评论 香港 深度 2019冠状病毒疫情

传媒误读数据 —— 认识疫情数据的常见误区

数据从选取、分析到呈现，每一步都涉及各种选择，依靠各种专业判断；只要略为不留神，就会产生误导而不自知。

2022年2月11日，香港土瓜湾一间私家诊所外，有不少轮候接种疫苗的市民。摄：林振东/端传媒

特约撰稿人 梁启智 发自台北 | 2022-02-13

Simpson’s paradox 辛普森悖论 Omicron变种病毒 Delta变种病毒 COVID-19 2019冠状病毒疫苗

疫情爆发至今已逾两年，尽管疫情影响我们日常生活的每一环节，公众对疫情的认识和掌握仍然十分有限。原因一方面是因为疫情随变种病毒的出现而不停改变，专家以科学方法所得的答案往往一转眼便已跟不上最新的发展。而当社会本身缺乏互信，当权者欠缺认受性，则各种有意或无意的假消息散播更会变得一发不可收拾，为控制疫情增添困难。

在一个理想的民主社会当中，我们会期望传媒可帮助公众更知情地参与公众事务，决定自己的未来。而在信息纷乱之际，从数据出发本应可让我们较客观地分析事实。毕竟数据讲求定义，只要定义清晰则应可减少误解。

不过两年来，很不幸，传媒理解和呈现数据的能力明显受到挑战，有时越讲越糊涂。数据虽然越来越多，却不见得社会中各种对疫情理解的矛盾有所减少。

苹果要和苹果比较

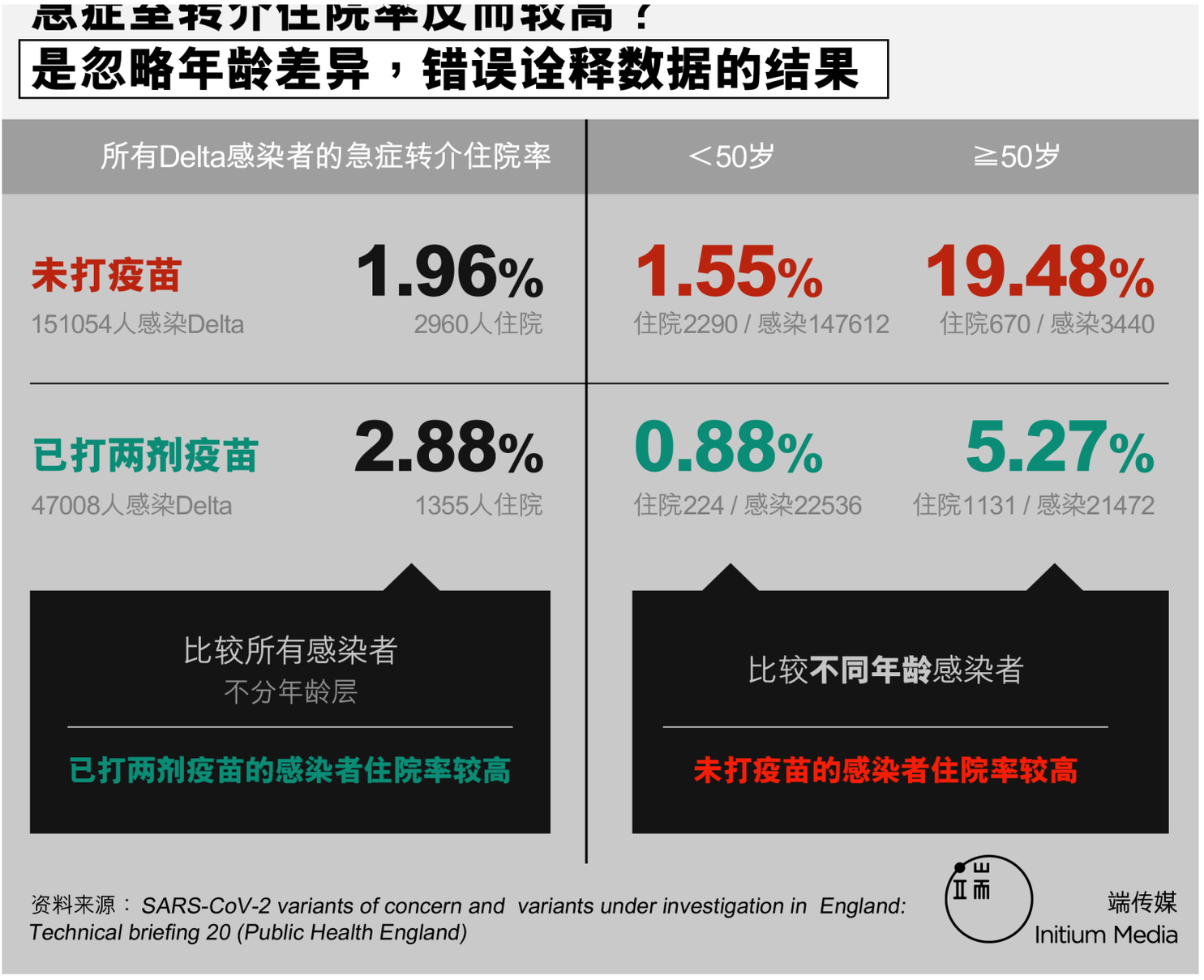
我们从一宗疫苗效用的消息开始谈起。去年夏天，网上流传一份英国政府的研究报告，指已打两针的感染者住院率比未打疫苗的感染者还要高。有网上名嘴拿着这份报告宣称数据“令专家不开心”，更声言打疫苗的风险比不打更大，帖文被广泛传播。

