QALYs due to Investment**).

The three stock variables of the model in Fig. 2 are measured in units of QALYs and are formulated as the accumulation of their input flows minus their output flows over time, plus their initial values at time zero. All the flow variables of the model are measured in QALYs/Year. As with the case in Fig. 1, the exact definitions of the flow variables depend on the special circumstances of each individual.

**TOC**’s local performance measures of inventory, throughput, and operating expenses are represented in Fig. 2 by **Invested QALYs**, **Gain of Additional QALYs due to Investment**, and **Waste of QALYs due to Investment** variables respectively.

With regards to the global performance measures, similar to the **Total Cash stock** in Fig. 1, the **Remaining QALYs** stock serves as the survival performance measure, equivalent to cash flow in **TOC** language. Similar to the company going bankrupt if the **Total Cash** stock hits zero, the individual is dead if the **Remaining QALYs** stock reaches zero.

**TOC**’s other two global performance measures of net profit and return on investment are represented by
auxiliary variables of Net QALYs Gain and Return on Health Investment respectively, and are formulated as follows (the terms in brackets are the units of measurement):

Net QALYs Gain \[\frac{\text{QALY}}{\text{Year}}\] = Gain of Additional QALYs due to Investment – Waste of QALYs due to Investment

Return on Health Investment[Annual] = \[\frac{\text{Net QALYs Gain}}{\text{Invested QALYs}}\]

Similar to the case of publicly traded for-profit companies, there is a hierarchy amongst the global performance measures. At the bottom is the survival performance measure of Remaining QALYs. A person normally does not participate in a system that would kill them. At the next level, the more the health program creates QALYs (the absolute measure of Net QALYs Gain), the more appealing it is to participate in. But the ultimate performance measure is the relative performance measure of Return on Health Investment or how many QALYs are created relative to the amount invested.

A final note here is discounting future QALYs. Unlike in engineering economics where there is a general consensus around the need to discount the value of money over time, health economists are divided on their opinion around discounting health status. An interesting example of an argument against discounting is that “discounting years of life gained in the future gives less weight to future generations in favor of the present one” [8]. There are of course arguments in favor of discounting as well, including “simple numerical examples showing that leaving effects undiscounted while discounting costs, or discounting costs and effects at different rates, can lead to inconsistencies in reasoning” [8]. We advocate discounting QALYs since it better differentiates between those interacting with the health system who experience zero return on heath investment versus those not interacting with the health system at all. For instance, imagine two individuals with the same initial remaining QALYs; the first one invests in a health program and gains back the exact amount of QALYs invested (i.e. zero Gain of Additional QALYs due to Investment or Return on Health Investment) while the second one does not interact with the health system at all. Without discounting and assuming all else equal, both individuals would enjoy the same level of the QALYs Lived Outside the Health System. Discounting allows us to depreciate the value of QALYs enjoyed later in life by the first individual (as compared to the second individual) and might better reflect the fact the first individual might have chosen to not go through the trouble and interact with the health system at all if she knew that a zero Return on Health Investment is awaiting her. While the debate is definitely in the realm of health economics and not the subject of this paper, suffice it to say that if a decision to discount is made, the model can be easily modified to incorporate auxiliary variables representing the equivalents of the model variables at time zero, using the continuous compound discount formula [7]. For simplicity, we have not included these auxiliary variables.

To elevate the model to a societal level, all we need to do is to aggregate the values of the various flow variables across the whole population. With the exception of Total Net QALYs Gain External to the Health System flow, each of the flow variables for the societal model (i.e. variable names that include the term “total”) are the summation of the equivalent flow variables in the individual model for all the individuals in the society. For instance, the Total Waste of QALYs due to Investment flow is the summation of all the Waste of QALYs due to Investment flow values for all the individuals in the society.