这份报告是真的，数据也确实是从报告的资料中推算出来。统计显示有47008名已接种两剂疫苗者感染Delta病毒，当中有1355人住院；另有151054名未接种疫苗者感染Delta病毒，当中有2960人住院。换言之，已接种两剂疫苗者的住院比率是2.88%，未接种疫苗者则是1.96%，即是已接种者反而更高。

在这案例中，数据虽然正确，但结论却是错误的。以上数字是整体情况，但报告也提供了不同年龄层的情况。如果分开年龄层计算，50岁以下已接种两剂疫苗的感染者和未接种疫苗的感染者，其住院比率分别为0.88%和1.55%；50岁或以上的，则分别为5.27%和19.48%；两者都是未接种疫苗的感染者住院率较高。

为什么分开计算是未接种疫苗的感染者住院率较高，合起来计算却是相反呢？这是因为当时英国的政策是让年纪较大的先接种疫苗，所以已接种疫苗者当中有很多是年纪较大的人。由于他们本来身体状况比较差，所以入院的比例也较高。因此，合起来算的结果其实没有显示疫苗带来风险，而是显示了年龄带来的风险。没有考虑这点，便是误读数据。

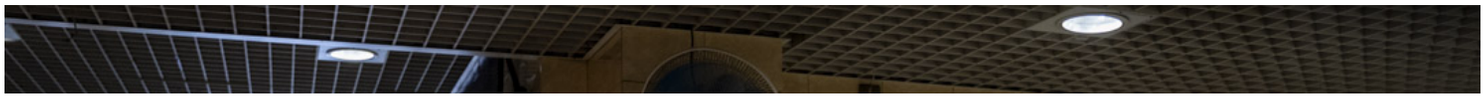
英国接种疫苗者感染Delta后，
急症室转介住院率反而较高？



同样的数据，合起来算和分开算的结果互相矛盾，在数据分析当中相当普遍，学术上称为“辛普森悖论”（Simpson’s paradox）。它会让分析结果看起来违反直觉，研究者要小心考虑数据应如何分类和比较才比较贴近事实。换句话说，我们要学懂“苹果和苹果比较、橙要和橙比较”。困难的地方，是要先意识到在当前的研究问题下，面前的到底只是一堆生果，还是应分为苹果和橙。

早前有香港网媒处理葵涌邨的疫情爆发时，亦犯了类似的错误。

该网媒统计葵涌邨疫情爆发的数字，发现有55人没有接种疫苗，曾接种疫苗的则有61人，有接种疫苗者略为多一点（占整体53%）。即使这些数字本身正确，网媒不加处理就把两个数字并列出来，仍会产生误导效果。



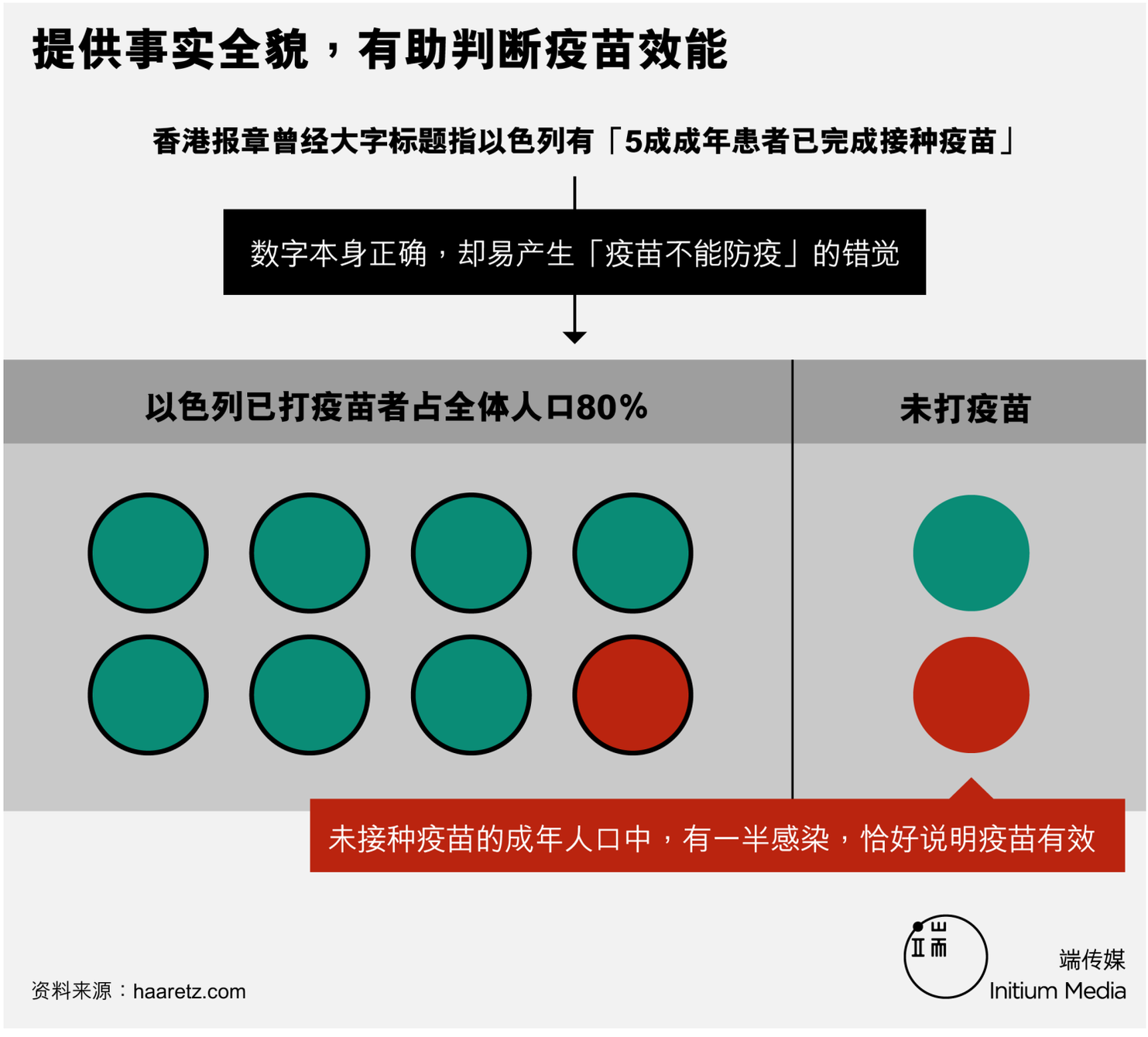
2022年1月25日，消毒人员为葵涌邨的大厦大堂消毒。摄：林振东/端传媒

事实上，该网媒依此组数字制作的社交媒体图片迅速被网民转贴到各网上讨论区，被视为“疫苗不能防疫”的所谓“证据”。

回到“苹果要和苹果比较”的立足点，上述的理解当然是错误的。要知道疫苗能否抗疫，我们需要把整条葵涌邨当中有接种疫苗者和未接种疫苗者分为两组处理，再看每一组当中染疫者的比例。虽然我们没有这些数据，但我们可假设葵涌邨的情况和香港的整体相差不太远，即约有七成居民曾经接种疫苗。如是者，如果疫苗真的不能防疫，我们应预期染疫者当中也该大约有七成曾经接种疫苗。现在染疫者的比例低于七成，在未考虑其他因素影响的前提下，其实恰恰反过来说明疫苗很可能发挥了作用。