At the societal level, the Total Net QALYs Gain External to the Health System also captures the net QALYs gained by accepting new immigrants (or lost by emigration). Immigrants only bring (or emigrants only take away) their Remaining QALYs at the time of immigration (or emigration) through this flow to (or out of) the Total Remaining QALYs stock, and not the Remaining QALYs they had upon birth. The stock variables are still the accumulation of their input minus their output flows over time, plus their initial value at time zero. The initial value at time zero is determined by summing the initial values of the corresponding stock variable of all the individuals in the society. The auxiliary variables are defined in the same way that they were defined for the individual model, except that they are based on the stock and flow variables of the societal model.

\[
\text{Total Net QALYs Gain} = \text{Total Gain of Additional QALYs due to Investment} - \text{Total Waste of QALYs due to Investment}
\]

\[
\text{Return on Total Health Investment} = \frac{\text{Total Net QALYs Gain}}{\text{Total Invested QALYs}}
\]
5 A framework for applying TOC’s process of ongoing improvement in health systems

Having completed the prerequisite steps of defining the goal and performance measures, we should now be able to articulate TOC’s process of ongoing improvement (the five steps discussed under the introduction section). The first step is to identify the bottleneck: What limits the performance of the system (as measured by the defined performance measures) the most. The goal and performance measures are fundamentally defined by the various flow variables of the model in Fig. 2. However, since we have not formulated the flow variables, it is not possible to determine what is most limiting the goal through the flow variables.

While the factors affecting flow variables differ from system to system and as such developing accurate formulations applicable to any health system is not possible without extensive further research, it is our intention to present some of the likely factors and their possible effects on the flow variables and use them to discuss, at a conceptual level and through a hypothetical example, how a more accurate and complete model can potentially guide TOC’s process of ongoing improvement.

Figure 3 is a snapshot of the model presented in Fig. 2 with some additional variables. Please note that any variable shown in <and> signs are the repetition of the variable with the same name and not between <and> signs presented elsewhere in the model. Moreover, since we are not formulating the flow variables here, only the direction of the relationship is shown in Fig. 3. A positive sign on an arrow connecting A to B means that, all else equal, if A increases (decreases), then B “increases (decreases) above (below) what it would have been” [6]. A negative sign means that, all else equal, if A increases (decreases), then B “decreases (increases) below (above) what it would have been” [6].

Given that in publicly funded health systems there is usually more demand than capacity, Total Investing QALYs in Health Programs flow is mostly affected by Available Capacity of the health system. Capacity is an aggregate term encompassing physical capacity (facilities, equipment, devices, etc.), health human resources (physicians, nurses, psychiatrists, public health professionals, etc.), and possibly soft types of capacity such as knowledge base of practice (i.e. knowledge of how to best invest QALYs in health programs). The higher the value of Available Capacity is, the higher the Total Investing QALYs in Health Programs can be.

Given a level of capacity, the manner in which the capacity is planned to be utilized, which we have titled Capacity Operational Planning, can affect the Total Waste of QALYs due to Investment. For instance, given a level of doctors’ availabilities at community clinics (e.g. two clinics with a full time doctor practicing in each), the manner in which the doctors’ time is planned to be used (i.e. appointment booking) as well as the location and the manner in which the physical capacity is planned out determine the waiting around and travel times that the patients would be experiencing. The better this Capacity Operational Planning is performed, the lower would be the Total Waste of QALYs due to Investment; hence the negative sign on the arrow connecting Capacity Operational Planning to Total Waste of QALYs due to Investment.

The values of the Total Gaining Back Invested QALYs and the Total Gain of Additional QALYs due to Investment flow variables depend on the Total Investing QALYs in Health Programs flow, since QALYs need to be invested in health programs before they can be gained back or additional (albeit sometimes negative) ones achieved. The value of Total Gaining Back Invested QALYs is positively affected by Total Investing QALYs in Health Programs flow, as the former is the rate at which the non-wasted portion of the latter is
The value of Total Gaining Back Invested QALYs is also negatively affected by Total Waste of QALYs due to Investment, as the fewer QALYs invested are wasted, the more of them would be available to be gained back.

The effect of the Total Investing QALYs in Health Programs flow on the Total Gain of Additional QALYs due to Investment can be either positive or negative and depends on the value of Health Program Effectiveness. Health Program Effectiveness captures the clinical appropriateness and quality aspects of the health service provided. Highly effective health programs return high levels of QALYs for every QALY invested, whereas less effective health programs return little or even take away QALYs (when the health program does more harm than good).

With a positive Health Program Effectiveness, the more the Total Investing QALYs in Health Programs flow, the more the Total Gain of Additional QALYs due to Investment flow will be. If the Health Program Effectiveness is negative (i.e. health program doing more harm than good), then the value of the Total Gain of Additional QALYs due to Investment flow will be negative and the more the Total Investing QALYs in Health Programs flow, the less (or indeed the more negative) the value of the Total Gain of Additional QALYs due to Investment Flow will be.

Total Gain of Additional QALYs due to Investment Flow is also positively affected by an improvement in Health Program Effectiveness, which can be improved through health research (both fundamental and clinical research), technological innovations resulting in more effective medical devices and technologies, and quality improvement initiatives such as those aimed at reducing adverse events or enforcing best practice adoptions.

The Total Living QALYs Outside the Health System flow is affected by the health system as well: during a certain period of time, the more QALYs are spent on health programs (Total Investing QALYs in Health Programs), the less QALYs will be available to be spent on other purposes of life; in other words, the value of the Total Living QALYs Outside the Health System flow will be reduced.

We admit that while the model as presented with these additional variables in Fig. 3 provides a framework to address high level health policy issues, extensive research is still needed to create an accurate and complete model populated with evidence based estimates for the various model parameters from a wide variety of literature in health services research, such as clinical, health economics, workforce, health technology, epidemiology, and operational research studies. However, one of the strengths of the model is that it does not necessarily need to be applied at the whole health system level. The conceptual model can be regarded as a framework and used at different levels: the high national or regional health system level, meso-levels for specific diseases (e.g. cancer or chronic kidney disease) or specific populations (e.g. seniors or particular ethnic groups), and the micro level of understanding individual patient decision making. It is possible that, within the current literature, there is sufficient evidence to populate the model at a specific level.

To demonstrate how the model can be used to guide TOC’s process of ongoing improvement, we present a hypothetical example around how policy makers can use the model to guide resource allocation for the care of chronic kidney disease (CKD) patients. CKD patients comprise pre-dialysis CKD patients (who might or might not know that they have some degree of kidney failure) and End Stage Renal Disease patients (ESRD) who are not able to survive without dialysis or a successful kidney transplant.

As a first step in demonstrating the value of the model in guiding CKD care resource allocations, we describe a hypothetical progression of the disease and various options available to an individual patient, and then look at the aggregate societal level model of CKD care and how the model guides decision making at the policy level.

An individual who does not know whether or not she has kidney failure may choose to invest her Remaining QALYs (through the Investing QALYs in Health Programs) to be screened, via blood and/or urine tests, for kidney failure once in a while. The investment is likely associated with some minimal Waste of QALYs due to Investment, for instance the travel and wasted time to get these diagnostic tests and consult with a family physician to discuss the findings. The rational for this investment is that the earlier the disease is discovered, the more it can be done to maintain the level of kidney function and through that a higher quality of life. If the individual learns the unfortunate outcome that her kidneys are not functioning as well as they should, but that fortunately the disease has not yet progressed to the ESRD stage, the individual then has to invest more QALYs in consulting a nephrologist. The patient might likely be put on a regimen of steroids to prolong the kidney function. These drugs may have side effects, but the patient takes them with the hope that overall, she will experience a positive Gain of Additional QALYs due to Investment (i.e. the individual hopes that the desired effects of these drugs in maintaining kidney function overweight the undesirable side effects).

Now assume that the unlucky patient, though successful in maintaining the kidney function for a while through investments in health, has finally progressed to the ESRD state. Given the gap between demand and supply of kidneys, it is most likely that the patient has to take the dialysis path first, even if she is medically a suitable candidate for a kidney transplant.