类似的情况，在整场疫情当中不停出现。例如有香港报章曾经大字标题指以色列有“5成成年患者已完成接种疫苗”，虽然这个数字本身正确，但标题却无助读者理解当时以色列的疫情，甚至容易产生误解，读者一不留神，同样很容易产生“疫苗不能防疫”的错觉。

实情是当时未接种疫苗者只占成年人口的五分之一左右，所以当有一半患者属未接种疫苗，则同样恰恰反过来说明疫苗有效。媒体没有提供这个比较的背景，就好像一宗片面的选择性报导一样，只说出事实全貌的一半，还振振有词说自己没有半句是假话，其实明明就是新闻专业失当。



数据需要处境

推而广之，媒体引用任何数据，都应该提供该数据的处境。例如说香港男子组合Error成员郭嘉骏有193厘米高，我们都会得出他身型很高的印象。但我们能有此印象，是因为我们是地球人，知道地球人一般有多高；如果对方是外星人的话，仅仅跟对方说“郭嘉骏有193厘米高”而不同时提供地球人的平均身高，则这句说话本身没有意义。偏偏在疫情当中，这样的问题俯拾皆是。

在疫苗接种初期，不少媒体争相报导民众接种疫苗后的不良反应，例如有台湾传媒就曾以表列方式呈现各县市的“疫苗开打死亡统计（死因待查）”。这儿最少有两个问题。第一，既是死因待查，那么媒体列举案列时是否最少应该把标准说清（例如说14天内不论任何情况死亡）？毕竟说白了传媒在此是在发明一套非官方的点算方法，很有必要交代清楚定义。

第二，就算假设定义合理，纯粹点算各县市的数目本身的意义也很有限，因为没有同时告诉读者同期各县市已经施打的疫苗数目。疫苗安全性是一个概率问题，只有分子没有分母又如何推算概率呢？

即使有了概率，也要提供处境来协助读者理解。例如有研究说接种阿斯利康疫苗有百万分之5的机会产生血栓反应，然而一般读者恐怕很难想像百万分之5是什么意思。这时候，我们要把数字化为一些较易理解的概念，例如说如果整个台北市每人打1次，概率上就会有13人有反应。我们也可以拿其他概率事件作比较，例如台湾每年交通事故的死亡数字是台湾人口的百万分之126，这样相对起来阿斯利康的百万分之5算是多还是少，就有个比较的基础。