The patient will likely receive dialysis on a frequent basis (usually 3–5 times per week, for several hours in each session). Dialysis is an investment in health that the patient
hopes to gain back through extended life (i.e. without the
dialysis, the patient is likely to die). If the dialysis is per-
formed at a dialysis center, there might be some significant
Waste of QALYs due to Investment that occurs because of
frequent travels to the dialysis center. Home dialysis patients
save this waste of QALYs, as they are dialyzed at home.

Our patient might at some point get lucky and be
matched with an available kidney for transplantation. While
the preparation for and the actual operation and post-
operative period amount to a considerable Investing QALYs
in Health Programs, the hope is that the operation would
give the patient a well-functioning kidney and as such a
huge Gain of Additional QALYs due to Investment. Besides
kidneys from cadaver donors, there are living donors who
decide to forgo a kidney (the body can function with the one
remaining kidney). While at first sight the decision of these
people to donate one of their kidneys might seem contrary to
the normal case of trying to maximize one’s own Remaining
QALYs, assuming a level of rationality for these altruistic
people would suggest that they are gaining a pleasure from
helping a loved family member or simply a fellow human
being and through that pleasure increase the quality of their
lives. Moreover, in some health systems, financial rewards
for selling one’s kidney might incentivize these donors to
forgo some QALYs (by loss of one kidney) in the hope to
get more QALYs through the higher quality of life expected
to be associated with a higher level of wealth.

Our hypothetical patient might enjoy the benefit of the
new kidney to the end of her life, or just for a while (if at all)
before her body rejects it, at which point she would go back
to receiving dialysis and may choose to be put back on the
wait list to receive another transplant.

Having pictured a possible trajectory of the CKD disease,
let’s hypothetically explore how a policy maker can use the
model to identify improvement initiatives that result in the
highest level of Return on Total Health Investment.

A very possible constraint in the CKD care system is the
available soft knowledge capacity to identify and care for
CKD patients before they progress to the ESRD stage.
Many health systems, especially those in the developed
economies, might have the physical and financial capacity
of having at risk patients undergo blood and urine tests in
their annual check-ups with their family physician, or even
at more frequent intervals. However, there might be a soft
capacity gap or knowledge gap regarding the identification
of the at-risk population to be screened. In US, for instance,
only 28 % of individuals with hypertension underwent
serum creatinine levels measurement in 2003 [15], while
hypertension is a known CKD risk factor [16]. Many
patients may not even have seen a nephrologist before they
are diagnosed as ESRD.

A policy maker might want to explore how to exploit the
constraint (available soft knowledge capacity) by making
sure the knowledge (soft capacity) already available is best
disseminated among various stakeholders, especially family
physicians. Screening tools such as SCORED [16] can be
disseminated to ensure at risk population are identified and
screened for CKD.

In subordinating everything else to the constraint, policy
makers should identify whether the infrastructure to best
care for these pre-dialysis patients once they are identified
by their family physicians is at place. For instance, is there a
clear referral mechanism to refer these patients to the care of
a nephrologist, who can help maintain their level of kidney
function for as long as possible? Are the nephrologists
aware of the importance of caring for pre-dialysis CKD
patients and give them as high quality care as they should?

Only if the constraint is not elevated at this stage, for
instance if the official statistics show that still a high propor-
tion of patients are not diagnosed with CKD until they
are at the ESRD stage, should the policy makers think about
rationalizing the constraint. Elevating this soft knowledge cap-
acity constraint could include funding wide screening cam-
paings and general population education sessions. Once
finally it is determined that the system is indeed identifying
and caring for as many pre-dialysis CKD patients as pos-
sible in a timely fashion, or that further investments in these
initiatives would not result in the highest Return on Total
Health Investment compared to investing in other initiatives,
the focus should then move to the next step of the improve-
ment process, which is to avoid inertia and find the new
constraint (which would return the highest Return on Total
Health Investment at this point).

Assume that the new constraint is then identified as
inadequate levels of successful transplants, conceptualized
by Available Capacity in our model (in this instance the
capacity refers to available kidneys for transplants and the
medical and logistical processes to make transplants hap-
pen), as well as the success rate (i.e. long term organ
acceptance) in kidney transplant, conceptualized in our
model with Health Program Effectiveness.

To exploit the constraint, the policy maker might want to
ask if the current available kidneys are being best used by
the system. Are the kidneys that become available matched
against a large pool of potential recipients to make sure the
available kidneys are received by a recipient whose body is
most likely to accept the kidney (as much as it is possible to
predict organ acceptance with our current level of knowl-
edge). If not, a major step in exploiting the limited capacity
of available kidneys is to match available kidneys with the
most promising recipients.

The policy maker should also explore whether or not the
other processes subordinate themselves to improving the
successful rate of kidney transplants. Are the logistics of
harvesting and transporting kidneys supportive of achieving
best results? For instance, does everybody understand the
importance of lowering “cold ischemia time (i.e. chilling of the donor kidney when its blood supply was cut off prior to transportation)”, which “allows the transplant kidney to begin working quickly” [17]? Are the recipient notification processes agile and do these processes enable the system to act in a timely manner that maximizes the chances of a successful kidney transplant?

Only when the available kidneys have been allocated appropriately and in a timely manner to the most promising recipients should the policy maker explore whether an elevation of the constraint is needed, or indeed possible. If the society is lacking sufficient donors, policy makers can fund campaigns to increase awareness and improve donation rates. Policy changes such as passing laws that assume all cadaver donors have agreed to donate their body parts unless they have stated otherwise during their life time, as opposed to policies that assume cadaver donors did not want their body parts donated unless otherwise stated, is another possible avenue for increasing cadaver donor rates. Policy makers may also invest in more clinical and fundamental research to better understand predictors of successful kidney transplants, which would result in higher success rates through an improved Health Program Effectiveness.

Assuming that the constraint has been elevated to the level that there is now a new constraint (i.e. there is now a new constraint, investments in which are expected to produce a higher potential level of Return on Total Health Investment), the policy maker should then start the process again by identifying the new constraint, exploiting it, subordinating everything else to the constraint, and if needed and possible, elevating the constraint, and repeat this cycle of TOC process of ongoing improvement.

6 Discussion and future research

A fundamental principle of TOC is that all of the multiple objectives of a publicly traded for-profit enterprise can be reduced to “making money now and in the future” [1]. For example, corporations may display “social responsibility” because they believe that this will improve their image and bring in more customers in the future. While we agree that this is highly debatable, the purpose of the assumption is to help construct a simple corporate model that we can then test and evaluate. This simple model fails to describe behaviour in a public enterprise like healthcare. The real purpose of our paper is to attempt to find a single objective (increasing the quality and quantity of lives, now and in the future) that we can use to construct a paradigm for TOC in healthcare. Although we need to make many simplifying assumptions to build our model, the result is a simple model that clarifies key issues through all the complexity and brings stakeholders to common understanding.

The model should be used to inform health policy and management decision-making. Any use beyond improving the health care system (such as using the model to inform immigration policies) is debatable, as the model does not capture the complexities beyond those present in the health care system.

With regards to the unit of measurement for the goal, there have been controversies around the concept and use of QALYs in the literature. Some interesting discussions can be found in the literature [18–24]. Some question the QALYs methodology, mainly due to measurement difficulties, without proposing any other evidence based method [24]. Others propose alternatives to QALYs, with each method having strengths and weaknesses associated with it, including healthy-year equivalents (HYE) [25–27] and Saved-Young-Life equivalents (SAVE) [28–30]. For the sake of our conceptual model, a unit of measurement is needed that captures both quality gains and quantity gains of life. While the choice of the right unit to measure quality and quantity of lives is in the realm of health economics and not the topic of this research, we chose to use QALY due its prevalence in the literature.