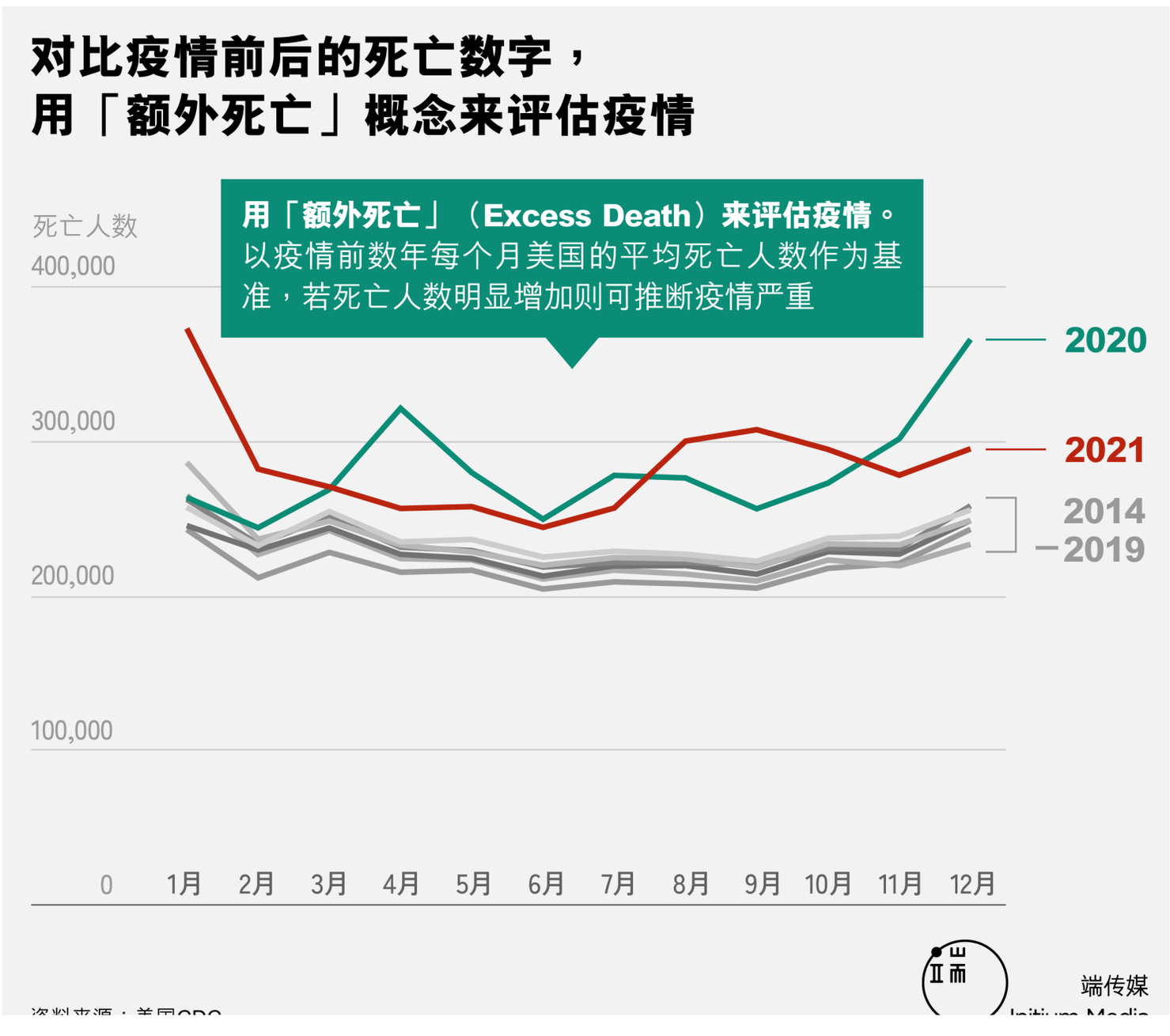
2022年2月10日，沙田大围新翠邨检测站有大量市民轮候排队。摄：林振东/端传媒

提到比例，有些时候我们又不能只看百分比，同时也要与确实数字一起分析。例如在英国一月初Omicron的高峰期时，每日入院人数只是染疫人数的约1%，看起来染疫也不是那么可怕。但因为Omicron本身的传染性极高，就算入院的比例很低但因为染疫的人数极多，仍会为医疗系统带来沉重压力。一个数字到底是大还是细，总是在一个处境当中回答的。

说起Omicron，坊间有不少舆论认为病毒已经弱化，疫情已和流感无异，这点同样可以通过提供处境验证。《纽约时报》统计美国于2022年1月约有6万人因染疫死亡，而当时Omicron已占感染个案的99%，说明感染Omicron还是有死亡风险的。那么6万人算多还算少？在疫情发生前，美国通常在每年的1月份有大约有26万人死亡。有这个比较基础，则说明2022年1月的6万人其实是个相当庞大的数目了。

善用处境去解说数据还有一个好处：可以为数据要解释的现象提供额外的评估。例如疫情的死亡率一直备受争议。有说很多染疫个案没有被发现或被呈报，所以分母被低估了；又有说很多死亡个案不一定是因为染疫而死，而是同时有其他疾病，所以分子被高估了。有些阴谋论者更认为全世界的政府都在合谋用虚假疫情恐吓公众，后面实有其他目的。

要绕过这些困难，有些外国媒体会用“额外死亡”（Excess Death）来评估疫情。举个例，我们可以拿在疫情前数年每个月美国的平均死亡人数作为基准，比较疫情开始后每个月的死亡数字。毕竟过去两年疫情是最明显的额外死亡因素，如果死亡人数明显增加则可推断疫情严重，不用担心会受医院呈报方法所影响。



教育自己 教育公众

提到死亡数字，现时不少国家对疫情的理解已开始转向：不再着重确诊数字，而更重视重症和死亡数字。这是因为疫苗普及后确诊和重症数字开始脱轨，加上Omicron的重症比例相对较低。例如英国在2022年1月最高的平均每日确诊数字约是21万宗，远远比2021年1月的6万宗要高；但是，2022年1月中的高峰的平均每日死亡数字一直少于300宗，而2021年1月的高峰却是超过1200宗。一来一回，如果我们只和读者讲英国的确诊数字不断创新高，却不同时讲死亡数字的升幅远远较低，则同样有误导读者之嫌。