Once the unit of measurement is appropriately chosen, the model is capable of providing an overall guideline for discussing some of the most controversial issues in health policy. Take the public discussions around equity of access for instance. While in our conceptual model, we gave all members of the society the same weight in aggregating the individual level models to construct the societal model, a similar idea to that of Williams [23] by weighing individuals of the society differently (e.g., giving more weight to the poor or to the young) is also possible. We are not advocating (nor objecting to) assigning different weights to individuals in the society. All we say is that the model is capable of handling this policy issue, should a decision be made to weigh individuals differently.

The model is also capable of addressing rationing issues. Should the available capacity be less than that desired (as is the case for most health services in publicly funded health systems), the decision as to who should receive the health services should be made. While the conceptual model of TOC performance measures does not explicitly specify how to make these decisions, it includes the essential elements to address it. The TOC process of ongoing improvement would suggest exploiting and subordinating everything else to the available capacity, as the constraint of the system. Therefore, the TOC process of ongoing improvement would favour providing services to those most likely to benefit from them (i.e. those with higher Return on Health Investment on every QALY invested in health services). Assuming the same level of Waste of QALYs due to Investment across the population, higher Return on Health Investment would correspond with higher clinical priorities, as conceptualized
with the Health Program Effectiveness element of the model. We admit that it would take a bold politician or decision maker to implement such explicit rationing rules. But, on the positive side, it does bring the rationing discussion to the front and help the public make explicit decisions on how to maximize the goal. Of course, other considerations of the public, such as favoring segments of the population, can be made prior to moving to the societal model by giving different sub-populations different weights.

Given the difficulties of quantifying the model at the aggregate levels (for instance at high national or meso levels of certain populations or regions), the model can be used as a qualitative tool to help communicate major ideas across different stakeholders of the system. Indeed, influence diagrams (also known as cause and effect diagrams) used in system dynamics could have been used to present the model in the first place. Our choice of a stock-and-flow diagram was due to our desire to bring some more clarity to the details of defining the performance measures, as it is sometimes difficult to understand all the units of measurement in an influence diagram. Now that the level of clarity has been achieved, should decision makers see that there is not enough evidence to populate the model quantitatively at certain levels of its application, the model can be used as a qualitative tool to guide TOC’s process of ongoing improvement.

A final discussion that we would like to present here is the potential applicability of our conceptual model to health systems that are not publicly funded. To the extent that the goal of such health systems is similar to that of the publicly funded health systems (e.g. some non-profit privately funded health systems), our conceptual framework of the system-level TOC performance measures presented might be useful for such health systems as well (though it is possible that the model would need some modifications to be fully applicable to these settings). However, if the goal of the health system is to maximize profit, then the classical TOC performance measures of cash flow, net profit, and return on investment might be more appropriate.

7 Conclusions

Theory of constraints has the potential to improve the resource-constrained publicly funded health systems by focusing improvement efforts on system constraints. However, to date, the literature has not adequately addressed the required customizations to make the theory applicable to this environment. Building upon a system dynamics representation of TOC’s classical goal and performance measures for publicly traded for-profit companies, we developed a similar model that defined the goal and performance measures for publicly funded health systems. The model was then expanded, at a conceptual level, to include various factors affecting health system performance with the aim to guide TOC’s process of ongoing improvement.

Acknowledgments The authors would like to thank the generous financial support by MDS Inc, and the Discovery Grant (M.W.Carter) from NSERC.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Goldratt EM, Cox J (1984) The goal: excellence in manufacturing. North River Press
2. Womack DE, Flowers S (1999) Improving system performance: a case study in the application of the theory of constraints. J Healthc Manag / American College of Healthcare Executives 44(5):397–405, discussion 405–7
3. Breen AM, Burton-Houle T, Aron DC (2002) Applying the theory of constraints in health care: Part 1–The philosophy. Qual Manag Health Care 10(3):40–46
4. Rahman S (1998) Theory of constraints: a review of the philosophy and its applications. Int J Oper Prod Manag 18(4):336–355
5. Goldratt EM (1990) The haystack syndrome. North River Press
6. Stermann JD (2000) Business dynamics: systems thinking and modeling for a complex world. Irwin/McGraw-Hill
7. Riggs JL (1996) Engineering economics. McGraw-Hill, New York
8. Drummond MF (1987) Methods for the economic evaluation of health care programmes. Oxford University Press, New York
9. Schaeffers J, Colin J, Aggoune R, Kucina M (2007) A contribution to performance measurement in the healthcare industry: the industrial point of view. Int J Bus Perform Manag 9(2):226–239
10. Wright J, King R (2006) We all fall down. North River Press
11. McNutt RA, Owczarezy MC (2004) The theory of constraints and medical error: a conversation with Robert A. McNutt. QMHC 13(3):183–187
12. Hunink MG (2001) In search of tools to aid logical thinking and communicating about medical decision making. Med Decis Mak- ing 21(4):267
13. Ronen B, Pliskin JS, Pass S, NetLibrary I (2006) Focused operations management for health services organizations. Jossey-Bass
14. Sadat S (2009) Theory of constraints for publicly funded health systems. Dissertation: University of Toronto (Canada)
15. Stevens LA, Fares G, Fleming J, Martin D, Murthy K, Qiu J, Stark PC, Uhlig K, Van Lente F, Levey AS (2005) Low rates of testing and diagnostic codes usage in a commercial clinical laboratory: Evidence for lack of physician awareness of chronic kidney disease. J Am Soc Nephrol 16(8):2439–2448
16. Bang H, Vupputuri S, Shoham DA, Klemmer PJ, Falk RJ, Mazumdar M, Gipson D, Colindres RE, Kshirsagar AV (2007) Screening for occult renal disease (SCORED): a simple prediction model for chronic kidney disease. Arch Intern Med 167(4):374
17. Porter ME, Baron JF, Chacko JM, Tang RJ (2010) The UCLA medical center: Kidney transplantation. Harvard Business School Case: 711-410
18. Carr-Hill RA (1991) Allocating resources to health care: is the QALY (Quality Adjusted Life Year) a technical solution to a political problem? Int J Health Serv 21(2):351–363

Springer
19. Carr-Hill RA, Morris J (1991) Current practice in obtaining the “Q” in QALYs: a cautionary note. Br Med J 303(6804):699
20. Spiegelhalter DJ, Gore SM, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR (1992) Quality of life measures in health care. III: Resource allocation. BMJ: Br Med J 305(6863):1205
21. Broome J (1993) Qalys. J Public Econ 50(2):149–176
22. Nord E (1993) Toward quality assurance in QALY calculations. Int J Tech Assess Health Care 9(1):37–45
23. Williams A (1993) Economics, QALYs and medical ethics—a health economist’s perspective. Health Care Anal 3(3):221–226
24. McGregor M, Caro JJ (2006) QALYs: are they helpful to decision makers? Pharmacoeconomics 24(10):947
25. Mehrez A, Gafni A (1989) Quality adjusted life years (QALYs), utility theory, and healthy years equivalent (HYE). Med Decis Making 9:142–149
26. Mehrez A, Gafni A (1990) Evaluating health related quality of life: an indifference curve interpretation for the time trade-off technique. Soc Sci Med 31(11):1281–1283
27. Mehrez A, Gafni A (1991) Healthy years equivalent: how to measure them using standard gamble approach. Med Decis Making 11:140–146
28. Nord E, Richardson J, Macarounas-Kirchmann K (1992) Social evaluation of health care versus personal evaluation of health states: evidence on the validity of four health state scaling instruments using Norwegian and Australian survey data. National Centre for Health Program Evaluation
29. Nord E (1995) The person-trade-off approach to valuing health care programs. Med Decis Making 15(3):201
30. Nord E (1996) Health status index models for use in resource allocation decisions. Int J Tech Assess Health Care 12(1):31–44