当然，哪一个时候用哪一个标准去衡量疫情，并无划一定律。任何的数据最终都是一个被简化的真实，必然有所偏差，问题是这个偏差是否涉及提问重点，以及有否向读者清楚交代。很多时候，新闻中的数据呈现出了问题，往往基于媒体本身对数据背后的各种属性掌握得不够充分，自己也未搞清那些数字的背后所指，甚至弄出各种笑话。

在疫苗接种的早期，曾有香港网上名嘴质疑政府的疫苗数据作假。他发现政府公布的疫苗接种数字已有240万，同时说香港已接种疫苗的比例不足两成。他按香港700多万人口计算，认为两项数字当中必定是有一项是错误的。事实是政府公布的数字是240万剂而不是240万人，由于有些人已打了两针，所以当时的接种人数其实是142万才对。此外，政府计算接种比例的是以合资格人口算而不是全港人口算，所以分母也不是700多万。该网上名嘴犯上如此基本的错误，未免过于粗心大意。然而类似的简单理解错误，在疫情中数之不尽。

因此，我认为传媒在疫情面前很有迫切提升自身和社会大众的数据素养，既要教育自己，也要教育公众。常说“数据会说话”（The data speaks for itself），专业的数据新闻工作者却恐怕大多不会同意。数据从选取、分析到呈现，每一步都涉及各种选择，依靠各种专业判断；只要略为不留神，就会产生误导而不自知。

我提议三个传媒呈现疫情数据时可以立即改善的方向。

2022年2月11日，香港的地铁站有不少政府呼吁接种疫苗的广告。摄：林振东/端传媒

第一，处理数据时行文用字应务必力求准确。例如曾有政府官员说疫苗“不一定能防感染”时被传媒略写为“不能防感染”，就可以带来极大的误解；后者很容易被误读为完全不能。事实上，英国的数据显示接种3针复必泰对Omicron有症状感染，有约七成保护力；虽然不是百分之百，却亦绝非毫无作用。

传媒应教育自己和教育公众以概率来理解疫情，避免简单的二分法。正如乘客配带安全带也“不一定能”防止乘客于遇上交通意外时受伤，但无阻政府规定乘客配带，而大多数乘客都会乐于遵守。

第二，提醒公众注意数据有时会受其他和疫情不相关的因素影响。例如1月初美国单日确诊数字曾突破100万宗，但其实是因为之前数天刚好放假，地方数据上报中央出现滞后，几天的数据被叠在一起，当时的平均每日确诊数字不足50万宗。美国的大型媒体报导疫情时，会清楚把这些不正常的数值标记出来并作解释，以免读者误解。

类似的情况也可在疫苗接种数字中看到：香港的疫苗接种数量往往是逢星期六特别高，因为很多人觉得星期六接种后可以利用星期日休息。因此，报导相关数字时的常见做法是加入7天平均数，避免一星期内不同天数之间的差异所产生的影响。

第三，提醒公众注意数据有时未能立即反映后来的影响。例如英国的确诊数目在2022年1月1日到达最高峰，但同期死亡数目的最高峰则要到了1月16日才出现，毕竟病人从确诊到死亡之间是有时差的。传媒的责任，就是当确诊数目已经上升，死亡数目却仍然维持低水平之际，提醒读者不要因为那一刻两组数字差距巨大而误以为死亡率低。

类似的情况也可应用在一些知名人士染疫的消息当中。当一些名人公布自己染疫而又病情轻微，便很容易让很多人误以为是必然现象。现实是名人之所以为名人，是因为他们数目有限，但这也暗示他们的代表性亦有限。正如一颗骰子摇6次，结果1至6各出现一次的机会不大；然而如果你摇600万次，结果1至6的出现次数相若的可能性就很大。有时得承认眼前有限的观察并不足以让我们立即跳到结论。

最后，我必须要强调一点：有些事情是数据解决不了的。数据可以告诉你有什么事情已发生，或可能发生，但不能告诉你什么事情应该或不应该发生。特别是一些涉及价值观的问题，数据分析不能完全取代主观判断。例如数据可以告诉你采取不同防疫手段对疫情的影响，但不能代替社会回答是否愿意接受这些手段。

疫情是科学，但公共政策却不能离开政治，而政治讲求公信。